UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE MATEMÁTICA

CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ANNANDA DANDI DE FREITAS SOUSA

REFINANDO A EXTRAÇÃO DE GRAFOS DE RELACIONAMENTO SEMÂNTICO PARA TEXTOS EM PORTUGUÊS

RIO DE JANEIRO

2019

ANNANDA DANDI DE FREITAS SOUSA

REFINANDO A EXTRAÇÃO DE GRAFOS DE RELACIONAMENTO SEMÂNTICO PARA TEXTOS EM PORTUGUÊS

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. João Carlos Pereira da Silva, D.Sc.

RIO DE JANEIRO

2019

CIP - Catalogação na Publicação

S725r

Sousa, Annanda Dandi de Freitas
Refinando a extração de grafos de relacionamento semântico para textos em português / Annanda Dandi de Freitas Sousa. -- Rio de Janeiro, 2018.
70 f.

Orientador: João Carlos Pereira da Silva.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciência da Computação, 2018.

1. Dados ligados. 2. Dados estruturados. 3. Grafos de relacionamento. 4. Processamento de linguagem natural. I. Silva, João Carlos Pereira da, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

ANNANDA DANDI DE FREITAS SOUSA

REFINANDO A EXTRAÇÃO DE GRAFOS DE RELACIONAMENTO SEMÂNTICO PARA TEXTOS EM PORTUGUÊS

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 29 de junho de 2018.

BANCA EXAMINADORA:

Prof. João Carlos Pereira da Silva, D.Sc.

Danilo Silva de Carvalho, Ph.D.

Fabrício Firmino de Faria, M.Sc.

AGRADECIMENTOS

Agradeço ao meu paizinho (João Silva), meu modelo de vida, esse ser humano incrível que tenho a sorte de chamar de paizinho. Obrigada por sempre me apoiar, me amar, acreditar em mim e me contar histórias dos meus antepassados que me inspiram todos os dias. Sem você eu não seria quem sou e não estaria aqui. Um dos maiores desejos da minha vida é fazer você se orgulhar de mim.

Agradeço à minha mãe (Cicera de Freitas), irmãs e irmão. Vocês são o tesouro da minha vida e me ensinam muito. Minha mãe me ensinou a servir por amor, minha irmã Dani me ensinou generosidade, minha irmã Tutu me ensinou aceitação incondicional e justiça, e meu irmão Vitinho me ensinou que amor à primeira vista existe, eu o amei desde o primeiro segundo que vi aquele bebezinho que ele foi, te amo meu Nandinho. Carrego vocês no meu coração, ansiando pelo nosso próximo encontro.

Obrigada ao meu noivo e amor da minha vida Felipe. Obrigada por tudo o que vivemos juntos, por me amar exatamente como eu sou, por entender e aceitar meu jeito, por sempre estar lá para mim, por me abraçar apertado quando a ansiedade e o medo me fazem chorar, você é meu lar. Obrigada aos pais e irmã do Felipe, que me adotaram como filha e irmã. Obrigado por me fazerem experimentar de novo o que é estar em família.

Obrigada a todos os professores que me inspiraram na minha jornada de aprendizado desde a infância. Obrigada ao professor João Bosco, por sua causa eu me apaixonei por Física e comecei a sonhar em estudar na UFRJ, por sua influência eu sonho grande até hoje. Obrigada a professora Jonice por, após a separação dos meus pais, me escrever um e-mail que me ajudou a não desistir. Obrigada a professora Maria Luíza que sempre fez tudo para me ajudar em relação aos processo burocráticos da faculdade. Obrigada às professoras Luziane e Juliana, vocês são para mim modelos de mulheres na carreira acadêmica. Espero um dia ser como vocês. Obrigada ao professor Luís Menasche porque graças às discussões durante sua disciplina sobre Filosofia da Ciência eu tive certeza do meu desejo de seguir na

área acadêmica e me tornar uma pesquisadora.

Obrigada a todos os meu amigos pela amizade. Sou feliz por cada pessoa que posso chamar de amigo.

Obrigada ao professor João Carlos porque além de me guiar por todo esse trabalho, me ensinando, corrigindo, dando ideias e apoiando, também aceitou me orientar remotamente pelos últimos 10 meses, o que me permitiu fazer um estágio acadêmico em pesquisa fora do Brasil que com certeza mudou minha vida. Muito obrigada!

Agradeço ao universo ou qualquer que seja a força que traz tanta sorte e coisas boas para mim. A todos as pessoas que já cruzaram meu caminho e trouxeram bondade, ensinamentos e ajuda. Citando Newton eu digo: "se cheguei até aqui é porque me apoiei no ombro de gigantes". Obrigada a todos que foram os meus gigantes.

RESUMO

A tecnologia de interação homem-máquina evolui a cada dia. Atualmente, é cada vez mais comum termos acesso a aplicações às quais os comandos são dados em linguagem natural, e o *software* é capaz de processar as sentenças recebidas, gerando outras semanticamente coerentes ao que foi requisitado. Para ser capaz de usar linguagem natural como entrada e saída, os programas de computador precisam fazer uso de dados estruturados. É a navegação através de dados estruturados, como grafos de relacionamento semântico, que permite que conhecimento seja acessado rapidamente por computadores. Por razão do grande volume de dados não-estruturados sendo produzidos a cada segundo na Web, a geração de dados estruturados é inviável manualmente. Como conseguinte, é de vital importância a criação de ferramentas para a extração automática de dados estruturados.

O trabalho apresentado tem como foco o problema de geração automática de dados estruturados. Ele é construído a partir do Graphia, propondo melhorias ao mesmo. Graphia é um extrator automático de grafos de relacionamento. Ele recebe como entrada sentenças em linguagem natural e produz dados ligados na forma de grafos de relacionamento. Este documento apresenta uma nova versão para o Graphia aplicado a sentenças em português. Também reporta uma avaliação que compara as duas saídas do Graphia: antes e depois da aplicação das modificações sugeridas. Os resultados revelam que de fato houve uma melhora significativa na extração das relações ao utilizar as alterações propostas.

Palavras-chave: Dados ligados. Dados estruturados. Grafo de relacionamento. Processamento de linguagem natural. Extração de conhecimento..

ABSTRACT

In each passing day, new advances regarding human-computer interaction occur. Nowadays, it is very common for one to deal with computer programs capable of receiving and processing sentences into semantic appropriated responses. In order to use natural language as input and output, programs need structured data. It is the navigation through structured data, such as semantic graphs, that allows the computer to access the available knowledge. The structured data extraction is only viable through automatic techniques, due to the huge amount of non-structured data being produced on the Web in each minute. Therefore, it is very important the development of tools for structured data automatic extraction.

This work addresses the problem of structured data automatic extraction. It builds on Graphia work by proposing enhancements to it. Graphia is an automatic extractor of semantic graphs. It accepts as input sentences in a natural language and produces structured data. This document proposes a new version for Graphia applied to Portuguese sentences. In addition, it brings an evaluation to compare the different Graphia's outputs: the original and the one produced with the modifications applied. Results show that a relevant improvement happens when using the new approach to extract semantic graphs.

Keywords: Linked data. Structured data. Semantic graphs. Natural language processing. Knowledge extraction.

LISTA DE FIGURAS

Figura 1: Árvore sintática para a sentença do Exemplo (5): ”Maria salvou o coelho” 23
Figura 2: Esquema de um grafo de relações em que o sujeito e objeto são vértices, enquanto a relação é a aresta do grafo. 24
Figura 3: Grafo semântico para a sentença do Exemplo (5): ”Maria salvou o coelho” 24
Figura 4: Grafo semântico com reificação para a sentença do Exemplo (7): ”Maria salvou o coelho na semana passada”. 25
Figura 5: Esquema do pipeline do Graphia 26
Figura 6: Saída do pipeline do Graphia versão 1 para a sentença do Exemplo (8): ”Donald John Trump (born June 14, 1946) is the 45th and current President of the United States.” 27
Figura 7: Pipeline do Graphia versão 3 com destaque para as duas últimas fases que foram modificadas para a versão em português 31
Figura 8: Saída do pipeline do Graphia versão 3 para a sentença do Exemplo (9): ”Thiago conhece algumas pessoas na cidade e tem vinte amigos.” 32
Figura 9: Grafo gerado pelo Graphia versão 3 para a sentença do Exemplo (10): ”A porcentagem de população ativa empregada no setor terciário depende bastante do nível de desenvolvimento de cada país”. Complementos nominais aparecem enumerados. 33
Figura 10: Grafo gerado pelo Graphia português para a sentença do Exemplo (11): ”A luz nesse comprimento de onda é absorvida pela atmosfera da Terra, então as observações devem ser feitas na atmosfera superior ou no espaço”. Os complementos nominais extraídos aparecem enumerados. 35
Figura 11: Grafo esperado para a sentença do Exemplo (12): ”Os aeroportos da Antártica estão sujeitos a severas limitações por causa das condições climáticas e geográficas” 37
Figura 12: Pipeline do Graphia versão 4 com destaque para a etapa modificada neste trabalho. 38

Figura 13: Grafo produzido usando a abordagem inicial de regras de extração para da sentença do Exemplo (13): "A decepção com o governo Moreira Franco, que não cumpriu a promessa de acabar com a violência em seis meses, levou o eleitorado fluminense a eleger Leonel Brizola novamente, em 1990" 41
Figura 14: Esquema apresentando a proposta para a construção básica de relacionamento no grafo semântico. 42
Figura 15: Árvore sintática da sentença do Exemplo (13): "A decepção com o governo Moreira Franco, que não cumpriu a promessa de acabar com a violência em seis meses, levou o eleitorado fluminense a eleger Leonel Brizola novamente, em 1990". 43
Figura 18: Grafo extraído utilizando as novas regras propostas para a sentença do Exemplo (11): "A luz nesse comprimento de onda é absorvida pela atmosfera da Terra, então as observações devem ser feitas na atmosfera superior ou no espaço". 46
Figura 16: Grafo extraído utilizando as novas regras propostas para a sentença do Exemplo (13): "A decepção com o governo Moreira Franco, que não cumpriu a promessa de acabar com a violência em seis meses, levou o eleitorado fluminense a eleger Leonel Brizola novamente, em 1990". 48
Figura 17: Grafo extraído utilizando as novas regras propostas para a sentença do Exemplo (10): "A porcentagem de população ativa empregada no setor terciário depende bastante do nível de desenvolvimento de cada país". 48
Figura 19: Grafo gerado para a sentença do Exemplo (14): "A densidade populacional é de 15 pessoas por quilômetro quadrado de área de terra". O sujeito simples (1), verbo (2) e complementos nominais (3) e (4) foram extraídos corretamente. 50

Figura 20: Grafo gerado da sentença do Exemplo (15): ”Poisson, Liouville, Fourier e outros mais estudaram as equações em derivadas parciais e a análise harmônica”. O sujeito composto ”Poisson, Liouville, Fourier e outros” foi corretamente extraído (marcado em vermelho). 51

Figura 21: Grafo extraído da sentença do Exemplo (10): ”A porcentagem de população ativa empregada no setor terciário depende bastante do nível de desenvolvimento de cada país”. Ele é avaliado como incorreto segundo a categoria informação redundante. É possível observar as informações redundantes marcadas em vermelho. . . . 53

Figura 22: Árvore sintática da sentença do Exemplo (17): ”Poisson, Liouville, Fourier e outros mais estudaram as equações em derivadas parciais e a análise harmônica” 67

LISTA DE TABELAS

Tabela 1: Quantidade em frequência relativa de erros em extrações por categoria de erro. Tabela extraída do trabalho [4] 29

Tabela 2: Estruturas sintáticas e onde estão presentes na sentença do Exemplo (12): "Os aeroportos da Antártica estão sujeitos a severas limitações por causa das condições climáticas e geográficas" 37

Tabela 3: Resultados da comparação de desempenho entre os sistemas Graphia versão 3 e versão 4 em porcentagem de acerto para cada categoria avaliada. 54

Tabela 4: Detalhes sobre a avaliação das sentenças, apresentando seu número no website, seu ID no Banco de Dados, o número de instâncias de cada categoria avaliada e o número de acerto. 65

Tabela 5: Tabela com as nomenclaturas usadas para se referir às versões do Graphia 68

LISTA DE ABREVIATURAS E SIGLAS

ART	Artigo
CN	Common Nouns
DA	Definite Articles
EIA	Extração de Informação Aberta
EN	Entidade Nomeada
HTML	Hyper Text Markup Language
N	Noum
NP	Noun Phrase
PLN	Processamento de Linguagem Natural
PNM	Part of Name
PNT	Punctuation Mark
POS	Part-of-speech
RDF	Resource Description Framework
REN	Resolução de Entidade Nomeada
S	Sentence
SBAR	Subordinate Clause
SN	Sintagma Nominal
URI	Uniform Resource Identifier
V	Verb
VP	Verb Phrase

SUMÁRIO

1 INTRODUÇÃO **14**
1.1 MOTIVAÇÃO 14
1.2 OBJETIVO DO TRABALHO 16
1.3 TRABALHOS RELACIONADOS 16
1.3.1 Contextualização na área de extração de relacionamento semântico **17**
1.4 ESTRUTURA DO TRABALHO 18

2 CONCEITOS **19**
2.1 MORFOLOGIA 19
2.2 SINTAXE 19
2.3 SEMÂNTICA 21
2.4 PROCESSAMENTO DE LINGUAGEM NATURAL 21
2.4.1 Árvore sintática **23**
2.4.2 Grafo Semântico **23**

3 GRAPHIA **26**
3.1 GRAPHIA - SENTENÇAS EM INGLÊS 26
3.2 GRAPHIA - ANÁLISE DE DESEMPENHO E UMA NOVA ALTERNATIVA 30
3.3 GRAPHIA - SENTENÇAS EM PORTUGUÊS 31

4 PROBLEMAS E SOLUÇÕES APRESENTADAS **36**
4.1 NOVA VERSÃO GRAPHIA PARA SENTENÇAS EM PORTUGUÊS 36
4.1.1 Grafo de relação para representação de sentenças **36**
4.1.2 Refatoração do código **39**
4.1.3 Alteração nas regras de extração de relações **40**
4.1.4 Aplicando a nova versão para extrair grafos de relacionamento **46**

5 AVALIAÇÃO E RESULTADOS **49**
5.1 MÉTODO DE AVALIAÇÃO 49
5.1.1 Categorias de avaliação **49**
5.1.2 Seleção de amostra de sentenças **52**
5.1.3 Resultados e discussão **54**

6 CONCLUSÃO E TRABALHOS FUTUROS **57**

REFERÊNCIAS **60**

ANEXO A **62**
ANEXO B **66**
ANEXO C **68**
ANEXO D **69**

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

A cada dia temos acesso a avanços no modo como nós humanos interagimos com os computadores e máquinas. Nos dias de hoje, nós podemos perguntar para *Alexa*[1] como está o tempo antes de sair de casa, juntamente com recomendações sobre qual restaurante próximo está aberto para almoço. E ainda podemos falar com o *Google Duplex* [2] requisitando que ele faça ligações telefônicas (conversando com pessoas reais) para realizar nossos agendamentos de compromissos.

Mesmo a tecnologia evoluindo ao ponto de nos fazer parecer que o computador entende o que uma pessoa está falando, a verdade é que computadores não conseguem compreender a linguagem humana. Os computadores podem, entretanto, processar a linguagem natural, recebendo entradas e produzindo saídas a fim de gerar a ilusão de compreensão e executar as tarefas requisitadas. Para tal, é necessário o chamado dado estruturado. Dados estruturados são as informações que se enquadram em um conjunto de organização pré-determinado. Essa formatação permite que o atual aparato computacional possa facilmente navegar e extrair conhecimentos dos mesmos. É ao navegar por dados estruturados (na forma de grafos semânticos) que *Alexa* é capaz de processar que você quer saber quantas calorias tem um pedaço de pizza e consegue dar a correta resposta para tal pergunta.

O problema dos dados estruturados é que essa não é a forma em que os dados em linguagem natural são produzidos. Conteúdos em linguagem natural são gerados como dados não-estruturados, que são os dados não organizados e formatados, não sendo possível ao computador navegar e extrair conhecimento dos mesmos. É necessário então que o dado não-estruturado seja convertido em dado estruturado. Somado a isso, nos tempos modernos de redes sociais, o volume de dados em linguagem natural não-estruturado produzido é muito grande, tornando inviável que essa conversão seja feita de forma manual. Ao não ter esses dados em forma estruturada,

[1] `https://developer.amazon.com/alexa` acessado em 12 de Junho de 2018.
[2] `https://www.youtube.com/watch?v=D5VN56jQMWM` acessado em 12 de Junho de 2018.

perde-se potenciais conhecimentos que poderiam ser obtidos se os mesmos estivessem em outro formato. Por essa razão se faz necessário a criação de ferramentas que extraiam automaticamente dados estruturados a partir de dados não-estruturados.

Quanto maior a quantidade de informação estruturada disponível, mais significado e ligação podem ser obtidos delas. *Linked data web*[3] atualmente chamada de *web of data* é um termo criado por Tim Berners-Lee (criador da *web*) em 2006 e diz respeito a ação de disponibilizar na *web* dados estruturados e ligados entre si seguindo padrões de estruturação. O objetivo é transformar a *web* baseada em documentos (que só humanos podem ler) na *web* baseada em dados ligados (que os computadores podem navegar, processar e extrair conhecimento). Para isso foi preciso padronizar o formato dos dados estruturados. Hoje, uma linguagem muito conhecida e usada para descrição dos dados estruturado é a RDF[4]. O RDF é uma linguagem que usa triplas para descrever o relacionamento entre entidades. As triplas contém: sujeito, verbo e objeto. E uma coleção de triplas RDF é chamada de grafo RDF.

Atualmente já existem grandes bases de dados disponíveis em dados estruturados, como por exemplo a DBPedia[5] (uma base de dados com informações extraídas da Wikipédia) e a FOAF[6] (uma base de dados sobre pessoas, suas atividades e relacionamentos). Entretanto, estima-se que em torno de 80% dos dados da *web* ainda são não-estruturados [15]. Como consequência, a geração de dados estruturados é uma tarefa ainda muito necessária e pertinente. Contudo, por mais que esse seja um tópico de pesquisa relevante no momento, a maioria dos estudos conduzidos e soluções disponíveis são para a língua inglesa. Pelo nosso atual conhecimento, poucos dos trabalhos publicados tiveram como principal atenção a língua portuguesa. Então, dado o ainda limitado foco em extrair conhecimento aplicado à essa língua, a motivação desse trabalho é contribuir com a geração automática de dados ligados com conteúdo em português.

[3] `https://www.w3.org/standards/semanticweb/data` acessado em 12 de Junho de 2018.
[4] `https://www.w3.org/TR/rdf-concepts/` acessado em 9 de Junho de 2018.
[5] `https://wiki.dbpedia.org/` acessado em 12 de Junho de 2018.
[6] `http://lov.okfn.org/dataset/lov/vocabs/foaf` acessado em 12 de Junho de 2018.

1.2 OBJETIVO DO TRABALHO

O objetivo desse trabalho é produzir uma solução que extraia automaticamente dados estruturados de textos em linguagem natural em português, a fim de torná-los disponíveis a sistemas capazes de obter informações de tais estruturas. Tornando possível então, extrair conhecimento automaticamente de texto em linguagem natural escritos na língua portuguesa. O Graphia para sentenças em português [2] cujo objetivo era o mesmo, deixou como sugestão de trabalhos futuros a implementação correta da regra do complemento nominal e uma avaliação formal do desempenho do sistema. É então, parte do escopo do trabalho aqui apresentado contribuir para a extração de conhecimento ao atacar ambos esses pontos deixados em aberto pelo trabalho anterior. Como consequência, esse trabalho tem o intuito de apresentar melhorias para o sistema Graphia, assim como propor um *framework* para a avaliação do mesmo.

1.3 TRABALHOS RELACIONADOS

Como mencionado antes, o problema de extrair dados estruturados a partir de conteúdo em linguagem natural tem recebido atenção, com vários estudos e ferramentas já publicados a esse respeito. Alguns exemplos de trabalhos relacionados a esse tópico são: Graphene[7], ReVerb [6], ArgOE [7] e o trabalho de LIMA SENA et al.[10]

Graphene é um pipeline de extração de informação que gera grafos de conhecimento a partir de sentenças em inglês. Enquanto outras ferramentas de extração existentes focam na relação mais importante expressa em uma sentença, o Graphene busca maximizar a extração do contexto das relações ao encontrar tempo, quantidade e causa. O ReVerb é um programa que identifica e extrai relações binárias (fatos) de sentenças em inglês. Ele é escalável, aplicável para casos em que as relações desejadas não são especificadas previamente e tem foco em velocidade. É baseado em heurísticas criadas manualmente que usam padrões usando *POS-tags*

[7] `https://github.com/Lambda-3/Graphene` acessado em 12 de Junho de 2018.

das sentenças originais.

ArgOE é um extrator de informação multilingual (inglês, português e espanhol). Ele usa árvores de dependência a fim de criar regras para extrair relações binárias (fatos). Embora esse sistema inclua a língua portuguesa, os resultados são impressionantes apenas para sentenças em inglês [10]. LIMA SENA et al.[10] extraem automaticamente fatos de textos escritos em português e fazem inferências sobre os fatos inicialmente extraídos para criar novos fatos. Esse sistema usa regras manualmente criadas que se baseiam em reconhecimentos de padrões sobre *POS-tags* e *NP-chunker*[8] *parser* para extrair os fatos. Ele usa aprendizado de máquina para inferir os novos fatos.

1.3.1 Contextualização na área de extração de relacionamento semântico

O Graphia é um sistema de extração de informação aberta[9] (EIA). Extração de informação aberta é uma das abordagens existentes para obter informação semântica. O objetivo de sistemas desse tipo é extrair triplas de relacionamento a partir de textos não estruturados. WU e WELD[17] definem um sistema de extração de informação aberta como uma função de um documento d para um conjunto de triplas *relação(sujeito, objeto)*, que representam a relação, determinada pelo verbo, entre os dois argumentos.

Diferente de outras abordagens, EIA não é limitada por um conjunto pré-definido de relações alvo. Ela extrai todos os tipos de relações binárias (definidas pelo verbo) encontradas no texto. São características dessa abordagem: (i) independência de domínio, (ii) métodos de extração não supervisionados e (iii) são escaláveis para grandes quantidades de texto.

Sistemas de EIA podem ser classificados em duas grandes categorias: (i) aqueles que precisam de dados para o treinamento da extração e (ii) sistemas baseados em regras manualmente criadas e heurísticas, não necessitando de dados para treinamento. Cada uma dessas grandes categorias pode ser subdividida em dois subtipos:

[8] Um parser que identifica as partes da sentença que correspondem a NP - Noun Phrases

[9] Em inglês Open Information Extraction - OIE

baseado em análise sintática (e.g. *POS-tagging*, *chunking*, árvores sintáticas) ou baseado em árvores de dependência[7].

O Graphia utiliza regras manualmente criadas e análise sintática. Outro trabalho que é classificado na mesma categoria é o ReVerb, entretanto ele é aplicado apenas para sentenças em inglês. O trabalho que mais se aproxima do Graphia para sentenças em português é o do LIMA SENA et al.[10] Ele se encontra nas mesmas categorias, mas não tem como produto final grafos de relacionamento e sim fatos. Além disso, o Graphia é capaz de extrair contexto das relações como tempo e quantidade e outras estruturas sintáticas como complemento nominal. Características que o trabalho de LIMA SENA et al. não compartilha. Então o Graphia é um sistema mais completo em relação a tipos de extração se comparado ao trabalho de LIMA SENA et al, por essa razão foi escolhido como base para o presente trabalho.

1.4 ESTRUTURA DO TRABALHO

Esse documento está dividido em 6 capítulos. O capítulo 2 apresenta conceitos básicos necessários para o entendimento das discussões conduzidas adiantes. O capítulo 3 fala sobre o sistema Graphia e os trabalhos anteriores relacionados a ele. Já no capítulo 4 é descrito as modificações e melhorias aplicadas ao Graphia, enquanto que o capítulo 5 traz uma avaliação que compara o desempenho entre as versões do Graphia em português. Por fim, o Capítulo 6 apresenta a conclusão juntamente com perspectivas para trabalhos futuros. Adicionalmente, o trabalho inclui 4 anexos contendo detalhes da avaliação, sugestão de melhoria futura relacionada a sujeito composto, tabela guia para os nomes das versões do Graphia e detalhes sobre os testes unitários, respectivamente.

2 CONCEITOS

Neste capítulo são apresentados concisamente alguns conceitos necessários para os tópicos e discussões posteriores. Para os conceitos sobre língua portuguesa foram consultados as gramáticas [14] e [12].

2.1 MORFOLOGIA

Morfologia é a parte da linguística que estuda a estrutura, a formação e a classificação das palavras. Na língua portuguesa, quase todas as palavras se distribuem entre 10 classes. Substantivo, verbo, e adjetivo são alguns exemplos dessas classes. As classes de palavras mais importantes são substantivo e verbo.

2.2 SINTAXE

Sintaxe é a parte da linguística que estuda a relação e combinação das palavras entre si para formarem frases dentro de um discurso lógico. Dentro de uma sentença existem diversas funções sintáticas, como por exemplo: sujeito, predicado, predicativo do sujeito e complemento nominal.

Oração. Na língua portuguesa, oração é todo o conjunto linguístico que se estrutura em torno do verbo (ou locução verbal), apresentando opcionalmente um sujeito, mas obrigatoriamente um predicado [12]. As sentenças do Exemplo (1) apresentam tipos de orações.

(1) a. Alessandra estuda bastante. [Oração com sujeito]

b. "Havia formigas na casa." [Oração sem sujeito]

Sujeito. Sujeito é o agente praticante da ação verbal na oração. Existem oficialmente três tipos de sujeito: simples, composto e indeterminado. Também é possível haver orações sem sujeito. No sujeito simples há apenas um núcleo, no sujeito com-

posto há mais de um núcleo. Já no sujeito indeterminado, não se conhece ou se quer determinar a identidade do praticante da ação verbal. Existem também casos em que o sujeito não está explícito na oração, mas ele pode ser identificado pela flexão do verbo. Nesses casos ele é chamado sujeito simples implícito na desinência verbal (também conhecido como sujeito oculto) [14]. As sentenças do Exemplo (2) apresentam os tipos de sujeito.

(2) a. "As **aulas** começam cedo." [Sujeito simples]

b. "As **aulas** e a **ginástica** começam cedo." [Sujeito composto]

c. Roubaram minha carteira. [Sujeito indeterminado]

d. "Fomos roubados" [Sujeito implícito/oculto]

Predicado. É tudo aquilo que se diz sobre o sujeito em uma oração. Ao encontrar o sujeito da oração, tudo o que sobra é o predicado.

Verbo. É a classe de palavras que indica processos como: ação, estado, fenômeno, etc. Os verbos podem aparecer sozinhos nas sentenças ou vir junto com outros verbos criando um sentido único, quando isso acontece, há a **locução verbal**. O verbo é uma classe de palavras dentro da morfologia, e ele pode desempenhar diferentes funções sintáticas. Existem basicamente três tipos de funções sintáticas que o verbo desempenha: verbo de ligação, verbo transitivo e verbo intransitivo. O verbo de ligação tem a função de ligar o sujeito e suas características, não expressando ação. E nisso ele se distingue dos verbos transitivos e intransitivos pois os mesmos expressam ação praticada ou sofrida. As sentenças do Exemplo (3) apresentam os diferentes tipos de verbo.

(3) a. "O carro **atropelou** um pedestre" [Verbo transitivo direto]

b. "Eu **gostei** do livro" [Verbo transitivo indireto]

c. "A borboleta **morreu**." [Verbo intransitivo]

d. "Juçara **está** bonita" [Verbo de ligação]

Objeto, núcleo do objeto e predicativo do sujeito. Quando se tem um verbo do tipo transitivo direto ou indireto, significa que o verbo não tem sentido completo por si só, ele precisa de algo que o complemente. O que complementa os verbos desse tipo é chamado objeto. Núcleo do objeto é a palavra mais importante e que carrega o maior significado dentro do objeto. Predicativo do sujeito é o que vem depois de um verbo de ligação, ele contém características do sujeito. Nas sentenças do Exemplo (3), "um pedestre"é objeto direto, "pedestre"é núcleo do objeto, e "bonita"é predicativo do sujeito.

Complemento nominal. É todo termo na oração que complementa o sentido de um nome (substantivo, adjetivo ou advérbio). Sempre há presença de preposição no complemento nominal. No Exemplo (4), é possível ver uma oração com complemento nominal destacado.

(4) "Tenho saudades **de Cristina**."

2.3 SEMÂNTICA

Em linguística, a semântica estuda o significado de uma palavra, sintagmas ou frases em determinado contexto. Para sintetizar, podemos dizer que a morfologia se ocupa com a forma e a classe das palavras, a sintaxe com a combinação das palavras entre si e a semântica com o significado das palavras e suas combinações em diferentes contextos.

2.4 PROCESSAMENTO DE LINGUAGEM NATURAL

Processamento de linguagem natural[1] (PLN) é uma área de pesquisa e aplicação que explora como computadores podem ser usados para entender e manipular

[1]Linguagem falada ou escrita por seres humanos, por exemplo, texto escrito em língua portuguesa

texto ou discursos em linguagem natural para fazer tarefas úteis [5]. Para tal, geralmente é necessário cumprir algumas tarefas, elas são *tokenization*, *POS tagging* e *constituency parsing*. Para ilustrar vamos usar a sentença do Exemplo (5).

(5) Maria salvou o coelho.

Tokenization. Etapa em que uma sentença é dividida em partes chamadas tokens. No geral, os tokens de uma sentença são equivalentes às palavras da mesma. Os tokens da sentença do Exemplo (5) são: "Maria", "salvou", "o"e "coelho".

POS Tagging. Nessa etapa cada token é etiquetado com uma *POS-tag* que representa suas classes gramaticais (MITKOV, 2004)[13]. Usando um *POS-tagger*[2] na sentença do Exemplo (5) obtemos (6):

(6) Maria/PNM[3] salvou/V[4] o/DA[5] coelho/CN[6] ./PNT[7]

Lemmatization. É o processo de agrupar diferentes formas flexionadas de uma palavra em sua forma inflexionada a fim de serem analisadas como um único item. Por exemplo, as palavras "jogaram", "jogou"e "joguei"todas ao passar pelo processo de *lemmatization* produziriam "jogar", que seria a palavra inflexionada dessas formas verbais.

Constituency Parsing. É nessa etapa que a árvore sintática[8] apresentada na seção seguinte é gerada. Então o produto final são árvores sintáticas como a da Figura 1.

[2] `http://lxcenter.di.fc.ul.pt/tools/en/LXTaggerEN.html` e `http://lxcenter.di.fc.ul.pt/services/en/LXServicesSuite.html` acessados em 9 de Junho de 2018.

[3] Part of Name, representa substantivos próprios

[4] Verb, representa verbo

[5] Definite Articles, representa artigo definido

[6] Common Nouns, representa substantivos comuns

[7] Punctuation Marks, representa pontuações

[8] Também chamada Constituency Tree.

2.4.1 Árvore sintática

Uma árvore sintática é uma estrutura que representa os constituintes sintáticos de uma string[9] de acordo com alguma Gramática Formal[10]. Nessas árvores, o nó raiz é o um nó chamado "S", que representa toda a sentença, os sintagmas[11] são representados por nós não-terminais e as palavras são representadas pelas folhas da árvore. A classe gramatical das palavras também está presente na árvore sintática. Vamos usar o Exemplo (5) para ilustrar esse conceito. A Figura 1 apresenta a árvore sintática para este exemplo.

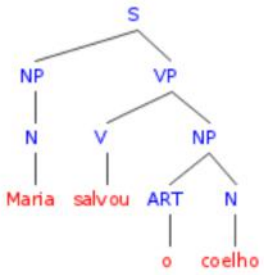

Figura 1: Árvore sintática para a sentença do Exemplo (5): ”Maria salvou o coelho”

Como observado na Figura 1, a árvore sintática começa com a raiz "S"e termina nas folhas que contém as palavras da sentença. No nós intermediários estão presentes as estruturas NP (Noun Phrase) indicando o sujeito que é "Maria"e VP (Verb Phrase) indicando o predicado que é o restante da sentença "salvou o coelho". Cada uma das palavras está marcada com sua classe gramatical, indicada pelas tags "N"(substantivo), "V"(verbo) e "ART"(artigo).

2.4.2 Grafo Semântico

Também chamado grafo de relações ou grafo de relacionamentos, é um grafo direcional que representa os relacionamentos semântico entre conceitos. É uma forma de armazenar conhecimento de entidades do mundo material. Essencialmente,

[9]No contexto do trabalho essa string é uma sentença

[10]Em teoria de linguagens formais é o conjunto de regras de produção de cadeias de uma linguagem formal.

[11]Sintagma é um segmento linguístico da oração

grafos de relacionamento são bases de dados que são estruturados a fim de a se assemelhar com o pensamento humano. Durante o pensamento, a mente humana "navega"através de redes de pensamentos similares, por exemplo, se uma pessoa está pensando sobre o lago Michigan é muito mais provável que ela mude o pensamento para o lago Superior do que para Decartes[12]. Isso porque informações sobre lagos estão muito mais próximas uma da outra do que informação sobre um lago e um filósofo.

Da mesma forma, programas de computadores ou agentes inteligentes podem percorrer rapidamente um grafo semântico para descobrir relacionamentos e conexões entre as entidades em questão. Grafos semânticos são o que está por trás de ferramentas que processam perguntas e respostas, como *Watson*[13] e *Siri*[14], dentre outras.

Os grafos de relações seguem o modelo RDF de triplas, no formato: ***relação(sujeito, objeto)***, em que o sujeito e objeto se tornam vértices do grafo e a relação se torna aresta direcionada ligando os dois. A Figura 2 exemplifica o esquema de uma relação no grafo. E a Figura 3 apresenta o grafo de relacionamento para a sentença do Exemplo (5).

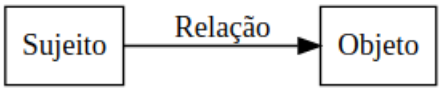

Figura 2: Esquema de um grafo de relações em que o sujeito e objeto são vértices, enquanto a relação é a aresta do grafo.

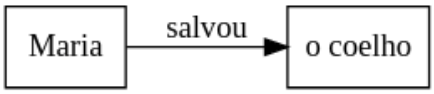

Figura 3: Grafo semântico para a sentença do Exemplo (5): ”Maria salvou o coelho”

[12]Exemplo retirado de `http://davidvandegrift.com/blog?id=62` acessado em 16 de junho de 2018.

[13]`https://www.ibm.com/watson/` acessado em 16 de junho de 2018.

[14]`https://www.apple.com/ios/siri/` acessado em 16 de junho de 2018.

Reificação. Uma reificação acontece num grafo de relacionamento quando um novo nó é ligado por uma nova aresta a uma tripla já existente. A Figura 4 mostra o grafo de relacionamento com uma reificação para a sentença do Exemplo (7). Nela é possível ver que a informação "semana passada"está semanticamente conectada a "Maria salvou o coelho"pela preposição "em_a"(na) dando ideia de tempo, ou seja é a indicação do tempo em que a relação da tripla aconteceu.

(7) Maria salvou o coelho na semana passada.

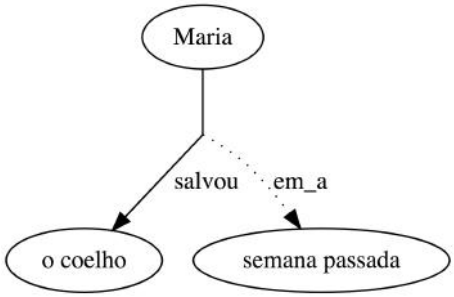

Figura 4: Grafo semântico com reificação para a sentença do Exemplo (7): ”Maria salvou o coelho na semana passada”.

Neste capítulo foi discutida a fundamentação teórica necessária ao entendimento do trabalho. O próximo capitulo vai explanar o que é o Graphia e os trabalhos anteriores sobre ele, incluindo a aplicação para sentenças em português.

3 GRAPHIA

3.1 GRAPHIA - SENTENÇAS EM INGLÊS

O Graphia [4] é um extrator de grafos de relação a partir de estruturas sintáticas para sentenças em inglês. Para fins de clareza, ao longo deste documento vamos se referenciar a esse trabalho como Graphia versão 1 (veja Anexo C). Para gerar um grafo semântico, há basicamente quatro problemas que precisam ser resolvidos: resolução de entidades nomeadas (REN), resolução de correferências, análise sintática e extração de relações. A estrutura do extrator foi feita seguindo uma linha de montagem (pipeline) em que cada um dos problemas acima foi tratado e a saída de cada fase é a entrada da seguinte. O pipeline pode ser sintetizado no diagrama da Figura 5. E a Figura 6 mostra a saída do pipeline do Graphia versão 1 tendo como entrada a sentença do Exemplo (8).

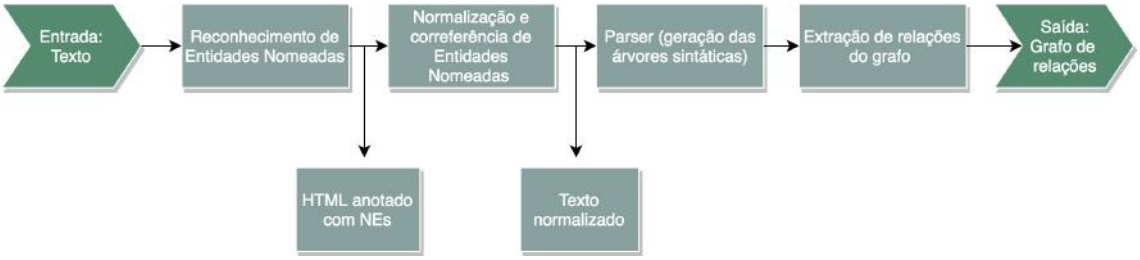

Figura 5: Esquema do pipeline do Graphia

(8) "Donald John Trump (born June 14, 1946) is the 45th and current President of the United States."

Resolução de Entidades Nomeadas

Uma entidade nomeada é um objeto do mundo real, como pessoas, empresas, locais, produtos, etc que pode ser denotada como um nome próprio. Exemplos

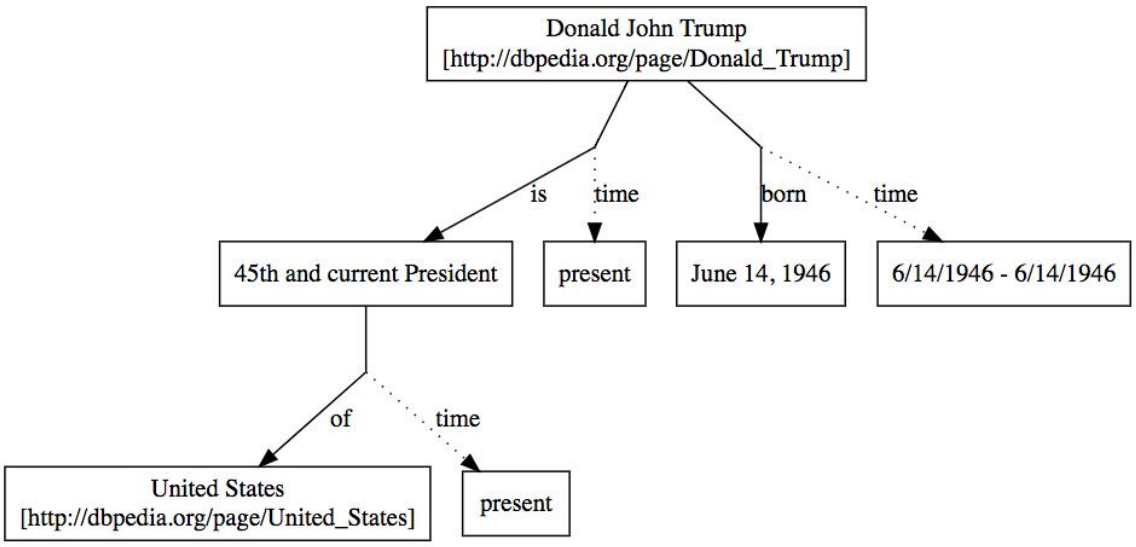

Figura 6: Saída do pipeline do Graphia versão 1 para a sentença do Exemplo (8): "Donald John Trump (born June 14, 1946) is the 45th and current President of the United States."

de entidades nomeadas incluem: Donald Trump (pessoa), Rio de Janeiro (lugar), Amazon (empresa) ou qualquer outra coisa que pode ser nomeada. A REN é o trabalho de identificar, extrair e normalizar entidades nomeadas em textos. Ao fazer a extração de entidades nomeadas, se identifica uma mesma entidade que por ventura aparece de forma diferente no texto. Essa tarefa no Graphia versão 1 foi feita utilizando a ferramenta Spotlight[11]. Nesse componente, a entrada é um texto em linguagem natural e é gerado como saída um arquivo HTML onde as Entidades Nomeadas estão marcadas com links para suas respectivas URI na DBpedia[9].

Resolução de Correferências

No Graphia versão 1, um script foi criado para realizar a resolução de correferências baseado em uma estratégia heurística. A heurística usa uma lista de nomes obtida do censo dos Estados Unidos, onde constam a maior parte dos nomes dos cidadãos americanos e também uma lista de pronomes da língua inglesa, ambos se-

parados por gênero. Essas listas são usadas para identificar os gêneros das Entidades Nomeadas encontradas na fase anterior do pipeline, assim como os pronomes a fim de normalizar as referências.

Análise Sintática

Nesse ponto do pipeline, a análise sintática é obtida fazendo as ações de: *tokenization*, *POS-tagging*, *lemmatization* e *parsing*. Todas essas ações são executadas usando a respectiva ferramenta de Stanford [8], produzindo um arquivo representando as árvores sintáticas das sentenças.

Extração de Relações

Último passo do pipeline é um script que utiliza regras criadas manualmente para extrair relações a partir das estruturas sintáticas. Cada regra contém condições em que será ativada e ações correspondentes caso a ativação ocorra. O script recebe a árvore sintática e através de uma busca em profundidade, percorre todos os nós podendo, em cada nó, ocorrer a aplicação das regras.

No Graphia versão 1 foram criadas 8 regras para a extração de relações. A **regra do início de oração** vai marcar o nível do nó da árvore sintática do início de cada oração. A **regra do sujeito** tem como objetivo encontrar o sujeito da oração e adicioná-lo no grafo de relações para posteriormente relacionar o mesmo com o predicado. A **regra do predicado** tem como objetivo encontrar o predicado da oração. Ao encontrar o predicado, a regra retira o verbo do conteúdo do predicado e o usa como aresta para ligar o sujeito encontrado previamente com o objeto. A **regra do Objeto, núcleo do objeto e predicativo do sujeito** tem como objetivo encontrar o núcleo do predicado, facilitando a ligação entre Entidades Nomeadas no grafo de relações.

A **regra do complemento nominal** serve para encontrar relações de posse e pertinências tanto nos sujeitos quanto nos objetos. A **regra do sintagma preposicionado** tem como objetivo capturar relações de causalidade e de lugar, ela ignora as preposições já encontradas na regra do complemento nominal e foca em achar complementos para os verbos.

A **regra do tempo** foca em dois elementos nas sentenças: datas explícitas (contendo dias, meses ou anos) e verbos. Essa regra produz uma reificação, criando um novo nó e o relacionando com uma tripla de relação já existente. A **regra do fim da oração** serve para delimitar o contexto a fim de relacionar corretamente o sujeito, os predicados e o tempo.

Resultados

CARVALHO[4] conduziu experimentos com o Graphia versão 1 em que foram extraídas 300 sentenças de textos da Wikipedia. Também foram criadas categorias de erros relacionadas tanto às fases do pipeline quanto às regras de extração das relações. Os resultados obtidos estão listados na Tabela 1.

Erro	Total
Sem erros encontrados	14,3%
Resolução de correferência	16,7%
Resolução de entidade nomeada	54,4%
Resolução de referência temporal explícita	11,0%
Construção do sujeito	17,0%
Construção do predicado	13,3%
Construção do objeto	29,6%
Construção de reificação	13,0%
Construção do caminho de triplas	16,7%
Faltando informação	20,0%

Tabela 1: Quantidade em frequência relativa de erros em extrações por categoria de erro. Tabela extraída do trabalho [4]

Os resultados mostram que os erros foram bem distribuídos com exceção da REN e da extração do objeto. O resultado foi considerado positivo dado o pequeno conjunto de regras que o criou (apenas 8), mostrando o potencial que a extração de relações tem usando a abordagem de regras semânticas.

3.2 GRAPHIA - ANÁLISE DE DESEMPENHO E UMA NOVA ALTERNATIVA

Posteriormente, BARBOSA[1] analisou o Graphia versão 1 para melhor entendimento da geração dos erros mais comuns e propôs nova alternativa para o sistema, criando assim o Graphia versão 2. Assim na versão 1, a versão 2 se focou em sentenças em inglês. O trabalho proposto nesse documento se foca em sentenças em português e não tem relação direta com o trabalho [1]. Portanto, apenas uma breve explanação do mesmo vai ser apresentada nessa seção.

Foi objetivo de [1] identificar as características de sentenças que geravam má extrações no Graphia versão 1, e então fazer um filtro para eliminar as sentenças problemáticas. A utilização do módulo de filtro aumentou a precisão da extração do Graphia versão 1, mas a custo da eliminação de maior parte das sentenças de um texto. Entretanto, a ideia da construção do filtro era apenas o passo inicial para entender e conceber uma solução para cada caso bloqueado. Fazendo assim com que o número de tipos de sentenças que o Graphia versão 1 extrai corretamente aumentasse.

Então, uma nova abordagem (Graphia versão 2) foi proposta pelo autor em [1]: criar novas regras de extração de relação não baseadas nas estruturas sintáticas, como o Graphia versão 1, mas baseadas nas dependências gramaticais. A versão 2 também adicionou um novo sistema de REN e um novo mecanismo de identificação de reificações temporais.

Segundo a conclusão do autor, o melhor resultado seria obtido combinando as duas ferramentas: Graphia versão 1 e 2. Se fosse usada uma espécie de filtro, similar ao inicialmente construído, para direcionamento das sentenças com determinadas características para o processamento pelo Graphia versão 1 ou pela versão 2 a precisão de extração do sistema aumentaria. Embora o trabalho [1] melhorou os resultados do Graphia, ele criou um pipeline alternativo cujo esforço para acoplar ao Graphia versão 1 seria muito dispendioso e não faz parte do escopo desse trabalho.

3.3 GRAPHIA - SENTENÇAS EM PORTUGUÊS

Como já dito anteriormente, o projeto Graphia versão 1 foi criado visando sentenças em inglês, porém [2] posteriormente iniciaram o processo de extensão do Graphia para abranger sentenças em português, nascendo assim a chamada versão 3. Como o Graphia versão 1 foi criado com módulos separados, o trabalho em português se ateve em modificar as duas últimas fases do pipeline, deixando as demais inalteradas. A Figura 7 mostra em destaque as fases do pipeline modificadas para a aplicação do Graphia em português (versão 3). E a Figura 8 apresenta o grafo de saída do pipeline para a sentença do Exemplo (9).

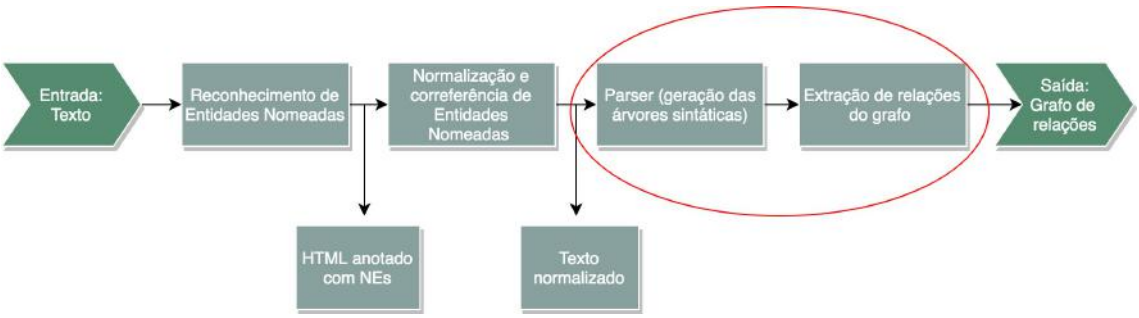

Figura 7: Pipeline do Graphia versão 3 com destaque para as duas últimas fases que foram modificadas para a versão em português

(9) Thiago conhece algumas pessoas na cidade e tem vinte amigos.

O Graphia versão 3 produziu diversos itens, dentre os quais, os utilizados neste trabalho são listados a seguir:

- *POS-Tagger* e *parser*
- Banco de dados de sentenças em português
- Sistema web para avaliação colaborativa de sentenças
- Regras de extração de relacionamentos para sentenças em português

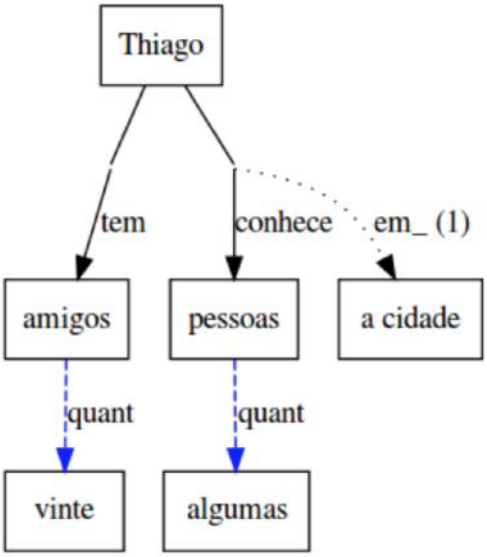

Figura 8: Saída do pipeline do Graphia versão 3 para a sentença do Exemplo (9): "Thiago conhece algumas pessoas na cidade e tem vinte amigos."

Os autores decidiram usar o mWANN-Tagger [3] como *POS-tagger* e o LX-Parser [16] como *parser*. Adicionalmente, um total de 100 mil sentenças com quantidade de caracteres entre 100 a 600 foram obtidas da Wikipédia e adicionadas em um banco de dados. Além disso, criou-se um Sistema Web que permite a avaliação colaborativa tanto das sentenças quanto de suas árvores sintáticas.

Regras de extração de relação

No Graphia versão 3, das 8 regras originais aplicáveis em inglês na versão 1, 3 permaneceram sem alteração, enquanto as outras foram modificadas para a versão 3. As regras inalteradas foram: *início da oração*, *predicado* e *fim da oração*. As regras modificadas são:

- Sujeito
- Objeto, núcleo do objeto e predicativo do sujeito
- Complemento nominal
- Sintagma preposicionado
- Tempo

Detalhes sobre as regras para o português serão discutidos mais adiante juntamente com as modificações feitas para a presente versão.

Avaliação

Os autores [2] sugeriram em seu trabalho a avaliação formal do sistema como um trabalho futuro. Os autores apresentaram alguns casos em que as sentenças não eram boas. Vamos utilizar os Exemplos (10) e (11) para entender alguns erros do Graphia versão 3.

(10) "A porcentagem de população ativa empregada no setor terciário depende bastante do nível de desenvolvimento de cada país."

(11) "A luz nesse comprimento de onda é absorvida pela atmosfera da Terra, então as observações devem ser feitas na atmosfera superior ou no espaço."

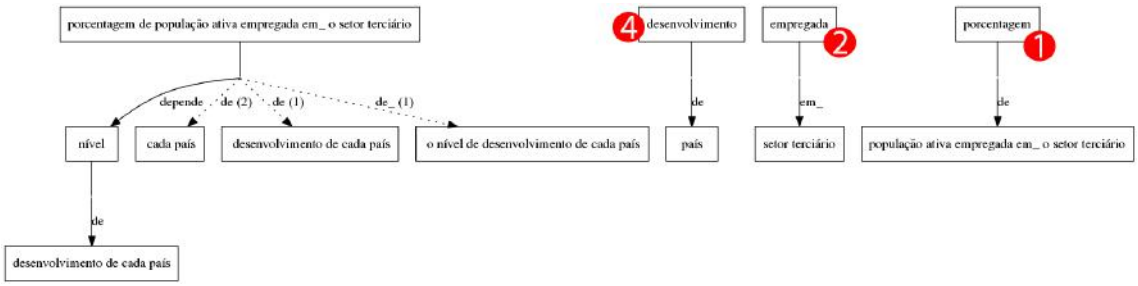

Figura 9: Grafo gerado pelo Graphia versão 3 para a sentença do Exemplo (10): ”A porcentagem de população ativa empregada no setor terciário depende bastante do nível de desenvolvimento de cada país”. Complementos nominais aparecem enumerados.

A Figura 9 apresenta o grafo gerado para a sentença do Exemplo (10). O problema desde grafo ocorre quando o sistema trata o complemento nominal. Nesta sentença, há quatro complementos nominais: 1) "porcentagem de população", 2) "empregada no setor terciário", 3) "nível de desenvolvimento"e 4) "desenvolvimento de cada país". Ao observar a Figura 9, é possível notar que a regra de extração do

complemento nominal embora extraia corretamente os complementos nominais, em três deles (enumerados na Figura 9), não há comunicação apropriada com o restante do grafo, gerando nós com informação redundante e desconexos do restante da sentença.

A Figura 10 apresenta o grafo da sentença do Exemplo (11). Nela há cinco complementos nominais: 1) "luz nesse comprimento", 2) "comprimento de onda", 3) "absorvida pela atmosfera", 4) "atmosfera da Terra"e 5) "feitas na atmosfera superior ou no espaço". O Graphia versão 3 consegue reconhecer apenas três deles (enumerados na Figura 10). Além dos erros de complemento nominal, há erros na extração das reificações, uma série de informações redundantes e o novamente o problema apresentado antes dos complementos nominais não estarem conectados ao grafo.

Com os dois exemplos apresentados, é possível concluir que, embora o trabalho de extensão do Graphia para português tenha dado um importante passo, ele ainda necessita de melhorias a fim de cumprir o papel de extração de relações em sentenças em português. Um exemplo é a melhoria nas regras de extração, como a do complemento nominal. E também uma avaliação formal a respeito do desempenho do Graphia versão 3. Ambos os problemas foram tratados aqui, como veremos em detalhes nos capítulos seguintes.

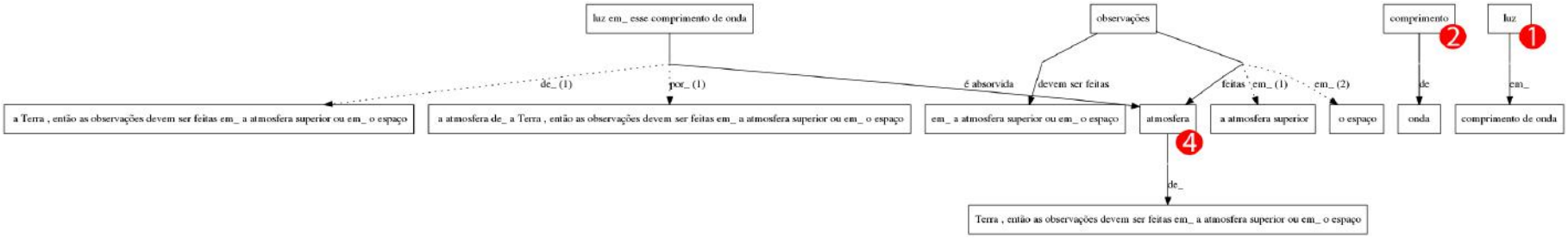

Figura 10: Grafo gerado pelo Graphia português para a sentença do Exemplo (11): ”A luz nesse comprimento de onda é absorvida pela atmosfera da Terra, então as observações devem ser feitas na atmosfera superior ou no espaço”. Os complementos nominais extraídos aparecem enumerados.

4 PROBLEMAS E SOLUÇÕES APRESENTADAS

Este capítulo apresenta uma descrição detalhada das características esperadas para um grafo de relação correto. Também descreve o escopo do trabalho em relação às fases do pipeline, tipos de sentenças, etc. Por fim, são discutidos alguns problemas encontrados no Graphia versão 3[2] juntamente com as abordagens propostas para corrigir tais problemas e exemplos de grafos gerados pelo novo sistema proposto.

4.1 NOVA VERSÃO GRAPHIA PARA SENTENÇAS EM PORTUGUÊS

4.1.1 Grafo de relação para representação de sentenças

Assim como nos trabalhos [4, 1, 2], a estrutura escolhida para representar as relações das sentenças é o grafo de relacionamento (Seção 2.4.2). O grafo gerado deve conectar os atores de uma relação através do tipo de relacionamento existente. Além disso, estruturas sintáticas precisam ser identificadas e tratadas corretamente, como sujeito, verbo, predicado, objeto, predicativo do sujeito e complemento nominal, dentre outros.

Um grafo corretamente gerado é aquele em que o sistema extraiu as estruturas acima listadas e adicionou os nós e as arestas correspondentes a cada uma delas no grafo final sem cometer erros. A Tabela 2 apresenta as estruturas sintáticas correspondentes à sentença do Exemplo (12).

(12) "Os aeroportos da Antártica estão sujeitos a severas limitações por causa das condições climáticas e geográficas"

Estrutura	Parte da sentença
Sujeito	Os aeroportos da Antártica
Predicado	estão sujeitos a severas limitações por causa das condições climáticas e geográficas
Predicativo do sujeito	a severas limitações por causa das condições climáticas e geográficas
Complemento Nominal	"aeroportos da Antártica", "sujeitos a severas limitações", "limitações por causa" e "causa das condições climáticas e geográficas"

Tabela 2: Estruturas sintáticas e onde estão presentes na sentença do Exemplo (12): ”Os aeroportos da Antártica estão sujeitos a severas limitações por causa das condições climáticas e geográficas”

A extração correta das estruturas é fundamental para a geração dos nós e arestas do grafo de relação. Um grafo de relacionamento esperado para a sentença do Exemplo (12) pode ser vista na Figura 11.

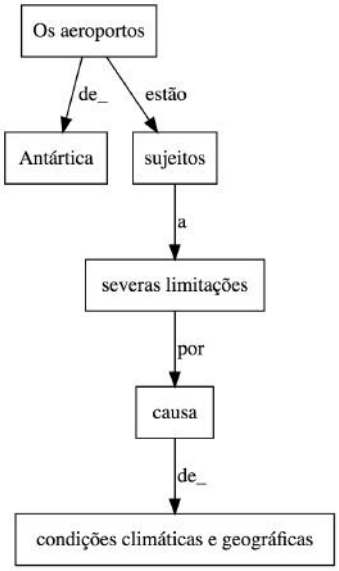

Figura 11: Grafo esperado para a sentença do Exemplo (12): ”Os aeroportos da Antártica estão sujeitos a severas limitações por causa das condições climáticas e geográficas”

No grafo da Figura 11, é possível notar que cada um dos complementos nominais mostrados na Tabela 2 foram corretamente extraídos e a preposição do complemento nominal se torna a aresta que liga o nome sendo complementado ao termo que o complementa. De forma similar acontece com o verbo "estão"que se torna uma aresta ligando o sujeito "Os aeroportos"a parte do predicativo do sujeito "sujeitos".

Seguindo [4], as regras para extração das relações das sentenças serão criadas baseando-se na árvore sintática das sentenças (Seção 2.4.1). É através das árvores sintáticas das sentenças que as regras para extração das relações identificam as estruturas sintáticas e podem ativar as ações, para por fim gerar o grafo de relacionamento.

O escopo do trabalho se limita a sentenças em português presentes no banco de dados criado em [2]. Essas sentenças foram extraídas de diversos artigos da Wikipédia[1], tratadas (eliminando estruturas que geravam erros) e disponibilizadas com as suas árvores sintáticas. Então, o trabalho aqui apresentado (chamado Graphia versão 4) se foca na parte final do pipeline, colaborando com o desenvolvimento das regras de extração das relações, deixando as demais fases do pipeline inalteradas. Na Figura 12 é possível ver destacado a fase do pipeline que é foco do atual trabalho.

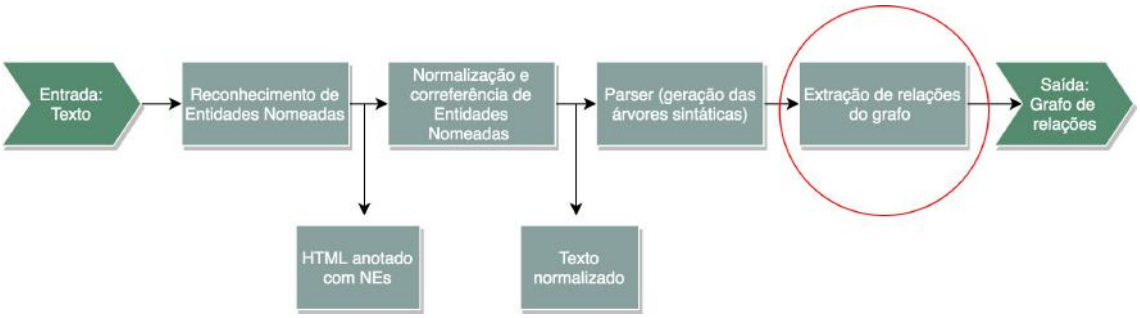

Figura 12: Pipeline do Graphia versão 4 com destaque para a etapa modificada neste trabalho.

[1] https://pt.wikipedia.org

4.1.2 Refatoração do código

A ideia original era consertar o problema da regra do complemento nominal que foi indicada por [2] como trabalho futuro. Entretanto, na tentativa de executar a correção, enfrentou-se um problema com o código de programação do Graphia versão 3. Ele estava altamente acoplado[2] e a coesão[3] era baixa, significando que na tentativa de consertar uma parte do código várias outras deixavam de funcionar. Era inviável acompanhar rigorosamente o que mudanças no código acarretavam nos grafos anteriormente corretos porque o código não tinha cobertura de testes unitários[4].

Assim, o primeiro passo do atual projeto foi uma refatoração completa do código da versão 3, a fim de tornar o sistema altamente modularizado, possibilitando a alteração de partes específicas do Graphia versão 4 sem a necessidade de alterar todo o sistema e facilitando a manutenção do código. Além disso, foi feita a atualização do código para usar Python versão 3 (em vez da original versão 2) porque a versão 2 do Python já tem data marcada para perder o suporte oficial da linguagem [5]. A mudança para Python 3 é mais um passo para facilitar a manutenção do Graphia no futuro.

Uma outra mudança adicionada ao Graphia versão 4 foi a utilização de uma biblioteca em Python para lidar com a geração e visualização dos grafos. Anteriormente, os grafos eram criados usando a linguagem DOT[6] diretamente dentro dos scripts Python. Isso gerava dois complicadores: primeiro a necessidade de que a pessoa a trabalhar com o Graphia soubesse programar na linguagem DOT para criar o grafo, e segundo adicionava uma variável a mais na possibilidade de gerar erro ao escrever

[2]Conceito de Engenharia de Software para expressar que cada parte de um script em linguagem de programação é altamente dependente dos demais, dificultando ou impossibilitando alterar uma parte do sistema modificando apenas uma área específica do código.

[3]Conceito de Engenharia de Software que diz que as classes e/ou funções deveriam ter responsabilidades únicas e realizá-las de maneira satisfatória

[4]Conceito de Engenharia de Software que diz respeito a um dos níveis de teste de Software. Nesse nível são testados componentes individualmente

[5]https://www.python.org/dev/peps/pep-0373/, acessada em 11/12/2017

[6]DOT é uma linguagem de descrição de grafos.

o grafo em DOT. Por essa razão, foi trocado a geração de grafos usando a linguagem DOT diretamente para uma biblioteca em Python que faz essa criação em alto nível. Assim, quem trabalhar com o Graphia não precisa adicionalmente aprender DOT, bastando chamar as funções e classes da biblioteca Graphviz[7] para criar mais facilmente e corretamente os grafos que são o produto final do sistema Graphia. Essa biblioteca gera como saída ambos a imagem do grafo de relacionamento e código do grafo em linguagem DOT.

Como mencionado anteriormente, o fato de o sistema não ter testes unitários, tornava inviável observar o impacto da mudança de uma regra específica nos grafos que antes estavam corretos. Por essa razão, outra adição ao código do Graphia versão 4 foi a implementação de testes unitários automatizados para garantir que alterações no código não impactassem no que já estava corretamente concluído. Por exemplo, ao mudar uma determinada regra, é possível verificar automaticamente se todos os grafos corretos antes da mudança continuam sem erros (veja Anexo D para detalhes sobre testes unitários no Graphia versão 4).

4.1.3 Alteração nas regras de extração de relações

Não apenas o conteúdo das regras de extração foram modificados mas a abordagem de aplicação das mesmas. Antes a abordagem das regras era: se buscava extrair os sujeitos e criava uma pilha de sujeitos, depois se tentava extrair o predicado e pegava o topo da pilha de sujeitos para ligar com o recém predicado extraído. Inicialmente foi seguida a abordagem dos trabalhos anteriores [4, 2] para a aplicação das regras de extração, mas essa abordagem, que funcionou bem para sentenças em inglês, não estava sendo efetiva em inúmeras sentenças em português, especialmente em caso de sentenças mais complexas.

Por essa razão, uma nova abordagem foi pensada sobre como aplicar as regras de extração em que o ponto de partida seja o verbo. Os verbos são a parte mais importante de uma tripla de relacionamento, pois são os verbos que carregam o significado das relações entre as entidades. Então ao invés de procurar sujeitos e

[7] https://pypi.org/project/graphviz/, acessado em 28 de Maio de 2018

predicados separadamente e depois tentar ligá-los, a nova abordagem se concentra em encontrar os verbos e quando um verbo é identificado, comandos são ativados para encontrar quem está praticando a ação do verbo e em direção ao que a ação está sendo aplicada. Como consequência, apenas os nós dos grafos de relacionamento que fazem parte de uma relação são gerados.

Vamos utilizar o Exemplo (13) para ilustrar os pontos sobre a mudança na abordagem e nas regras de extração dos grafos de relacionamento.

(13) "A decepção com o governo Moreira Franco, que não cumpriu a promessa de acabar com a violência em seis meses, levou o eleitorado fluminense a eleger Leonel Brizola novamente, em 1990"

Extração do sujeito da relação. Tanto na abordagem anterior quanto na nova, a árvore sintática das sentenças é percorrida em profundidade e cada nó pode ativar uma determinada regra de extração. Ne versão 3 do Graphia [2], a regra de extração do sujeito era ativada quando a busca em profundidade encontrava um nó do tipo NP filho de S ou SBAR. Na Figura 13 é possível ver o grafo extraído da sentença do Exemplo (13) utilizando a regra do sujeito anteriormente descrita.

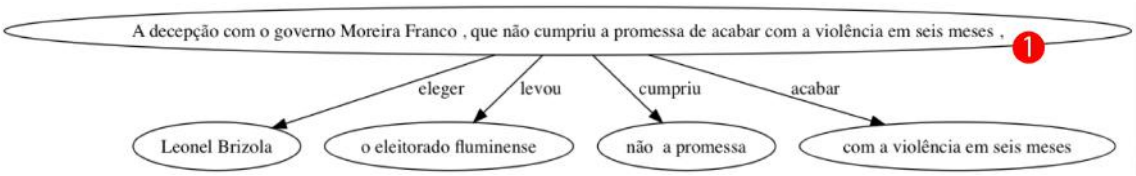

Figura 13: Grafo produzido usando a abordagem inicial de regras de extração para da sentença do Exemplo (13): ”A decepção com o governo Moreira Franco, que não cumpriu a promessa de acabar com a violência em seis meses, levou o eleitorado fluminense a eleger Leonel Brizola novamente, em 1990”

O problema com essa regra é que nem sempre é possível capturar os agentes das triplas de relacionamento com ela. Ao observarmos a árvore sintática do Exemplo

(13) na Figura 15, percebemos que apenas o NP marcado em 1 é capturado como sujeito, pois ele é o único que atente ao critério da regra. Como consequência, as relações não foram extraídas corretamente. O único nó de sujeito é o marcado em 1 na Figura 13, assim o sistema liga todos os verbos a esse sujeito. Mas nem todos os verbos estão expressando ação em relação a todo o conteúdo do nó 1. Portanto, é possível observar que nem sempre um agente da ação é um NP filho de S ou SBAR. Por exemplo "o governo Moreira Franco" é o praticante da ação de não cumprir "a promessa" e ele é justamente um NP que não é filho de S ou SBAR (marcado em 2 na Figura 15).

Pela limitação descrita anteriormente, na nova regra proposta, todos os NP's, independente de serem ou não filhos de S ou SBAR são candidatos a ser agentes de uma relação. A busca em profundidade percorre a árvore sintática e quando encontra um nó do tipo NP, ele é guardado numa pilha de NP's juntamente com a profundidade do nó. Quando a busca em profundidade encontra um nó do tipo V (representando um verbo), a pilha de NP's é percorrida, desempilhando os NP's, até encontrar o primeiro NP a ter profundidade igual ou menor que a do nó V. Quando isso ocorre, é esse o NP a ser o agente da relação com o verbo em questão. O conteúdo após o verbo é então extraído para completar a formação dessa relação verbo(sujeito, predicado). A Figura 14 apresenta um esquema de como é formado uma tripla de relacionamento básica na nova abordagem proposta.

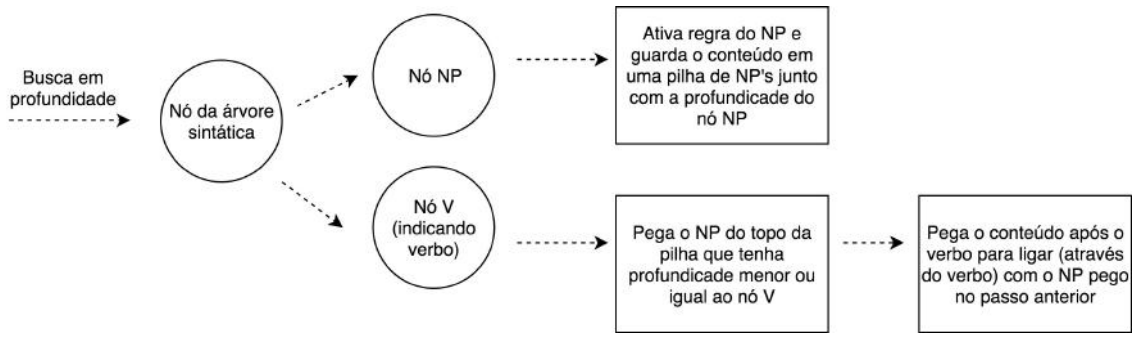

Figura 14: Esquema apresentando a proposta para a construção básica de relacionamento no grafo semântico.

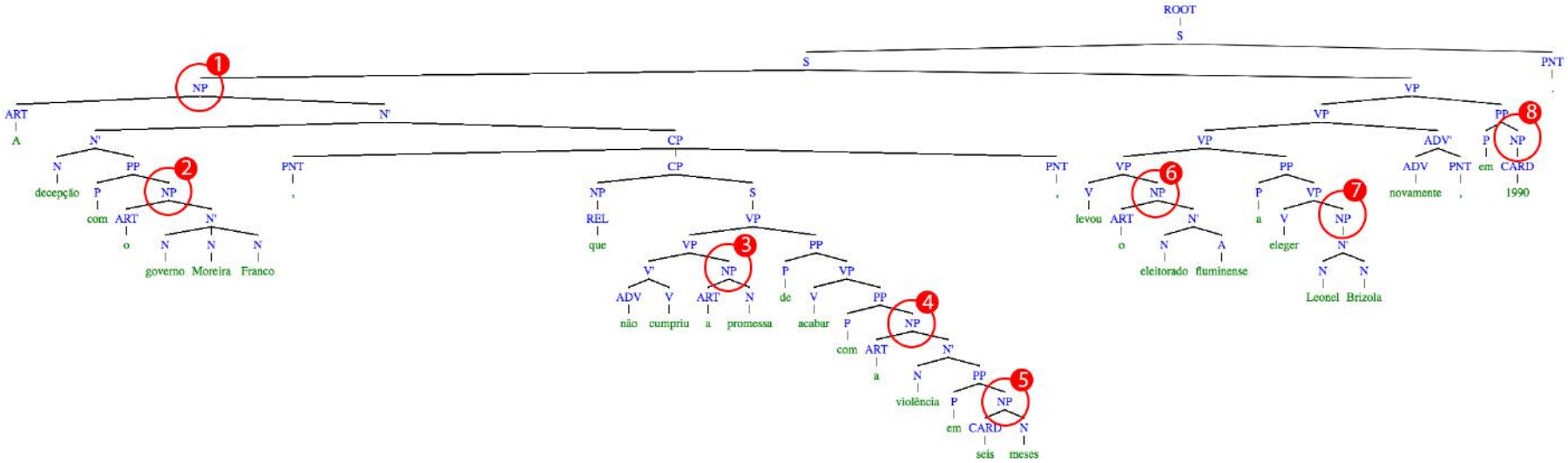

Figura 15: Árvore sintática da sentença do Exemplo (13): "A decepção com o governo Moreira Franco, que não cumpriu a promessa de acabar com a violência em seis meses, levou o eleitorado fluminense a eleger Leonel Brizola novamente, em 1990".

Duas são as regras responsáveis pela execução do apresentado na Figura 14: regra do NP e regra do predicado. **A regra do NP**, como já falado, é ativada quando a busca encontra um nó do tipo NP. Ela então pega o conteúdo abaixo do nó NP e armazena numa pilha de NP's juntamente com a profundidade do nó. A Figura 15 mostra os nós que fazem a regra NP ser ativada (marcados de 1-8).

A escolha para usar apenas NP com profundidade igual ou menor que a profundidade do nó de verbo foi baseada em observações e experimentos justificados pela estrutura da árvore sintática. Como as regras de derivação são hierárquicas, sabemos que na árvore sintática, o VP (constituinte que contém pelo menos um V) deve estar na mesma (ou quase na mesma) profundidade que o NP a ele relacionado. Assim, a profundidade de um nó de verbo é próxima da profundidade de nó NP que é agente da relação daquele verbo. Quando a profundidade desses nós é distante, é muito provável que o nó NP em questão esteja relacionado a um outro elemento da árvore sintática que está em uma profundidade mais próxima a ele, significando que ele não é o NP desejado para se conectar semanticamente com o verbo.

Extração do verbo, locução verbal e predicado da relação Aqui é onde o verbo ou locução verbal é extraída da árvore sintática da sentença. Essa regra não foi modificada em como ela é disparada, quando a busca em profundidade encontra um nó do tipo V, indicando que há um verbo. Após encontrar um verbo, o script continua a busca para verificar se há um outro nó do tipo V logo em seguida. Se esse for o caso, é identificada uma locução verbal, então a relação será construída com a locução verbal e não apenas com o primeiro verbo. A mesma regra se aplica quando a locução verbal é formada por mais que dois verbos.

A regra do predicado além de reconhecer o verbo e a locução verbal, também extrai o predicado da relação e por fim desempilha a pilha de NP's para encontrar o primeiro NP cuja profundidade seja menor ou igual ao do nó verbo. Assim, tendo o NP que é o agente da relação, o verbo e o predicado é possível construir uma relação do grafo.

Extração do complemento nominal A melhoria da extração do complemento nominal era o objetivo inicial do trabalho. Em [2] foi apontado que o complemente

nominal estava sendo extraído da sentença, porém o sistema era incapaz de conectar as relações extraídas com o restante do grafo de relacionamento. Então a regra do complemento nominal não foi modificada de como ela era ativada: quando encontra um nó do tipo N' que é pai dos nós N e PP. O empenho foi aplicado em fazer o necessário com o código do sistema para conectar os complementos nominais com os nós do grafo gerado, o que incluiu eliminar as partes dos textos correspondente ao complemento nominal dos nós originais e criar novos nós e arestas para os complementos nominais extraídos.

Para tal foi necessário separar a regra do complemento nominal em duas partes, uma a ser aplicada se o complemento nominal está nos candidatos a sujeito e outra a ser aplicada se o complemento nominal está no predicado. Ambas as regras do complemento nominal só são aplicadas após os candidatos a sujeito e o predicado serem separados em um sentença. Isso foi feito para facilitar a criação dos nós. Pois, primeiramente os nós de sujeito e predicado são criados e ligados pela aresta do verbo. Então as regras de complemento nominal, se ativadas, podem percorrer os nós de sujeito e de predicado e alterar o conteúdo desses nós para criar novos nós representando o complemento nominal. Então, durante a busca em profundidade, ao descobrir um nó que ativa a regra do complemento nominal no predicado, o script percorre todos os vértices que são do predicado no grafo e ao encontrar o conteúdo do complemento nominal extraído, os seguintes passos são executados:

- Apaga o texto do complemento nominal de dentro do nó
- Cria uma nova aresta com a preposição do complemento nominal
- Cria um novo nó com o complemento, ligando-o ao nó anterior de predicado através da aresta de preposição

Os mesmos passos são completados para complementos nominais encontrados dentro de sujeitos. Como consequência, o grafo gerado não possui informação redundante e o conteúdo dos complementos nominais é parte do grafo de relacionamento.

Regras eliminadas Por causa da nova abordagem algumas regras anteriores se tornaram dispensáveis. A regra de início e fim de oração são exemplos desse caso.

Essas regras serviam para marcar a profundidade do início e do fim das sentenças, a fim de ligar corretamente os predicados com os sujeitos das sentenças corretamente. Elas não são mais necessárias porque a nova abordagem usa a profundidade dos nós NP's e V's para saber quando conectar um V com um NP.

Regras não implementadas Algumas regras da versão 3 não foram implementadas no novo sistema proposto devido à restrição de tempo. Elas ficam como sugestão para trabalhos futuros. As regras a serem adicionadas no sistema proposto são a regra de quantidade e regra de tempo.

4.1.4 Aplicando a nova versão para extrair grafos de relacionamento

A Figura 16 mostra o grafo gerado para a sentença do Exemplo (13) usando as novas regras e a abordagem propostas. Na Figura 17 é apresentado o grafo de relações gerado para a sentença do Exemplo (10) também usando as novas regras. E por fim, a Figura 18 traz o grafo gerado da sentença do Exemplo (11). Podemos verificar que os três estão corretos semanticamente, diferente dos anteriormente gerados (Figuras 13, 9 e 10, respectivamente).

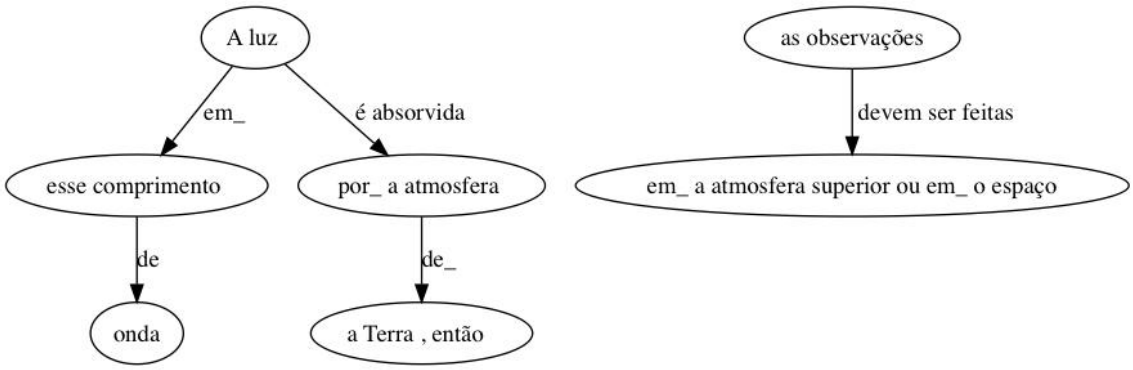

Figura 18: Grafo extraído utilizando as novas regras propostas para a sentença do Exemplo (11): "A luz nesse comprimento de onda é absorvida pela atmosfera da Terra, então as observações devem ser feitas na atmosfera superior ou no espaço".

Após as alterações das regras, a geração dos grafos é feita com basicamente três

regras: a do sujeito, do predicado e do complemento nominal (aplicada ao sujeito e ao predicado separadamente). Cinco sentenças foram usadas como base para a mudança das regras apresentadas nesse capítulo. Em 100% delas, as alterações criadas trouxeram melhorias para os grafos extraídos em comparação com o sistema [2]. Entretanto, ainda se fazia necessário avaliar formalmente o novo sistema, então o capítulo seguinte apresenta o método de avaliação e os resultados obtidos ao se comparar os dois sistemas Graphia para sentenças em português.

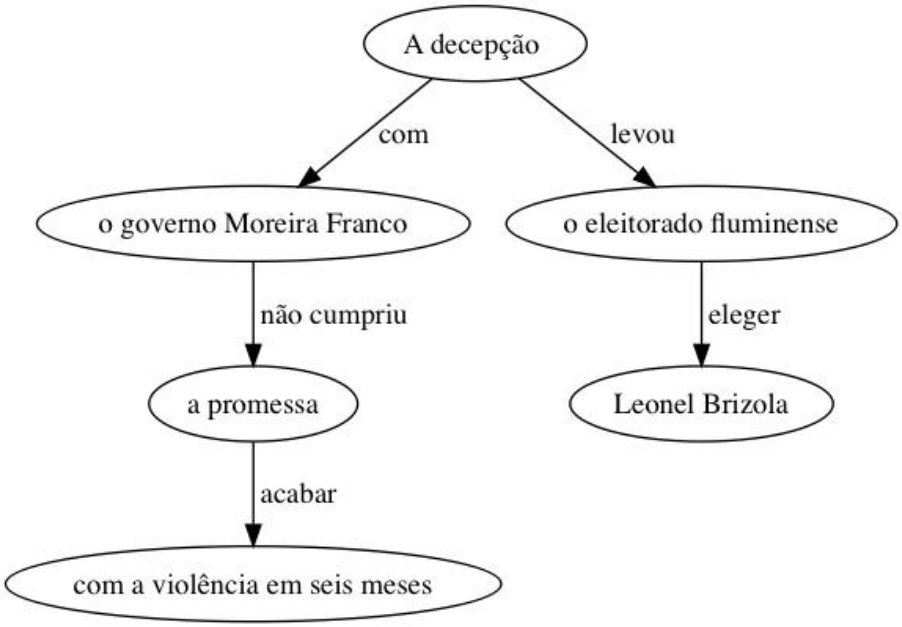

Figura 16: Grafo extraído utilizando as novas regras propostas para a sentença do Exemplo (13): "A decepção com o governo Moreira Franco, que não cumpriu a promessa de acabar com a violência em seis meses, levou o eleitorado fluminense a eleger Leonel Brizola novamente, em 1990".

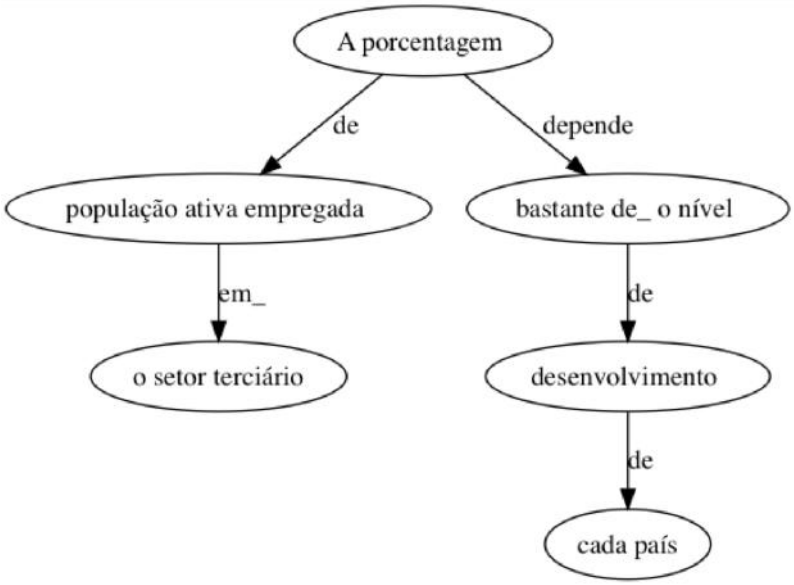

Figura 17: Grafo extraído utilizando as novas regras propostas para a sentença do Exemplo (10): "A porcentagem de população ativa empregada no setor terciário depende bastante do nível de desenvolvimento de cada país".

5 AVALIAÇÃO E RESULTADOS

5.1 MÉTODO DE AVALIAÇÃO

Uma avaliação foi conduzida a fim de obter a performance do sistema proposto em comparação com o já existente para sentenças em português (Graphia versão 3). Foram sorteadas 40 sentenças do banco de dados criado em [2]. Esse banco de dados contém um total de 100 mil sentenças com quantidade de caracteres entre 100 a 600 que foram obtidas da Wikipédia. Após o sorteio, o grafo de relacionamento de cada sentença foi extraído usando tanto o Graphia versão 3 quanto o Graphia versão 4 a fim de comparar ambas as versões.

As sentenças foram sorteadas pelos seus IDs no banco de dados utilizando a ferramenta online Randon.org[1]. A aleatoriedade dessa ferramenta vem de ruído atmosférico que para muitos propósito é melhor do que algoritmos de números pseudo-aleatórios que são geralmente usados em programas de computadores.

5.1.1 Categorias de avaliação

Após a extração dos grafos, eles são analisados e julgados como correto ou incorreto usando 5 diferentes categorias. As categorias são tratadas independentemente, ou seja, um grafo pode está correto segundo a categoria 1 e 2, mas está incorreto segundo a categoria 3. As categorias foram escolhidas baseadas nas regras de extração dos grafos, pois desse forma é possível gerar um melhor entendimento sobre o desempenho de cada regra em específico.

Sujeito simples. A categoria sujeito simples visa identificar se o grafo extraiu corretamente o sujeito simples da sentença. Estar corretamente extraído significa ter um nó no grafo de relações para o sujeito simples. A Figura 19 mostra o grafo extraído para a sentença do Exemplo (14). Nele é possível observar um nó (indicado pelo número 1) com o sujeito simples "*A densidade populacional*", significando que ele foi corretamente extraído. Esse grafo seria julgado como correto na categoria

[1] www.random.org Acessado em 4 de julho de 2018

sujeito simples.

(14) "A densidade populacional é de 15 pessoas por quilômetro quadrado de área de terra"

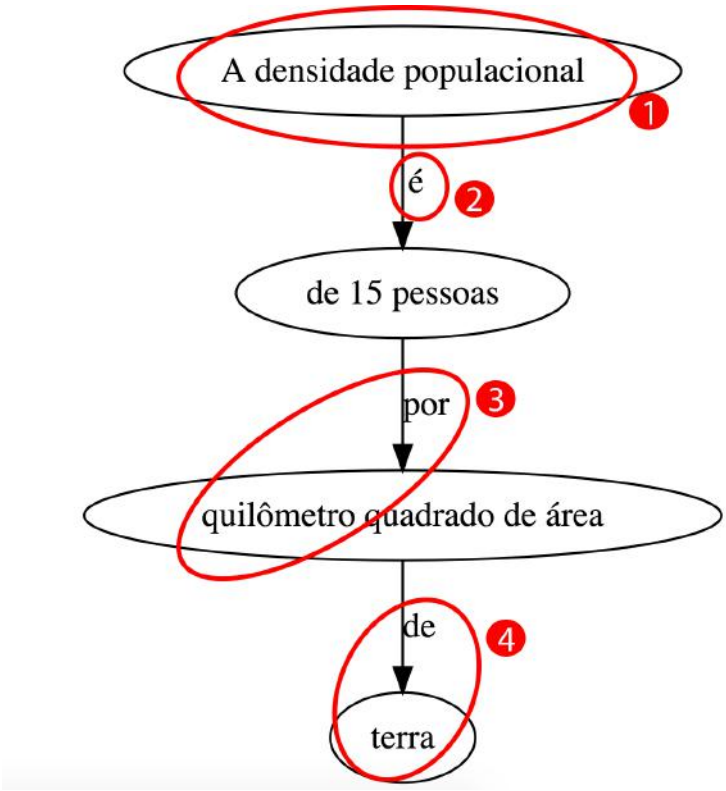

Figura 19: Grafo gerado para a sentença do Exemplo (14): ”A densidade populacional é de 15 pessoas por quilômetro quadrado de área de terra”. O sujeito simples (1), verbo (2) e complementos nominais (3) e (4) foram extraídos corretamente.

Sujeito composto. Similar ao critério de avaliação do sujeito simples, mas aplicado a sujeitos compostos. A Figura 20 mostra o grafo extraído para a sentença do Exemplo (15), em que é possível observar um nó com o sujeito composto "*Poisson, Liouville, Fourier e outros*", ou seja, ele foi corretamente extraído. Esse grafo seria julgado como correto na categoria sujeito composto. Uma outra possibilidade de o grafo estar correto na categoria sujeito composto seria se houvesse um nó para cada um dos componentes do sujeito composto e todos ligados pelo mesmo verbo a eles relacionado.

(15) Poisson, Liouville, Fourier e outros mais estudaram as equações em derivadas parciais e a análise harmônica

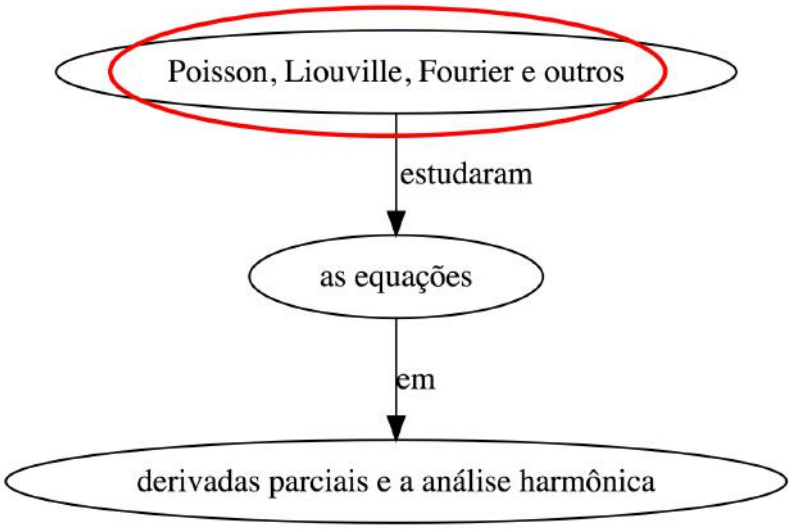

Figura 20: Grafo gerado da sentença do Exemplo (15): "Poisson, Liouville, Fourier e outros mais estudaram as equações em derivadas parciais e a análise harmônica". O sujeito composto "Poisson, Liouville, Fourier e outros" foi corretamente extraído (marcado em vermelho).

Verbo. A categoria verbo visa identificar se o grafo extraiu corretamente o verbo (ou locução verbal) da sentença. Estar corretamente extraído significa ter uma aresta no grafo de relações para o verbo. A Figura 19 mostra uma aresta do grafo (marcada com o número 2) com o verbo "*é*", significando que ele foi corretamente extraído. Esse grafo seria julgado como correto na categoria *Verbo*

Complemento nominal. Na categoria complemento nominal identificamos se o grafo extraiu corretamente complemento nominal da sentença. Estar corretamente extraído significa ter uma aresta para a preposição e um nó para o complemento nominal conectados com o restante do grafo de relações. Ter extraído o complemento nominal mas está desconectado do restante da sentença é considerado erro. A Figura 19 mostra dois complementos nominais corretamente extraídos (indicados pelos números 3 e 4). Esse grafo seria julgado como correto na categoria comple-

mento nominal.

Informação redundante. Na categoria de informação redundante verificamos se o grafo gerado contém informações redundantes. Essa categoria de avaliação foi criada pois no trabalho Graphia em português proposto em [2], a geração de informação redundante foi apontada como um dos problemas de vários grafos gerados. Para um grafo ser julgado como correto nessa categoria, as informações da sentença devem aparecer apenas uma vez (ou o número necessário para estar correto semanticamente) no grafo de relação, sejam elas nós ou arestas.

Ao comparar os grafos das Figuras 21 e 17 (Seção 4.1.4) percebemos claramente que na Figura 21 temos informações redundantes (marcadas em vermelho), enquanto que na figura 17 não há nenhuma informação redundante. O grafo da Figura 21 seria então classificado como incorreto na categoria informação redundante, enquanto que o grafo da Figura 17 receberia a classificação de correto na mesma categoria.

5.1.2 Seleção de amostra de sentenças

Foram selecionadas 40 sentenças para a avaliação através de sorteio. Um problema enfrentado na execução dessa etapa foi a alta quantidade de árvores sintáticas incorretas, especialmente na questão de classificar erroneamente um adjetivo como verbo (mais comum quando há verbo de ligação e predicativo do sujeito). Como as regras criadas para a extração do grafo de relacionamento partem do princípio que se tem árvores sintáticas corretas, uma árvore sintática incorreta não é válida para avaliar o sistema proposto. Consequentemente, a avaliação de uma sentença passou por duas etapas em sequência. Primeiramente, ao sortear uma sentença, uma análise qualitativa foi efetuada a fim de julgar se a árvore sintática da sentença apresentava algum erro. Sentenças com árvore sintática erradas foram eliminadas da avaliação. Um fato relevante adicional em relação à avaliação é que grande parte das sentenças sorteadas que continham mais de 300 caracteres possuíam a árvore sintática errada. Portanto, a maioria das sentenças avaliadas tem menos de 300 caracteres.

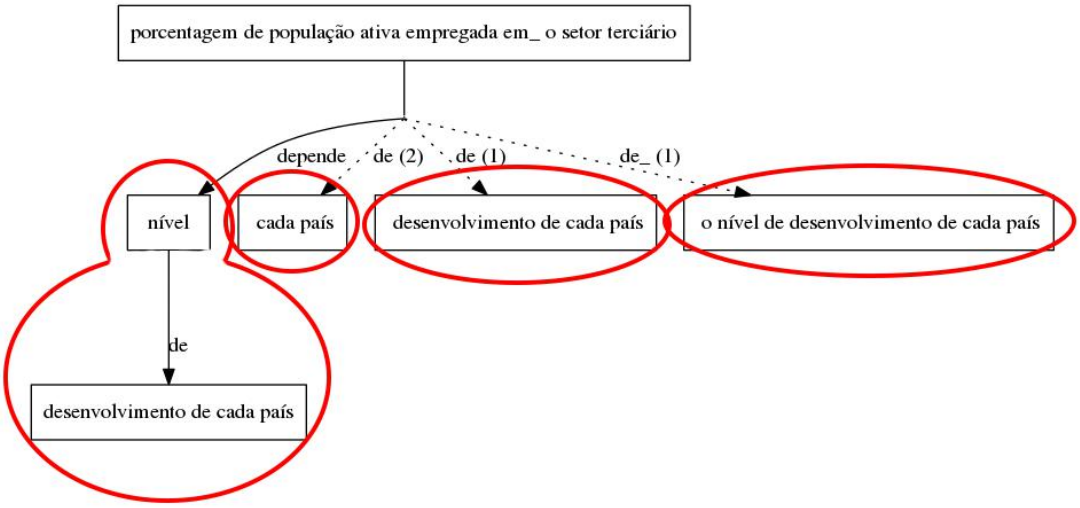

Figura 21: Grafo extraído da sentença do Exemplo (10): "A porcentagem de população ativa empregada no setor terciário depende bastante do nível de desenvolvimento de cada país". Ele é avaliado como incorreto segundo a categoria informação redundante. É possível observar as informações redundantes marcadas em vermelho.

Em seguida, uma avaliação quantitativa é aplicada. O grafo gerado foi julgado como correto ou incorreto em cada uma das categorias apresentadas na seção anterior. Um mesmo grafo pode apresentar várias instâncias da categoria a ser avaliada. O Exemplo (16) ilustra esse ponto. Ele tem 4 instâncias de verbo (destacadas na sentença), então ao julgá-lo na categoria verbo, cada uma das instâncias de verbo necessita ser julgada como correta ou incorretamente extraída no grafo de relacionamento.

(16) "O pampa **é** uma das regiões mais produtivas e férteis para a agricultura na Terra, no entanto, este fator também **é** responsável por **dizimar** grande parte do ecossistema original, para **abrir** caminho para a agricultura comercial"

5.1.3 Resultados e discussão

A Tabela 3 apresenta os resultados da avaliação dos grafos das 40 sentenças selecionadas, comparando o desempenho entre as duas versões do Graphia para português. Nela é possível ver o número de instâncias, assim como os valores em porcentagem de acerto para cada categoria avaliada.

	Número de instâncias	Graphia versão 3	Graphia versão 4
Sujeito Simples	60	75%	**96,66%**
Sujeito Composto	5	**80%**	60%
Verbo	87	59,77%	**83,90%**
Complemento Nominal	64	26,56%	**82,81%**
Informação redundante	40	17,5%	**82,5%**

Tabela 3: Resultados da comparação de desempenho entre os sistemas Graphia versão 3 e versão 4 em porcentagem de acerto para cada categoria avaliada.

Para todas as categorias, exceto sujeito composto, a porcentagem de acerto do Graphia versão 4 foi melhor que a versão 3. Como as sentenças foram sorteadas, apenas 5 delas apresentavam sujeito composto. Em 1 dessas sentenças, o trabalho versão 3 acertou enquanto o sistema proposto errou. O erro aconteceu por causa do critério de conectar o verbo com o NP (Seção 4.1.3). Como o NP escolhido para agente da ação é o primeiro encontrado que tem a profundidade igual ou menor que a do nó V, o Graphia versão 4 só seleciona um único NP, que neste caso é apenas um dos componentes do sujeito composto, enquanto que os os outros NP's contituintes do sujeito composto não são capturados. Como na versão anterior o agente da ação é sempre um NP filho de S ou SBAR, ele pega o NP pai dos NP's de sujeito composto. Assim, nesses casos a versão 3 consegue extrair o sujeito composto, mas a versão 4 não consegue. Veja o anexo B para visualizar um exemplo.

Esse problema pode ser corrigido ao se adicionar um critério a mais na seleção de NP's para se conectarem aos verbos. Se ao selecionar o NP, continuar-se a percorrer

a pilha de NP's e se observar que os NP's seguintes tem a mesma profundidade do NP primeiramente selecionado, isso seria um indício de se tratar de sujeito composto e então o verbo se ligaria a cada um desses NP's de mesma profundidade, ao invés de se ligar apenas com o primeiro deles. Essa modificação não foi implementada devido a limitação de tempo, mas fica como sugestão de trabalhos futuros a implementação dessa mudança e avaliação do impacto dela no sistema proposto.

Os grafos extraídos apresentaram uma grande melhora nas categorias complemento nominal e informação redundante, dois dos problemas listados do Graphia versão 3. Na categoria sujeito, algumas instâncias não foram corretamente extraídas em ambas as versões do Graphia porque eram casos de sujeito oculto (sujeito implícito) ou indeterminado. E nenhuma das duas versões do Graphia em português foi desenvolvida pensando nesses casos em específico. É importante destacar que esses tipos de sujeito não existem na língua inglesa, não tendo sido então um problema a atacar nas versões 1 e 2 do Graphia.

Acredita-se que a grande melhora apresentada pelo trabalho proposto se dê por três razões principais: a mudança da abordagem ao aplicar as regras, a mudança de como se seleciona o agente da relação e a correção da regra do complemento nominal. A mudança da abordagem e a correção do complemento nominal foram majoritariamente responsáveis por gerar grafos mais limpos, sem informações irrelevantes. Pois essas duas regras trabalharam em extrair as partes da sentença relevante aos nós e aresta em questão, deixando de fora partes da sentença não relevante ao grafo final gerado. A mudança no critério de seleção do agente da relação foi a grande responsável por gerar um número maior de triplas (sujeito, relação e objeto) corretas nos grafos. Isso porque capturou-se com maior precisão os agentes em relação aos verbos extraídos.

Como mencionado anteriormente, foi parte da avaliação dos grafos também avaliar as árvores sintáticas das sentenças. A quantidade de árvores sintática erradas encontradas foi bastante grande. Em média foi necessário sortear cerca de 5 sentenças até obter uma com a árvore sintática correta possibilitando a avaliação do grafo. Esse fato é um forte indicativo de que o parser usado para extrair as árvores

sintáticas das sentenças não está adequado.

6 CONCLUSÃO E TRABALHOS FUTUROS

Esse trabalho produziu uma nova versão do Graphia em português com as seguintes características:

- Código refatorado, versão do Python atualizada para 3, utilização de testes unitários automatizados.
- Mudança de abordagem ao gerar as regras de extração de triplas de relacionamento, tornando o verbo a parte mais importante, como consequência, gerando melhora na extração do verbo nas relações.
- Melhora em corrigir dois erros apontados no Graphia versão 3: a extração do complemento nominal e a geração de grafos com informação redundante.
- Avaliação do sistema proposto em comparação com o existente anteriormente [2], apresentando uma melhora geral nas categorias avaliadas.
- Contribuição na avaliação das árvores sintáticas usando o sistema web produzido em [2], rendendo forte evidência da necessidade de obter mais árvores sintáticas corretas para as sentenças.

O resultado da avaliação mostra que as alterações efetuadas no Graphia para sentenças em português são uma estratégia promissora, pois trouxeram uma melhora significativa em relação ao sistema anterior, mesmo usando apenas 3 regras para a produção de grafos de relacionamento. É portanto possível concluir que os objetivos de criar uma versão melhorada para o Graphia versão 3 e prover uma avaliação para o mesmo foram alcançados com sucesso.

Deixa-se em aberto possibilidades futuras que podem melhorar ainda mais o Graphia. Uma sugestão seria estender a versão 4 para adicionar outros elementos do pipeline a fim de trazer melhorias para a extração dos grafos. Por exemplo, incluindo as duas fases a seguir:

- **Resolução de correferência dos textos antes de separá-los em sentenças.** As sentenças em português estão independentes umas das outras

no banco de dados [2], significando que há casos em que não se sabe qual o sujeito (Entidade Nomeada) praticante da ação, pois na sentença há apenas uma referência à entidade (por exemplo, o pronome "ele") e tendo apenas a sentença separadamente, não é possível saber quem é o "ele" presente, se tornando impossível fazer a Resolução de Entidades Nomeadas. Normalizar o texto completamente, trocando pronomes, etc pela NE referenciada, permite separar as sentenças posteriormente e armazená-las no banco de dados sem perder informação sobre os agentes das triplas de relação.

- **Resolução de Entidades Nomeadas.** Encontrar as NE das sentenças e adicionar no grafo de relacionamento a URI única da NE, semelhante como é no Graphia versão 1. Fazer isso permitiria que agentes automáticos navegassem pelo grafo de relacionamento gerado pelo Graphia buscando todas as informações reacionadas a uma determinada NE.

A melhora da extração do sujeito composto também pode ser explorada no futuro, seguindo as sugestões de melhora explicadas na Seção 5.1.3. Adicionalmente, criar uma solução de extração para outros tipos de sujeito que existem na língua portuguesa, além do simples e composto:

- **Sujeito oculto.** Nessa caso seria necessário uma forma de extrair a pessoa a quem o verbo se refere da conjugação do mesmo. Teria que ser feito provavelmente antes da etapa de normalização, assim durante a normalização, a pessoa do pronome a quem o verbo menciona poderia ser explicitada e posteriormente identificada como uma entidade nomeada.

- **Sujeito indeterminado.** Esse caso é mais delicado e exige um estudo detalhado, pois esse tipo de sujeito se refere a elementos que não se pode (ou não se quer) identificar, por exemplo na sentença "Acredita-se na existência de vida fora do planeta Terra." não se sabe quem acredita na existência de vida fora do planeta Terra.

Outra sugestão para trabalho futuro para o Graphia em português é testar outras possibilidades levando em questão a profundidade de NP's e V's como critério de

seleção de NP's que se conectam com os V's para gerar a relação semântica correta. Nesse trabalho foi usados os NP's cuja profundidade é menor ou igual que a profundidade de V. Mas pode ser promissora a ideia de trabalhar com diferenças absolutas em vez de números. Por exemplo, selecionar o NP cuja profundidade tenha uma diferença absoluta de no máximo 3 com a profundidade de V.

Mais uma sugestão para trabalhos posteriores é usar as saídas do grafia versão 4 para gerar grafos conectados do conteúdo total de um texto, de forma semelhante ao criado no Graphia versão 1. Os grafos continuariam a ser extraídos para cada sentença, mas um passo a mais poderia ser implementado para conectar os grafos das sentenças de um mesmo texto. Isso pode ser feito sem grandes esforço empregado ao usar a característica da linguagem DOT de automaticamente conectar nós e arestas que tem o mesmo conteúdo.

Também fica como proposta de ação futura a criação de um script que usa a representação interna do grafo no Graphia versão 4 para extrair o conteúdo do mesmo para o formato RDF. Essa tarefa já foi aplicada ao Graphia versão 1, então o script para o Graphia versão 4 poderia se basear nele. Outro ponto interessante seria criar uma solução para que a avaliação fosse executada de forma automática, para isso, poderia-se usar como ideia a forma em que os testes unitários são feitos e executados.

E por fim, a necessidade de sortear cerca de 5 sentenças para obter uma árvore sintática correta aponta que a qualidade das árvores sintáticas das sentenças não está satisfatória. Pode haver a necessidade de trocar o *parser* para aumentar a quantidade de árvores sintáticas corretas. Um passo menos dispendioso (pois já existe arsenal para tal) seria trocar o *POS-tagger*, pois a maioria dos erros observados nas árvores sintáticas eram de *POS-tags*. Uma hipótese a ser testada seria se ao fazer a normalização e a resolução das entidades nomeadas do texto antes de separar em sentenças faria o parser atual produzir árvores sintáticas de maior qualidade, pois as relações nas sentenças estariam mais explícitas.

REFERÊNCIAS

[1] BARBOSA, R. B. **Extração de grafos de relacionamento semântico a partir de estruturas sintáticas**: uma nova iteração. 2014. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação)-Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

[2] CARDOZO, A. F. P.; TEIXEIRA, K. T. **Extração de relações em sentenças em português**. 2016. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação)-Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016.

[3] CARNEIRO, H. C.; FRANÇA, F. M.; LIMA, P. M. Multilingual part-ofspeech tagging with weightless neural networks. **Neural networks**, New York, v. 66, p. 11-21, 2015.

[4] CARVALHO, D. S. **Extração de grafos de relacionamento semântico a partir de estruturas sintáticas uma abordagem de melhor esforço**. 2011. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação)-Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2011.

[5] CHOWDHURY, G. G. Natural language processing. **Annual Review of Information Science and Technology**, White Plains, v. 37, n. 1, p. 51-89, 2003.

[6] FADER, A.; SODERLAND, S.; ETZIONI, O. Identifying relations for open information extraction. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2011, Edinburgh. **Proceedings**... [S.l.: s.n.], 2011. p. 1535-1545.

[7] GAMALLO, P.; GARCIA, M. Multilingual open information extraction. In: PEREIRA, F. et al. (Ed.). **Progress in artificial intelligence**: 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings. New York: Springer, 2015. p. 711- 722. (Lecture Notes in Computer Science, 9273).

[8] KLEIN, D.; MANNING, C. D. Accurate unlexicalized parsing. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 41., 2003, Sapporo. **Proceedings**... New York: ACM, 2003. p. 423-430.

[9] BIZER, C. et al. DBpedia: a crystallization point for the web of data. **Journal of Web Semantics**, [S.l.], v. 7, n. 3, p. 154-165, 2009.

[10] SENA, C. F. L.; GLAUBER, R.; CLARO, D. B. Inference approach to enhance a portuguese open information extraction. In: INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS, 19., 2017, Porto. **Proceedings**... [S.l.: s.n.], 2017. p. 442-451.

[11] MENDES, P. N. et al. Dbpedia spotlight: shedding light on the web of documents. In: INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 7., 2011, Graz. **Proceedings**... [S.l.: s.n.], 2011. p. 1-8.

[12] MESQUITA, R. M. **Gramática da língua portuguesa**. São Paulo: Saraiva, 2014.

[13] MITKOV, R. (Ed.). **The Oxford Handbook of Computational Linguistics**. Oxford: Oxford University Press, 2004.

[14] SACCONI, L. A. **Novíssima Gramática Ilustrada Sacconi**. 24. ed. São Paulo: Nova Geração, 2011.

[15] SAFWAT, H.; DAVIS, B. **Knowledge extraction from unstructured text using embedded controlled language**. 2015. Disponivel em: <https://www.insight-centre.org/sites/default/files/basic_pages_file/knowledge_extraction_from_unstructured_text_using_embedded_controlled_language_hazem_safwat_abdelaal.pdf>. Acesso em: 13 fev. 2018.

[16] SILVA, J. et al. Out-of-the-box robust parsing of Portuguese. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 9., 2010, Porto Alegre. **Proceedings**... New York: ACM, 2010. p. 75-85.

[17] WU, F.; WELD, D. S. Open information extraction using Wikipedia. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 48., 2010, Uppsala. **Proceeding**... [S.l.: s.n.], 2010. p. 118-127.

ANEXOS

ANEXO A – Tabelas de avaliação das 40 sentenças selecionadas

Esse anexo apresenta detalhes de todas as sentenças selecionadas para avaliação, mostrando o número de instâncias de cada categoria avaliada e quantas foram classificadas como corretas por sentença. Sob requisição, é possível obter a imagem de todos os grafos gerados em ambas as versões do Graphia em português.

Essa tabela foi gerada com o objetivo de reprodutibilidade. A fim de que em trabalhos futuros seja possível recriar a avaliação aqui conduzida para analisar o desempenho do sistema ao aplicar modificações no mesmo.

Sentença[1]	ID no BD[2]	Versão[3]	SS[4]	SC[5]	V[6]	CN[7]	IR[8]
1	512	Versão 3	0/1	0	0/2	0/1	1/1
		Versão 4	1/1	0	2/2	1/1	0/1
2	573	Versão 3	2/2	0	2/3	0/1	0/1
		Versão 4	2/2	0	3/3	1/1	0/1
3	1118	Versão 3	0	1/1	1/1	0/2	1/1
		Versão 4	0	1/1	1/1	0/2	1/1
4	1471	Versão 3	2/2	0	4/4	0/3	0/1
		Versão 4	2/2	0	4/4	3/3	0/1
5	1482	Versão 3	1/1	0	1/1	1/2	0/1
		Versão 4	1/1	0	1/1	2/2	1/1
6	1484	Versão 3	1/1	0	1/1	0/2	0/1
		Versão 4	1/1	0	1/1	1/2	1/1
7	1485	Versão 3	2/2	0	1/2	0/1	0/1
		Versão 4	1/2	0	2/2	0/1	1/1
8	1487	Versão 3	1/2	0	1/2	1/1	0/1
		Versão 4	2/2	0	2/2	1/1	0/1

Sentença[1]	ID no BD[2]	Versão[3]	SS[4]	SC[5]	V[6]	CN[7]	IR[8]
9	1488	Versão 3	2/3	0	2/3	0/2	0/1
		Versão 4	3/3	0	3/3	1/2	1/1
10	4353	Versão 3	1/2	1/1	1/4	0/4	0/1
		Versão 4	2/2	0/1	3/4	2/4	0/1
11	5633	Versão 3	0	1/1	1/1	0/1	0/1
		Versão 4	0	0/1	1/1	1/1	1/1
12	11269	Versão 3	1/1	0	1/1	0	1/1
		Versão 4	1/1	0	1/1	0	1/1
13	13056	Versão 3	1/1	0/1	1/2	1/3	0/1
		Versão 4	1/1	1/1	1/2	3/3	1/1
14	15872	Versão 3	1/3	0	2/3	1/2	0/1
		Versão 4	3/3	0	3/3	2/2	1/1
15	15873	Versão 3	2/2	0	2/3	1/2	0/1
		Versão 4	2/2	0	3/3	2/2	1/1
16	18688	Versão 3	1/1	0	1/2	0	0/1
		Versão 4	1/1	0	1/2	0	1/1
17	42754	Versão 3	0/1	0	0/3	0	1/1
		Versão 4	0/1	0	1/3	0	1/1
18	58624	Versão 3	1/1	0	2/2	1/2	0/1
		Versão 4	1/1	0	2/2	1/2	1/1
19	67073	Versão 3	3/3	0	3/3	0	0/1
		Versão 4	3/3	0	3/3	0	1/1
20	69889	Versão 3	1/2	0	2/2	0/2	0/1
		Versão 4	2/2	0	2/2	2/2	1/1
21	72448	Versão 3	1/1	0	1/1	1/2	0/1
		Versão 4	1/1	0	1/1	1/2	1/1
22	81408	Versão 3	1/1	0	1/2	2/5	0/1
		Versão 4	1/1	0	2/2	5/5	1/1
23	86016	Versão 3	1/2	0	1/3	0	0/1
		Versão 4	2/2	0	3/3	0	1/1

Sentença[1]	ID no BD[2]	Versão[3]	SS[4]	SC[5]	V[6]	CN[7]	IR[8]
24	108290	Versão 3	0	1/1	1/1	0/1	0/1
		Versão 4	0	1/1	1/1	1/1	1/1
25	109570	Versão 3	1/1	0	1/1	0/2	0/1
		Versão 4	1/1	0	1/1	1/2	1/1
26	131328	Versão 3	2/2	0	2/2	0	0/1
		Versão 4	2/2	0	2/2	0	1/1
27	133376	Versão 3	0/1	0	0/3	2/2	1/1
		Versão 4	1/1	0	1/3	2/2	0/1
28	139781	Versão 3	1/2	0	2/3	0	0/1
		Versão 4	2/2	0	3/3	0	1/1
29	141569	Versão 3	1/1	0	1/1	1/2	0/1
		Versão 4	1/1	0	1/1	2/2	1/1
30	142341	Versão 3	1/2	0	2/3	1/3	0/1
		Versão 4	2/2	0	3/3	3/3	1/1
31	145664	Versão 3	1/1	0	1/1	1/2	0/1
		Versão 4	1/1	0	1/1	2/2	1/1
32	148992	Versão 3	2/2	0	2/2	1/2	0/1
		Versão 4	2/2	0	2/2	2/2	1/1
33	160768	Versão 3	1/1	0	0/4	0/4	1/1
		Versão 4	1/1	0	1/4	4/4	1/1
34	167680	Versão 3	1/2	0	1/2	0	1/1
		Versão 4	2/2	0	1/2	0	1/1
35	173056	Versão 3	1/1	0	1/2	1/2	0/1
		Versão 4	1/1	0	1/2	2/2	1/1
36	190872	Versão 3	1/1	0	1/2	0/1	0/1
		Versão 4	1/1	0	1/2	0/1	0/1
37	190875	Versão 3	1/1	0	1/1	0/3	0/1
		Versão 4	1/1	0	1/1	3/3	1/1
38	190942	Versão 3	1/1	0	1/1	0/3	0/1
		Versão 4	1/1	0	1/1	2/3	1/1

Sentença[1]	ID no BD[2]	Versão[3]	SS[4]	SC[5]	V[6]	CN[7]	IR[8]
39	190970	Versão 3	2/2	0	1/2	1/1	0/1
		Versão 4	2/2	0	2/2	0/1	1/1
40	190982	Versão 3	2/4	0	2/5	0	0/1
		Versão 4	4/4	0	4/5	0	1/1

Tabela 4: Detalhes sobre a avaliação das sentenças, apresentando seu número no website, seu ID no Banco de Dados, o número de instâncias de cada categoria avaliada e o número de acerto.

.

[1]Número da sentença no site `http://jacarepagua.dcc.ufrj.br/~andrekaren/`
[2]Banco de Dados criado em [2]
[3]Referencia às versões do Graphia, veja Anexo C
[4]Sujeito Simples
[5]Sujeito Composto
[6]Verbo
[7]Complemento Nominal
[8]Informação Irrelevante

ANEXO B – Avaliação do sujeito composto

Esse anexo traz uma explicação mais detalhada sobre o problema que o novo Graphia teve ao extrair o sujeito composto, assim como uma proposta para solucionar esse problema. Vamos utilizar o Exemplo (17) para ilustrar o motivo de o Graphia versão 1 [2] extraiu corretamente 2 dos sujeitos compostos enquanto o sistema versão 2 não conseguiu.

(17) Poisson, Liouville, Fourier e outros mais estudaram as equações em derivadas parciais e a análise harmônica

A Figura 22 destaca os três nós: 1) o nó do verbo, 2) o nó NP que o novo sistema seleciona para se conectar com o verbo de 1 e 3) o nó NP que o sistema anterior seleciona para se conectar com o verbo em 1. Perceba que o nó marcado em 2 é o nó NP do topo da pilha de NP's cuja profundidade é menor ou igual ao do verbo em questão (marcado em 1), por isso ele é o selecionado. Enquanto que o nó 3 é selecionando por ser um NP filho de S.

Como falado na Seção 5.1.3, se o sistema selecionasse os NP's seguintes cuja profundidade é a mesma do primeiro NP selecionado, os nós NP's marcados em 4 e 5 também seriam selecionados e se ligariam com o verbo em 1, fazendo com o que a extração do sujeito composto da sentença do Exemplo 17 se tornasse correta.

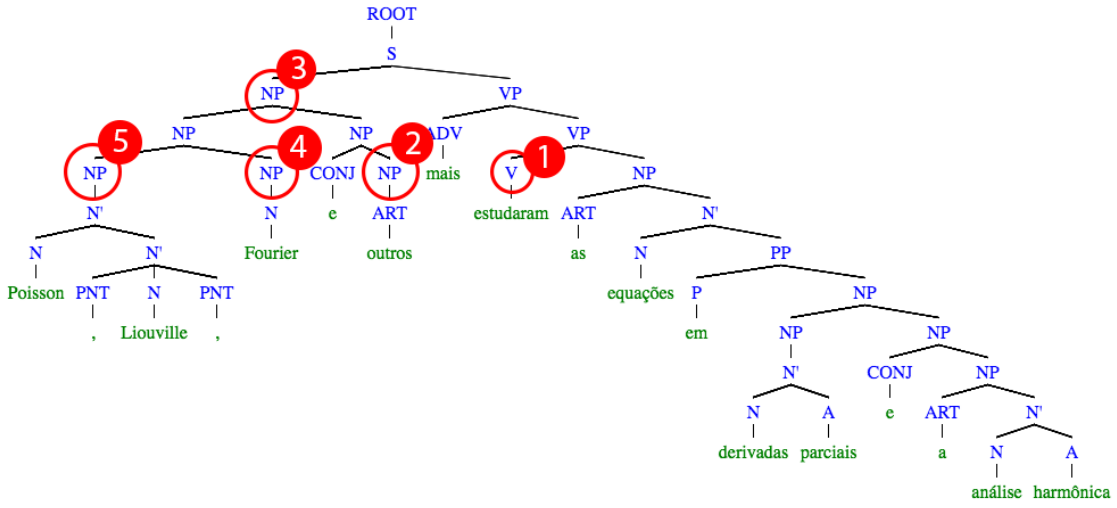

Figura 22: Árvore sintática da sentença do Exemplo (17): "Poisson, Liouville, Fourier e outros mais estudaram as equações em derivadas parciais e a análise harmônica"

ANEXO C – Tabela de versões do Graphia

Esse anexo enumera as versões do Graphia mencionadas durante a extensão do documento. Adicionalmente tem as informações de os autores, a língua aplicada no escopo e o ano de apresentação.

Versão	Autor(es)	Linguagem	Ano
Graphia versão 1	CARVALHO, D. S. [4]	Inglês	2011
Graphia versão 2	BARBOSA, R. B. [1]	Inglês	2014
Graphia versão 3	CARDOZO, A. F. P. e TEIXEIRA, K.T. [2]	Português	2016
Graphia versão 4	SOUSA, A. D. F.	Português	2018

Tabela 5: Tabela com as nomenclaturas usadas para se referir às versões do Graphia

ANEXO D – Testes unitários automáticos

Esse anexo apresenta a justificativa para a utilização de testes unitários automáticos no Graphia versão 4 e explica o processo utilizado para criar os testes. No fim traz um exemplo para melhorar o entendimento da criação dos testes.

Os testes unitários são importantes para assegurar um controle rigoroso da geração dos grafos de relacionamento. Faz parte da fase de desenvolvimento de um sistema, a alteração de várias partes do script e a realização de testes para avaliar as consequências das mudanças aplicadas. A abordagem mais ingênua para realizar esse processo é a manual, em que para cada alteração aplicada ao código, o programador gera o grafo novamente e verifica se ele continua correto.

O processo manual, viável para uma quantidade pequena de dados, se tornou inviável durante a implementação do Graphia versão 4, pois o número de sentenças a testar aumenta com as funcionalidades a implementar, assim como com o tipo e complexidade de sentenças a serem testadas durante o desenvolvimento. É comum sentenças cujo grafo extraído estava correto, se tornar incorreto ao implementar uma nova funcionalidade ou refatorar uma parte de código.

É inviável gerar manualmente cada grafo já testado e verificar se eles continuam correto a cada pequena mudança aplicada no código. Entretanto, essa limitação é facilmente corrigida ao se utilizar testes unitários automáticos. Por essa razão eles foram utilizados durante a fase de refatoração do código do Graphia versão 4.

Os testes unitários foram feitos seguindo o processo:

- Uma sentença é escolhida como caso de teste (inicialmente casos simples e gradualmente aumentando a complexidade das sentenças)
- Foi anotado manualmente o que se deseja verificar da sentença em questão, por exemplo: sujeito, verbo, predicado e complemento nominal
- A sentença é dada como entrada ao pipeline do Graphia versão 4
- Após o pipeline do Graphia versão 4 gerar uma saída para a sentença, os testes

automáticos comparam o que foi extraído pelo sistema com o que foi anotado manualmente. Se forem iguais o teste passa, senão há uma falha no teste e esta é indicada após a execução dos testes

(18) "Um vento solar de partículas de plasma corre constantemente."

Vamos utilizar a sentença (18) como exemplo. Nela, o sujeito anotado manualmente é “Um vento solar de partículas”. Durante a execução do Graphia versão 4, o sistema armazena internamente o sujeito extraído numa lista de sujeitos. Então a biblioteca de testes compara o sujeito da lista de sujeitos do grafo com a string anotada “Um vento solar de partículas”. Sendo iguais, o teste passa e é garantido que a modificação realizada não tornou a extração do sujeito dessa sentença incorreta.