Alfredo Antônio de Araújo Malheiros Filho

Semântica Inferencialista na Resolução de Anáforas Pronominais

Fortaleza - CE, Brasil

9 de abril de 2010

Alfredo Antônio de Araújo Malheiros Filho

Semântica Inferencialista na Resolução de Anáforas Pronominais

Dissertação apresentada como parte obrigatória para obtenção do título de Mestre em Ciência da Computação pela Universidade Federal do Ceará

Orientador:
Prof. Marcelino Cavalcante Pequeno, Ph.D.

Coorientadora:
Vládia Célia Monteiro Pinheiro, Dra.

DEPARTAMENTO DE COMPUTAÇÃO
CENTRO DE CIÊNCIAS
UNIVERSIDADE FEDERAL DO CEARÁ

Fortaleza - CE, Brasil

9 de abril de 2010

“Se procederes bem,
não é certo que serás aceito?”
Gn 4:7a

Agradecimentos

Agradeço principalmente a meu Deus que me apoiou durante todo este trabalho.

Agradeço a pessoas importantes que contribuíram com debates e discussões que findaram por dar o rumo atual deste trabalho. Entre estas pessoas gostaria de citar meu orientador Marcelino, minha coorientadora Vládia, os colegas Aragão, Cibele, Francicleber e os professores João Fernando e Ana Teresa.

Lembrando ainda que contei com a ajuda técnica imensurável de meus irmãos Rodrigo e Rodolfo e de minha noiva Monique.

Agradeço à Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico-FUNCAP, CE, pelo financiamento deste trabalho através da concessão da bolsa de Mestrado durante todo curso.

Resumo

Este trabalho utiliza informações semânticas inferencialistas na resolução de anáforas e aponta para a importância da união entre semântica e pragmática no âmbito do Processamento de Linguagem Natural. A Inteligência Artificial caminha para um novo rumo mais fenomenológico do que representacionalista. As ideias de Heidegger na filosofia [Hei96] e Brandom na linguística [Bra00] influenciam a Ciência da Computação dando uma nova opção de pesquisa e desenvolvimento de tecnologia. A linguagem natural carrega todo o potencial humano de criatividade e raciocínio, e por isto mesmo, representa um grande desafio para a Inteligência Artificial. A resolução de anáforas foi escolhida para implementação destas novas ideias em torno da Ciência da Computação. A resolução de anáforas é uma característica da comunicação humana e impõe dificuldades para os sistemas de PLN porque, na maioria das vezes, as informações para seu entendimento não estão explícitas no texto. Tiramos partido de informações semânticas inferencialistas no processamento de linguagem natural conforme Pinheiro [PPFF10]. Neste trabalho nos concentramos na resolução de anáforas pronominais.

Palavras-chave: Resolução de Anáforas, Linguística Computacional, Semântica Inferencialista, Processamento de Linguagem Natural, Inteligência Artificial.

Abstract

This work uses semantic inferential informations in the resolution of anaphora and points to the importance of marriage between semantic and pragmatic under the Natural Language Processing. The Artificial Intelligence goes to a new direction more phenomenological than representationalist. The ideas of Heidegger in philosophy [Hei96] and Brandom in linguistics [Bra00] influence Computer Science, providing a new option of research and technology development. Natural language carries the full potential of human creativity and skill, and for this reason, represents a major challenge for Artificial Intelligence. Anaphora resolution was chosen to implement these new ideas around the Computer Science. Anaphora resolution is a feature of human communication and imposes difficulties for NLP systems because, in most cases, the information for its understanding are not explicit in the text. We took advantage of semantic inferentialist information in natural language processing as Pinheiro [PPFF10]. In this work we focus on pronominal anaphora resolution.

Keywords: Anaphora Resolution, Computational Linguistics, Semantic Inferentialism, Natural Language Processing, Artificial Intelligence.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução p. 10

2 A pragmática linguística de Robert Brandom p. 17

3 Semantic Inferentialism Model (SIM) p. 24

3.1 Componentes do SIM . . . . . p. 27

3.1.1 Bases Semânticas do SIM . . . . . p. 27

3.1.2 SIA . . . . . p. 33

3.2 Aplicação em WikiCrimesIE . . . . . p. 38

4 Resolução de anáforas pronominais p. 40

4.1 Pronomes pessoais . . . . . p. 40

4.2 Anáforas . . . . . p. 46

4.3 Resolução de anáforas . . . . . p. 51

4.3.1 Lapin & Leass . . . . . p. 55

4.3.2 Mitkov . . . . . p. 58

4.3.3 Centering . . . . . p. 62

4.3.4 Algoritmo de Leffa, semântica sem conhecimento de mundo . . . . . p. 65

4.4 Comparação entre as abordagens explanadas . . . . . p. 70

5 Uma nova abordagem para a resolução de anáforas pronominais p. 74

5.1 Corpus linguístico e analisador sintático . . . . . p. 75
5.2 Processo de resolução de anáforas utilizando o SIA . . . . . p. 79
5.2.1 Testes preliminares . . . . . p. 79
5.2.2 Algoritmo de resolução de anáfora utilizando o SIA . . . . . p. 82
5.3 Resultados obtidos . . . . . p. 90
5.4 Análise dos resultados . . . . . p. 93

6 Conclusão p. 95
6.1 Trabalhos futuros . . . . . p. 96

Referências Bibliográficas p. 98

Lista de Figuras

3.1 modelo de arquitetura do SIM . . . . . p. 28

3.2 Grafo do Conceito “CRIME” no SIM . . . . . p. 30

3.3 Grafo da sentença-padrão “<X> <ser assassinar> <por> <Y>” no SIM . . . . p. 33

3.4 Modelo da execução do SIA . . . . . p. 35

3.5 Utilização do SIM no WikiCrimesIE . . . . . p. 38

4.1 Fluxograma do algoritmo de Leffa . . . . . p. 68

5.1 *Análise morfossintática do PALAVRAS* . . . . . p. 76

5.2 http://www.overmundo.com.br/overblog . . . . . p. 78

Lista de Tabelas

3.1 Tabela dos rótulos das arestas que representam relações entre os conceitos X e Y . . . . . p. 31

4.1 Valores iniciais dos fatores de saliência . . . . . p. 56

4.2 Valores dos fatores de saliência após a primeira sentença . . . . . p. 58

4.3 Valores dos fatores de saliência após a segunda sentença . . . . . p. 58

4.4 Possíveis transações entre o foco das sentenças . . . . . p. 63

5.1 Tabela dos *corpora* produzidos na Floresta Sintática . . . . . p. 77

5.2 Tabela dos *corpora* produzidos na Floresta Sintática . . . . . p. 77

5.3 Tabela dos pesos dos critérios sintáticos . . . . . p. 86

5.4 Tabela dos pesos dos critérios semânticos . . . . . p. 88

5.5 Tabela dos papéis presentes nos relacionamentos entre os conceitos X e Y . . p. 89

5.6 Tabela dos critérios sintáticos . . . . . p. 89

5.7 Tabela dos critérios semânticos . . . . . p. 90

5.8 Tabela dos critérios sintáticos . . . . . p. 91

5.9 Tabela dos critérios semânticos . . . . . p. 92

1 Introdução

Toda atividade laboriosa do homem esbarra na dificuldade de encontrar soluções satisfatórias para os problemas enfrentados. A muito a Computação tem investido esforços na descoberta de métodos que auxiliem o homem nas suas atividades. Um dos objetivos da Ciência da Computação é a resolução de problemas considerados difíceis ao ser humano. Problemas difíceis são aqueles em que, segundo o senso comum, o ser humano necessita esforçar-se na criação e manipulação de ideias, seguindo determinada linha de raciocínio, para então encontrar uma solução satisfatória ao problema. Um ramo da Ciência da Computação denominado Inteligência Artificial (IA) desdobra-se em esforços que compreendem a tentativa de imitar o ser humano na resolução desse tipo de problemas. Existem duas grandes correntes filosóficas na área de IA que se utilizam de modelos distintos e ambas enfrentam dificuldades [MCa07]:

1. Utilização de modelo conexionista: busca entender como funciona fisicamente o cérebro humano e copiá-lo, por exemplo, através de redes neurais artificiais. Os problemas residem em não existirem habilidades suficientes para observar eficientemente o cérebro humano. Esta abordagem depende de um avanço significativo de áreas como a psicologia e neurofisiologia.

2. Utilização de modelo simbólico: propõe conhecer de que forma os seres humanos abordam os problemas e assim desenvolver programas de computador "inteligentes" que sejam capazes de raciocinarem em busca da resolução do problema. Nesta área encontram-se os problemas relacionados ao entendimento de como o ser humano aprende e manipula o raciocínio. Esta área depende do avanço da lógica e da modelagem lógica do raciocínio.

Uma forma de raciocínio pragmático tem ganhado força na filosofia desde o final do século XIX e, a partir da segunda metade do século XX, também alcançou a Computação de um modo ainda pouco explorado. Entende-se como pragmatismo, a tentativa de compreensão do mundo a partir de um ponto de vista amplo, holista, levando em consideração as circunstâncias dos acontecimentos e acreditando que um mesmo fato pode ter significado diferenciado de acordo

com a situação em que ele está inserido. Este novo raciocínio propõe a utilização dos mais variados meios para a obtenção de resultados. São utilizadas técnicas e avanços obtidos nas duas grandes correntes filosóficas na área de IA citadas anteriormente. É neste ponto de vista que o presente trabalho se apoia, por considerar que a linguagem natural é altamente influenciada pelas situações em que é utilizada e por fatores sociais e práticos da comunidade que a utiliza.

Há uma grande diferença entre como os programas "inteligentes" realizam algumas atividades e como os seres humanos as realizam. A maior parte dos programas possuem pouco de inteligência e muito de computação [MCa07]. Falta, aos programas de IA, a percepção da prática de como algo é feito, de aspectos subjetivos e psicológicos, do tratamento do que seriam informações de senso comum, enfim falta conhecimento sobre o "significado" (em um sentido mais profundo) dos objetos manipulados. Ao invés disso os computadores trabalham com manipulações sintáticas superficialmente chamadas de "semântica", fazendo uso de abordagens que funcionam em situações de informações fechadas, ou seja, mundos fechados onde assume-se que toda a informação existente esteja disponível.

Quando se fala em informações de senso comum, refere-se a um recurso que o ser humano constrói, através do relacionamento de uns com os outros, empiricamente, observando, agindo e reagindo com o mundo e retendo, através de critérios subjetivos, aquilo que é considerado por ele mesmo, útil. Situações de informações de senso comum assemelham-se à forma natural de como as pessoas pensam e tomam decisões cotidianamente. Essas situações são mais complexas que qualquer situação onde todas as informações existentes são conhecidas, também chamadas de situações de informações fechadas. Sua complexidade se dá pela existência de fatos incompletos e não há, pelo menos *a priori*, limitações sobre as informações envolvidas. Sendo assim, é necessário o uso de conceitos aproximados que podem não ser totalmente definidos. Nessas situações as consequências das decisões não são totalmente determinadas e é necessário o uso de raciocínio não-monotônico, que consiste em uma forma de raciocínio em que se pode tomar decisões de acordo com a ausência de informações sobre algo. O termo não-monotônico se refere ao fato de, após conhecidas novas informações que faltavam, o raciocínio pode ser reconstruído [Ant96].

Atentando para o problema da atribuição de significado, um dos principais filósofos pragmáticos do início do século XX, Martin Heidegger, propõe mudanças bem mais profundas. A partir de sua visão do modo como os homens veem o mundo e a si mesmos, Heidegger descreve um modo de entender os significados através do que as coisas são no mundo e não através de representações. Ele inicia questionando sobre o que é o homem e caracterizando-o precisamente como um ser que se pergunta sobre o que é [Dre07]. O homem está fortemente influenciado

pelo seu passado e se relaciona com o mundo a partir de seu conhecimento, preocupação, angústia e complexo de culpa. O significado para Heidegger é independente da representação e se baseia nas funcionalidades das coisas e nas relações destas com o mundo [Hei96]. Um exemplo pode ser o significado do que é um martelo. Como a nossa cultura é fortemente influenciada por Descartes e pela sua ideia de representação, responder-se-ia a esta pergunta descrevendo as propriedades físicas e funcionais do martelo. Porém a mente, embora não explicite isto, trata o significado do martelo através de suas experiências com ele, como bater um prego, quebrar uma parede e a relação dele com o mundo: usado em construções, reformas e marcenaria. O mesmo se dá com um projetor multimídia. Entende-se o que ele significa porque experimenta-se o projetor na prática, ao visualizar as imagens que ele recebe ampliando-as em uma superfície vertical. Além disso, são conhecidas as suas relações com a exibição de filmes em salas de cinemas, com a apresentação de palestras em auditórios e aulas em escolas e universidades, bem como sua relação com palestrantes e professores. Não se entende como o projetor multimídia gera a luz que é projetada, mas entende-se o que é um projetor multimídia. O significado de algo não está em uma lista de suas propriedades materiais e funcionais, mas na percepção, através da vivência na prática, das funções e relações experimentadas.

Dreyfus critica fortemente a IA baseada no representacionalismo e atomismo lógico, indicando ser este um caminho fadado a repetir o fracasso obtido na filosofia [Dre07]. A representação como fundamento do pensamento pode servir para alguns tipos de objetos, para alguns nichos limitados de aplicação, no entanto, para os objetivos mais amplos de IA dever-se-ia observar as funções das coisas e dos seres e entendê-los como existentes independentemente de estarem vinculados à sua representação. É o que Heidegger chama de *readiness-to-hand* e que Dreyfus entende que deve significar as funções e relações que tornam algo útil. Para Heidegger existem duas categorias que se referem às características de existências das entidades, nomeadamente *presence-at-hand* e *readiness-to-hand*. O termo *presence-at-hand* se refere a aspectos contemplativos da entidade, por exemplo uma lista das propriedades materiais e funcionais de uma bicicleta de *Triathlon*: quadro carbono *FACT 9R*, canote *Specialized Carbono* de duas posições, rodas *ZIPP 404 clincher*, câmbio Traseiro *SRAM RED*, *shifter SRAM Force TT* 10 velocidades, mesa *Specialized S-Works Pro-Set* com ajuste de ângulo. Por outro lado, *readiness-to-hand* trata do conhecimento de senso comum em torno das funções experimentadas de uma entidade. Estas funções não são um conjunto fixo e pré-definido que responde a um determinado gatilho e resulta em sucesso ou falha, mas contempla a solicitação para algum uso, com respostas flexíveis a partir da significância de cada situação, resultando em uma melhoria ou piora, na visão do usuário, do contexto em que o usuário está inserido [Dre07]. Para uma criança a bicicleta pode ser simplesmente “um meio mais rápido de ir comprar pão e voltar para

casa" ou "um instrumento que a une aos seus amigos enquanto passeia junto com eles".

Nesta nova perspectiva, a Inteligência Artificial seria então capaz de basear-se no comportamento (experimentação) das coisas no mundo, ao invés de fundamentar-se na representação interna do mundo. As tentativas de representar o mundo são imperfeitas e carentes de realidade. Ao invés disso o ser humano evita uma representação interna olhando diretamente para o próprio mundo como sua representação [Dre07]. Então um agente inteligente se comportaria nesse mundo de acordo com o que ele vivencia do mundo, deparando-se com ideias de um racionalismo pragmático em IA.

A capacidade do ser humano de compreensão da linguagem natural consiste na mais diferenciada e decisiva faceta da nossa inteligência. Tattersall [Tat06] afirma que a aquisição da linguagem e a capacidade para arte simbólica podem estar no cerne das extraordinárias habilidades cognitivas que nos separam do restante da criação. Fellbaum [Fel98] afirma que *"a linguagem tem sido reconhecida como um dos mais interessantes aspectos do comportamento humano e talvez a mais desafiante manifestação da complexidade cognitiva humana"*. Neste sentido, a área de Processamento de Linguagem Natural (PLN) tem um lugar importante dentro da IA nessa busca por sistemas inteligentes.

As soluções atuais em PLN para entendimento de linguagem natural ou são estritamente sintáticas ou recorrem a uma representação do mundo em ontologias e bases léxico-semânticas como WordNet [Fel98] e CYC [Len95], confirmando seu caráter representacionalista. Em contraposição a essa linha representacionalista em PLN, Pinheiro et AL [PAP$^+$08][Pin9a][PPFN09] propõem o Semantic Inferentialism Model - SIM, um modelo computacional para raciocínio e expressão semântica em sistemas de linguagem natural baseado nas teorias inferencialistas de Brandom [Bra94] [Bra00], Sellars [Sel80] e Dummett [Dum78]. Estas teorias retiram a representação do centro da explicação do significado de conceitos e sentenças e privilegiam as inferências em que estes podem estar envolvidos, exercendo o papel de premissas ou conclusões. Sellars afirma que *"compreender ou entender um conceito é ter o domínio prático sobre as inferências que ele está envolvido - saber o que segue da aplicabilidade de um conceito e a partir do que ele pode ser aplicado"* [Sel80].

A teoria semântica inferencialista de Brandom (2000), na qual o SIM se baseia, define que o significado de uma sentença em linguagem natural é o conjunto de suas premissas (precondições) e conclusões (pós-condições), geradas a partir do conteúdo dos conceitos articulados em uma dada estrutura de sentença. Este conteúdo, denominado de conteúdo inferencial, são as precondições e pós-condições de uso de conceitos de uma língua, acordadas dentro da prática da comunidade linguística. O SIM contém componentes para expressão do conteúdo inferencial

de conceitos e de padrões de sentenças, bem como para expressão de conhecimento pragmático de uma determinada comunidade ou domínio específico. Seu componente principal é o agente para raciocínio semântico - *Semantic Inferentialist Analyser* (SIA).

O SIM foi aplicado em um sistema de extração de informações sobre crimes descritos em textos em língua portuguesa - *WikiCrimes Information Extractor* (WIkiCrimesIE) [Pin9a][PPFN09]. Basicamente, o sistema *WikiCrimes* oferece uma área comum para interação com usuários onde eles podem relatar e monitorar locais onde ocorreram crimes. O projeto necessita de uma ferramenta que auxilie o usuário no registro de crimes a partir de notícias da *web*. Para atender esta necessidade, *WikiCrimesIE* utiliza o analisador semântico do *SIM - SIA* e extrai informações como local do crime, tipo e causa do crime para alimentação da base de dados do *WikiCrimes*. As avaliações dos resultados do sistema nesta tarefa de PLN apontam para um problema em aberto: a necessidade de solução para resolver as anáforas e outros referentes nos textos, que será definido detalhadamente mais adiante. A seguir alguns exemplos de textos analisados pelo SIA nos quais não foi possível identificar o local do crime.

- Exemplo 1:

  C foi encontrado por populares, caído na rua Noca Dauzacker, no Nova Dourados. Ele levou de quatro a seis facadas nas costas. Foi encaminhado em estado grave para o Hospital de Urgência e Trauma de Dourados, onde permanece internado.

  Análise do problema: o pronome *ele* introduz a descrição da razão pela qual a vítima está internada em um hospital (levou de quatro a seis facadas nas costas) mas, não pode ser relacionado com *C*, automaticamente, sem a resolução da anáfora e o SIA não pode inferir o provável local do crime (rua Noca Dauzacler).

- Exemplo 2:

  O jovem foi executado quando caminhava pela Rua Américo Facó, na Bela Vista, próxima ao canal do bairro. À altura da residência de número 322, ele foi abordado por três homens que trafegavam em bicicletas.

  Análise do problema: o pronome *ele* introduz a descrição de onde exatamente ocorreu o assassinato (à altura do número 322) e como a vítima foi abordada (por três homens em bicicletas), mas não pode ser relacionado com *O jovem*, automaticamente, sem a resolução da anáfora.

- Exemplo 3:

  Um homem identificado como Fábio Raulino do Nascimento foi morto, a tiro, no começo

da madrugada de ontem... Ele morreu na hora, na Rua Tenente Benévolo, 232, enquanto os comparsas fugiram em direção ao canal do Lagamar.

Análise do problema: o pronome *ele* introduz a sentença onde encontra-se o local exato do crime (rua Tenente Benévolo, 232), mas não pode ser relacionado com *Fábio Raulino do Nascimento*, automaticamente, sem a resolução da anáfora.

- Exemplo 4:

A violência empregada pelos assassinos foi tanta, que chegou a desfigurar o rosto da vítima... Ela foi encontrada por volta das 5h, na Rua Ambrósio Boni, no bairro Cachoeira, em Almirante Tamandaré.

Análise do problema: o pronome *ela* introduz a sentença que descreve onde a vítima foi encontrada e o provável local do crime (rua Ambrósio Boni), mas não pode ser relacionado com *vítima*, automaticamente, sem a resolução da anáfora.

Um conjunto de palavras que definem um único conceito é denominado sintagma. Quando o núcleo deste conjunto é um nome, então este grupo é chamado de sintagma nominal. Anáfora é um recurso da linguagem natural utilizado para evitar que se repita um mesmo sintagma várias vezes durante o texto. Com a anáfora é possível utilizar-se de outros sintagmas para referir-se a um mesmo conceito descrito anteriormente deixando o texto mais coeso e a leitura mais agradável. A anáfora também pode se utilizar de pronomes para referir-se a sintagmas que por sua vez se referem a conceitos. Este sintagma portador do conceito ao qual o pronome se refere é denominado antecedente anafórico. Assim, uma anáfora pronominal é o caso onde um pronome é utilizado para referir-se a um conceito que já foi inserido no texto por um sintagma. Este processo de referenciação é resolvido pelo leitor durante a interpretação do texto, quando ele lê um pronome e relaciona tal palavra a um conceito presente no texto através de características sintáticas e semânticas com as quais o pronome está envolvido. Quando existem características semânticas que relacionam o pronome ao conceito a que ele se refere e há uma ausência de características sintáticas que definam a resolução da anáfora, tem-se então o caso de uma anáfora conceitual.

O presente trabalho apresenta uma proposta de evolução do analisador semântico do SIM através da implementação de uma solução para resolução de anáforas pronominais. Pretende-se ainda avançar as pesquisas em resolução de anáforas pronominais no que diz respeito à resolução de anáforas conceituais. O avanço está na inclusão de uma etapa capaz de identificar relações semântico-inferenciais entre os sintagmas e o pronome elucidando o antecedente anafórico da anáfora conceitual. Sejam as sentenças abaixo:

A equipe de atletismo continua treinando. As Olimpíadas são uma grande motivação para eles.

A única relação entre "eles" e "A equipe de atletismo" é do tipo conceitual, pois não há concordância de gênero ou número. Quando tais sentenças se encontram inseridas em um texto maior, a quantidade de sintagmas nominais candidatos a antecedente do pronome aumenta e não se consegue estabelecer nenhuma relação entre eles pelos métodos atuais fundamentados em análise sintática. Por outro lado, a teoria semântica inferencialista é capaz de expressar e manipular tal relação conceitual, pois esta, diferentemente das abordagens convencionais, não se restringe à concordância de gênero ou número ou outras restrições sintáticas, mas aborda o conteúdo conceitual dos sintagmas e o seu uso na prática linguística. Segundo Monteiro [Mon94] toda anáfora é marcada, antes de qualquer coisa, por uma ligação inferencial entre os conceitos relacionados.

Vê-se neste trabalho a utilização de *corpora* anotados. Um *corpus* literário escrito é composto por um grupo de textos escritos que são representativos de uma língua por inteira. Um *corpus* anotado é um grupo de textos escritos selecionado e analisado morfossintaticamente. Aqui utilizaremos somente este tipo de *corpus* e com a finalidade de possuir um recurso linguístico onde a partir do qual pode-se utilizar um grupo de textos selecionados previamente e fazer análises sobre ele como se estivesse analisando toda uma língua.

No capítulo 2 é explanada a teoria semântica inferencialista de Robert Brandom. O capítulo 3 explica o SIM, seus componentes e funções. O capítulo 4 relata a resolução de anáforas, explicando o funcionamento dos pronomes pessoais e das anáforas pronominais pessoais. No mesmo capítulo são apresentadas as propostas atuais para resolução de anáforas e é apresentada uma comparação entre tais abordagens. No capítulo 5 é descrita a abordagem proposta para resolução de anáforas pronominais pessoais que é a contribuição principal do presente trabalho. Ainda neste capítulo é apresentada a metodologia utilizada na implementação do resolvedor de anáforas pronominais pessoais, na avaliação dos resultados obtidos e no *corpus* utilizado para os testes. O capítulo 6 apresenta a conclusão, os resultados obtidos pelo algoritmo implementado e indicações de trabalhos futuros.

2 *A pragmática linguística de Robert Brandom*

Cada ser humano em sua essência toma parte na prática de "pedir e dar razões" [Bra00]. A comunicação, assim como a argumentação e a refutação, são traços marcantes e característicos de nossa existência. Quando pedimos razões, respondemos por nossas ações um ao outro. Além disso, nos deixamos afetar pelas razões dadas através da "força do melhor argumento". Quando praticamos esse jogo das razões, empregando conceitos que obedecem regras semânticas de pensamento inferencial, estamos mergulhando em um mundo onde as razões é que contam, como diz Sellars *"space of reasons"* [*apud* Habermas, 2004].

Making it Explicit, trabalho de Brandom que se tornou um marco na filosofia teórica [Bra94], traz uma elaboração precisa e detalhada de uma fusão inovadora de pragmática formal e semântica inferencial. Neste trabalho Brandom escreve: *"Tomar-nos por nossa capacidade para a razão e para o entendimento expressa um compromisso em tomar a sapiência, e não a senciência, como a constelação de características que nos distingue. Senciência é o que partilhamos com animais não-verbais, como gatos - a capacidade de estar conscientes no sentido de estar despertos... Sapiência refere-se ao entendimento ou à inteligência, mais que a irritabilidade ou ao despertar."* [Bra94]

Para analisar pragmaticamente a linguagem, Brandom a situa dentro do domínio dos atos de fala no jogo da argumentação. Estes atos de fala são os veículos que transportam as pretensões de verdades e as razões para suportar ou contestar tais pretensões. Uma boa razão é dependente de regras lógicas e conceitual-semânticas desenvolvidas e deduzidas na práxis de uma comunidade linguística. Esta abordagem concorda com Wittgenstein em dois pontos, ambos privilegiam o saber prático de como algo é usado e a práxis de uma comunidade linguística em detrimento do saber temático baseado em proposições e as intenções privadas individuais respectivamente [Hab04].

Na teoria de Brandom encontra-se a definição de significado sendo aplicada às práticas linguísticas. Brandom juntamente com Dummett consideram entender uma expressão linguística

sobre dois aspectos:

1. As circunstâncias nas quais ela pode ser usada, ou seja, o que autoriza a utilização de uma expressão linguística.

2. As consequências do uso, ou com o que uma pessoa se compromete ao utilizar uma expressão linguística.

Brandom relaciona essa forma de pensar com o inferencialismo, que é um modo de entender e atribuir significado realizando inferências a partir das circunstâncias de uso de uma expressão e das consequências de seu uso, fundamentando-se no uso da expressão. Dummett e Brandom se fundamentam na forma padrão dos conectivos lógicos de Gentzen. Uma regra de introdução de um conceito é o conjunto de condições suficientes para afirmá-lo (premissas) e uma regra de eliminação é o conjunto das consequências necessárias para afirmá-lo (conclusões) [Bra00][PAP$^+$08].

Através de um saber implícito, o mundo dos falantes é estruturado por uma linguagem holisticamente constituída que dispensa o conhecimento explícito de regras ou princípios. Eles adquiriram, com a prática, a possibilidade de explicitar o "saber como" e de transformá-lo tematicamente em um "saber o quê". Em princípio, ao ser capaz de falar e agir, pode-se também refletir sobre como se faz algo na prática e expressar este conhecimento em palavras através de um vocabulário lógico. É o que Piaget chama de "abstração reflexionante", a capacidade de explicitar um saber intuitivo sobre como se deve usar um vocabulário semântico segundas as regras. *"Num sentido fraco, todo ser que se engaja em práticas linguísticas, e portanto aplica conceitos, é um ser racional; no sentido forte, seres racionais não são apenas seres linguísticos, mas, ao menos potencialmente lógicos. É assim que devemos entender a nós mesmos: como seres que respondem a essa condição expressiva dual"* [Piaget *apud* Habermas, 2004].

Uma teoria da expressão deve explicar como o que é explícito surge do que é implícito, ou seja, como é formada uma sentença proposicional a partir de regras implícitas da prática discursiva e em que consistem as propriedades de uso dos conceitos presentes na proposição [Hab04]. Também deve ser explicado como estas normas implícitas podem ser explicitadas em regras e princípios. A semântica de Brandom admite que a única possibilidade de entender-se um elemento é sabendo como ele se comporta e que alterações ele causa em seu contexto. Não se pode ter um conceito sem ter muitos conceitos. Este holismo trata a explicação do significado de maneira *top-down*, registrando todo o uso de práticas linguísticas, marcando os seus conceitos e ligando-os através de uma rede de inferências. O significado é então explicado a partir do comportamento dos conceitos nesta rede inferencial. Embora alguns conceitos sejam

expressados de forma representacionalista, a atribuição de significado é sempre inferencialista. Palavras, expressões e conceitos podem funcionar como premissas ou conclusões dentro do raciocínio, pois ao usar um conceito seu conteúdo conceitual é explicitado.

É importante notar que a vivência particular do indivíduo não é enfatizada, mas sim a práxis pública da comunidade linguística. A compreensão dos conteúdos proposicionais assume o lugar da representação de objetos. Isto faz com que a comunidade linguística se reconheça como indivíduos responsáveis envolvidos em uma rede de relações. Uma vez entendidos como pertencentes a esta rede, eles devem responder uns aos outros pelas razões de seus proferimentos linguísticos, tornando o ato de dar e pedir razões a infraestrutura de suas comunicações cotidianas.

Para Brandom a pergunta sobre o que é a "verdade" é colocada de lado e assume ponto fundamental outra pergunta sobre o que deve-se fazer quando assume-se que algo é "verdadeiro" ou quando recomendamos a aceitação de uma sentença ou quando tratamos algo como útil [Hab04]. Esta estratégia antiobjetivista coloca o analista em uma posição onde ele deve se enxergar como interlocutor dos proferimentos. Mais ainda, ao analisar uma sentença ele deve ter uma atitude performativa de participante do discurso onde considera e trata a pretensão de verdade que está sendo analisada dentro da prática linguística. Outra consequência desta metodologia é a substituição da pergunta sobre o que significa compreender ou interpretar corretamente uma afirmação pela pergunta sobre o que faz um intérprete ao considerar e tratar corretamente um falante como alguém que traz uma pretensão de verdade através do seu ato de fala. O intérprete atribui ao falante a pretensão à verdade "P", o falante se compromete com as consequências de "P" e o intérprete atribui ao falante as consequências da pretensão à verdade "P". Após isso, dependendo do peso de "P" o falante se sente obrigado a dar razões para a aceitação de "P" como verdade. O intérprete por sua vez também atribui a "P" o peso que ele próprio considera adequado e analisa se o argumento dado pelo falante supre a sua necessidade de razões para o peso determinado e então, se for o caso, o intérprete reconhece o direito do falante de proferir "P". Note que ambos, intérprete e falante, se baseiam nas razões que existem para "P" ser proferido e nas consequências que "P" traz ao ser proferido.

Brandom afirma que é o proferimento de um falante que faz parecer adequado a um intérprete atribuir ao falante a pretensão à verdade e um comprometimento com a mesma. A mudança de status do proferimento de "pretensão à verdade" para "aceito como verdade" depende, portanto, de como o intérprete recebe este proferimento, sendo necessário então analisar como ele se apresenta ao intérprete. Esta abordagem fenomenológica dos atos de fala compara uma conversa com um jogo. Através da prática discursiva os participantes vão realizando pro-

ferimentos de pretensões à verdade e dando razões para serem tratadas como verdade. Cada participante atribui pontos para os outros participantes que dão razões suficientes para serem aceitos, no seu ponto de vista, como proferidores de verdades. A cada momento é contabilizado quem obteve pontos suficientes para continuar sendo acreditado e quem a partir de então será desacreditado.

Brandom atrela a essa descrição de práticas discursivas uma teoria semântica que se encaixa muito bem ao se apropriar da explicação de significado proposta por Dummett: *"compreende-se uma proposição quando se conhece tanto as condições em que ela pode ser aplicada como as consequências que sua aplicação implicaria para os envolvidos"* [Dummett *apud* Habermas, 2004]. Esta concepção de compreensão linguística se refere a uma segunda pessoa que pode pedir razões para a satisfação de suas condições de verdade e extrair consequências do proferimento aceito. As condições de emprego de uma asserção e suas consequências podem ser confirmadas por um tipo de inferência chamada de "inferência material" que se apoia no conteúdo semântico das asserções, porém Brandom não se limita ao uso de tais inferências, reconhecendo a existência de razões que não necessitam de fundamentação adicional. Não é o conhecimento empírico, mas a prática linguística que dá o conhecimento ao intérprete das regras de condições e consequências do uso correto de expressões linguísticas. A prática linguística é mais uma geradora de conceitos, condições, consequências e inferências materiais do que refém deles, sendo portanto ao mesmo tempo limitada por eles e criadora de novos conceitos que lhe trazem a liberdade de volta.

As circunstâncias em que um conceito pode ser utilizado também é denominado de precondições do conceito e as consequências do uso de um conceito de pós-condições. As precondições funcionam como regras de introdução ao se obter um conjunto de condições suficientes para afirmar um conceito. Elas dão o direito de utilizar um certo conceito e servem de premissa para inferências, argumentos, proferimentos e raciocínios. É tudo que é suficiente para que um conceito possa ser usado, e se tal conceito já foi usado, então pode-se inferir que suas precondições foram satisfeitas. As pós-condições funcionam como regras de eliminação. Ao usar um conceito, compromete-se com todo o conjunto de condições necessárias para que ele seja usado. Elas são as consequências do uso de um conceito. Permitem saber com o que se compromete ao proferir alguma sentença. Pode-se fazer inferências também das precondições para as pós-condições, servindo de premissas para outros conceitos serem utilizados. Sendo assim, pode-se fazer inferências a partir das consequências do uso de um conceito para as circunstâncias de uso de outro conceito aumentando o conhecimento adquirido com o proferimento da sentença. Os conceitos são portanto ligados a outros conceitos pelas suas pré e pós condições, formando uma rede inferencial de conceitos que permite entender o significado dos conceitos enquanto

participantes de sentenças [PAP$^+$08].

Para Brandom a sentença é a unidade fundamental do raciocínio e expressões sub-sentenciais só têm sentido quando participantes de sentenças na construção do argumento. Sentenças são expressões autônomas cujo significado não depende diretamente de um proferimento maior e que representa um ato de fala num jogo de linguagem, elas alteram os compromissos assumidos ao serem proferidas ou servem para dar direito a outros proferimentos ou a inferências. Expressões sub-sentenciais não têm nenhuma dessas funções mas possuem significado ao contribuírem com o significado das expressões em que estão inseridas.

As condições e consequências de proferimentos vão mudando e sendo atualizadas e validadas com a prática linguística. Uma rede inferencial vai sendo montada com os proferimentos e os conteúdos conceituais e em certo ponto no tempo, tem-se todos os conceitos interligados entre si com premissas e conclusões e várias inferências que vão sendo validadas a cada ato de fala do jogo de linguagem. O uso de termos para expressar desejos, vontades, preferências, obrigações e deveres servem para apoiar um padrão de inferências materiais, não sendo necessários para tais inferências nem servindo como premissas.

"Expressões veem a significar o que elas significam ao serem usadas como o são na prática, e estados intencionais e atitudes têm os conteúdos que têm em virtude do papel que desempenham na economia comportamental daqueles a quem são atribuídos. O conteúdo é entendido em termos de propriedades inferenciais, mas estas, por sua vez, são entendidas em termos de atitude de instituição normativa, que consistem em considerar ou tratar algumas atitudes como estando ou não estando apropriadas na prática. Uma via teórica é tornada acessível pelo que as pessoas fazem com o que querem dizer, a partir de sua prática dos conteúdos de seus estados e de suas expressões. Dessa maneira, uma teoria pragmática apropriada pode fundar uma teoria semântica inferencialista; suas explicações do que é na prática tratar as inferências como corretas ou incorretas são o que, afinal, autoriza o recurso às propriedades materiais da inferência (que pode então desempenhar o papel de elementos semânticos primitivos)." [Bra00]

O conceito de prática discursiva de Brandom se deve a primeira seção de "Ser e Tempo" de Heidegger. O "ser-no-mundo" de Heidegger é definido por significações contextuais da prática ao manipular as coisas. O significado do *readness-to-hand* é determinado pelas suas reações às coisas em ações do cotidiano e pelo que uma comunidade entende como sendo reações adequadas e apropriadas, atribuindo ao significado aquilo como ele é visto. Para o indivíduo a tendência é responder a estímulos da mesma maneira que seus semelhantes. O significado será então construído pela reciprocidade entre os indivíduos de uma comunidade ao tratarem semelhantemente as reações como adequadas ou apropriadas. A autoridade epistêmica dos in-

divíduos é vinculada a autoridade social da comunidade. Cada indivíduo se reserva ao direito de não dar à autoridade epistêmica social a última palavra quanto a validade epistêmica. Cada um deve adquirir a clareza de que todos os indivíduos são falhos, portanto qualquer comunidade linguística pode consequentemente falhar e ele mesmo deve aprender meios de detectar tais falhas. Note que ambas as coisas tornam-se plausíveis, tanto assemelhar-se a comunidade em sua práxis linguística quanto se reservar ao direito da racionalidade quanto às normas.

Seguindo um raciocínio intuitivo Brandom explica a existência de "termos singulares" onde se supõe a existência de objetos reais a que estes termos referenciam e lhe são negadas ou atribuídas qualidades. Esta reflexão apoia-se no fato de que, em alguns casos, cabe a substituição de uma expressão por outra equivalente. São utilizados termos como "... refere-se a" e "... é verdadeiro", que se tornam essenciais para a descrição do estado das coisas, como operadores para a construção de expressões anaforicamente dependentes.

Termos singulares no âmbito da linguagem marcam alguma coisa no mundo real, definindo do que se fala e sobre o que se fala, de modo que se possa referir a objetos através de outras descrições. Porém a referência ao mundo não se limita a objetos, ela se completa com a referência a fatos, que podem ser afirmados sobre um objeto. Isto é explicitado na forma de proposições sobre um certo estado das coisas enunciadas, este estado é então tematizado. Quando o intérprete emite uma outra opinião sobre o mesmo tema, ele deve se referir a este termo singular para garantir que embora esteja emitindo um outro ponto de vista, ele se refere ao mesmo objeto ou estado tematizado anteriormente. Se o intérprete toma um ponto de observação diferente do falante, ele pode ver os mesmos objetos e estados das coisas de modo distinto. Se for o caso, o intérprete passa a considerar a proposição do falante desmerecedora de crédito, pois ele acredita que o falante se engana em relação às consequências do que disse. Então o intérprete repudia a pretensão de verdade do falante, pois ele se apoia em um potencial de inferências, que não foi esgotado pelo falante, mas que está contido no proferimento do mesmo.

Brandom ainda explica que a observação ou percepção, é um tipo de exposição não-inferencial e que pode dar um fim à regressão em busca da validação de uma condição de uso de conceitos ou proferimentos. Uma observação ou percepção é um tipo de justificador que não necessita de nenhuma outra fundamentação. Neste contexto Brandom acredita que existam "observadores confiáveis" treinados empiricamente que respondem diferencialmente ao seu meio por um compromisso reconhecidamente legítimo.

"As inferências que extraem das circunstâncias as consequências de aplicação (implícitas nos conteúdos conceituais) estão sujeitas à critica empírica em virtude das conexões inferenciais entre conteúdos dos compromissos que podem ser assumidos não-inferencialmente. Pode

ocorrer que alguém use o termo "ácido" de um modo que o sabor amargo de uma substância seja condição suficiente para aplicá-lo, e que o fato de ele tornar vermelho o tornassol seja consequência necessária de sua aplicação. Encontrar uma substância que ao mesmo tempo tenha sabor amargo e torne azul o tornassol mostra que tal conceito é inadequado." [Bra00]

3 *Semantic Inferentialism Model (SIM)*

A *web* possui um conteúdo extraordinariamente vasto, não estruturado e em linguagem natural. Sistemas de informação capazes de automaticamente explorar este conteúdo são ferramentas extremamente úteis para o gerenciamento do conhecimento. A finalidade de tais ferramentas está em localizar e extrair informações relevantes para um determinado objetivo e estruturá-las facilitando o seu manuseio e análise. Como as informações se apresentam em linguagem natural faz-se necessário o uso de um módulo semântico que entenda o texto para poder extrair as informações desejadas [PAP+08].

Para atender a esta necessidade, as pesquisas em Processamento Natural de Linguagem (PLN) devem avançar no sentido de entender o significado dos termos e sentenças expressos em linguagem natural e não adotar abordagens puramente sintáticas. Pinheiro et AL [Pin9a] compreendem que grandes limitações desta área são geradas pela abordagem usual utilizada para expressão do conteúdo semântico nos sistemas de PLN. As abordagens usuais, em geral, utilizam uma representação do mundo através de classes, relações e características dos objetos representados pelos termos de uma língua, desconsiderando seus usos na prática linguística. Esta abordagem para expressão de conhecimento semântico é insuficiente em possibilitar que agentes inteligentes se aproximem da qualidade de entendimento de linguagem natural dos homens. As abordagens atuais são fundamentadas no paradigma representacionalista, concebido em Frege, Russell, Tarski e Descartes, estes, os mais influentes nomes [Pin9a]. Um segundo problema diz respeito ao raciocínio no nível semântico-pragmático: a predominância de inferências formais e o uso de uma abordagem atomista que entende que o significado da sentença é formado pela combinação do significado dos termos da sentença, mas não aborda o fato de que o significado dos termos é também determinado pelo significado da sentença como um todo.

Como proposta de solução aos problemas relatados, Pinheiro et AL propõem o *Semantic Inferentialism Model* (SIM) [PAP+08], um modelo semântico inferencialista que especifica os requisitos para expressar e raciocinar sobre conhecimento semântico inferencialista, habilitando os sistemas com a geração de premissas e conclusões das sentenças em linguagem natural. Essa abordagem propõe um novo modelo de expressão do significado que baseia-se nas teorias

semânticas inferencialistas de Dummett [Dum78], Sellars [Sel80] e Brandom [Bra94][Bra00]. Essas teorias fundamentam-se no fato de que entende-se uma sentença quando se está habilitado a defendê-la ou refutá-la dando argumentos e explicações. Isto acontece porque somos capazes de inferir as premissas que autorizam ou proíbem o proferimento de uma sentença e quais as conclusões do proferimento da mesma.

Inspirado nessa visão inferencialista, o SIM propõe que expressar o conteúdo de um conceito requer tornar explícito seu uso em inferências, como premissas ou conclusões de raciocínios. E o que determina o uso de um conceito em inferências que este conceito pode participar são:

- Suas precondições ou premissas de uso: o que dá direito a alguém a usar o conceito e o que poderia excluir tal direito, servindo de premissas para proferimentos e raciocínios.
- Suas pós-condições ou conclusões de uso: o que se segue ou as consequências do uso do conceito, as quais permitem saber com o que alguém se compromete ao usar um conceito, servindo de conclusões do proferimento em si e de premissas para futuros proferimentos e raciocínios.

Este conteúdo define o importe ou competência inferencial de um conceito. Esta visão sobre o conteúdo de conceitos evita a necessidade de uma representação do mundo *a priori*. O que precisa ser expresso sobre um conceito deve ser feito considerando seus usos em práticas linguísticas. Isto concorda com a ideia de que conceitos surgem dentro da prática linguística de uma comunidade, sociedade ou de uma área do conhecimento e são apreendidos pelos usuários de uma língua a partir de seus usos e não por que existem *a priori* no mundo com características pré-definidas. Por exemplo, o conceito "saidinha bancária" se originou dentro da prática linguística de se descrever assaltos em que os clientes são abordados após realizarem saques em agências bancárias. Tal conceito não se originou pelas representações deste tipo de crime, mas pelas circunstâncias e as consequências que guiaram seu uso pela comunidade linguística. Os usuários deste conceito aprenderam em que situações pode-se usá-lo e o que se segue do seu uso. Embora existam os conceitos "saidinha" e "bancária", a nova expressão linguística "saidinha bancária" denota um conteúdo com valor semântico distinto que é caracterizado pelo seu uso em sentenças.

Com relação ao problema de como computar o significado de uma sentença em linguagem natural, o paradigma inferencialista traz uma nova forma de pensar sobre semântica que é diferente da tradicional visão representacionalista. Enquanto no representacionalismo têm-se uma

visão atomista na atribuição de interpretações semânticas, no inferencialismo há uma predominância holista. A tradição na semântica formal representacionalista tem sido uma visão atomista no sentido de que a atribuição de uma interpretação semântica a um elemento é tratada de forma independente da atribuição semântica dos demais elementos de uma sentença. Ao contrário, a semântica inferencialista é essencialmente holista: não se pode definir o valor semântico de um elemento sem considerar os outros elementos relacionados em uma sentença e como todos estão estruturados. Na semântica inferencialista define-se "essencialmente holista" por esta característica que é uma consequência direta e simples da concepção inferencial do conteúdo de conceitos - *"ninguém pode ter qualquer conceito a menos que tenha muitos conceitos"* [Bra00]. Ao expressar as potenciais inferências em que um conceito pode estar envolvido nada mais fazemos que expressar as relações inferenciais deste com outros conceitos e, na medida em que conhecemos um conceito, conhecemos vários. Conceitos, portanto, são ligados a outros conceitos através de suas pré e pós-condições de uso, formando uma rede inferencial que nos leva a saber o que autoriza a aplicação do conceito e o que se segue de sua aplicação em sentenças.

A rede de potenciais inferências em que conceitos podem participar consiste em uma base para raciocínio material e holístico na apreensão do significado de sentenças. Raciocínio material porque temos à mão o conteúdo inferencial dos conceitos, que possibilita a realização de inferências autorizadas pelo conteúdo e argumentos para refutar e validar inferências. No primeiro caso a inferência "se" "um relâmpago é visto agora" "então" "um trovão será ouvido em breve", é autorizada pelo conteúdo dos conceitos "trovão" e "relâmpago", e, no segundo caso, a afirmação "A água é vermelha" pode ser refutada pela precondição de uso do conceito "água" que define que este conceito só pode ser usado em sentenças onde não é associado ao mesmo uma cor.

O raciocínio holístico, por sua vez, é baseado no conteúdo inferencial dos conceitos e em como os conceitos estão relacionados na estrutura da sentença autorizada por uma gramática. Neste raciocínio, deve-se considerar o todo (sentença) e como suas partes (elementos subsentenciais) estão estruturalmente relacionadas a fim de definir a contribuição semântica de cada parte para com o todo (sentença). Nesta abordagem holística, as estruturas de sentenças assumem um papel importante porque, para determinar o valor semântico de um elemento subsentencial, deve-se considerar os outros elementos relacionados e é imprescindível levar em conta a estrutura que os organiza na sentença e que define suas formas e funções sintáticas. Por exemplo, para refutar a sentença "A água é vermelha" é necessário conhecer que o conceito "vermelho" está qualificando o conceito "água" e isto é possível através da análise da estrutura da sentença. As duas qualidades de raciocínio semântico em linguagem natural, aqui explanadas, material e holístico, oferecem uma boa indicação de como computar o significado de sentenças.

Em síntese o SIM foi construído sobre os seguintes axiomas [Pin10]:

- AXIOMA 1: O conteúdo de conceitos deve ser explicado em termos do papel destes em raciocínios.
- AXIOMA 2: A sentença é a unidade fundamental da linguagem e as expressões subsentenciais só têm significado como constituintes de sentenças.
- AXIOMA 3: Práticas linguísticas envolvem inferências materiais e raciocínio não monotônico.

3.1 Componentes do SIM

Para expressão e raciocínio sobre conteúdo semântico inferencialista, o SIM conta com os seguintes componentes [Figura 3.1]: base conceitual, base de sentenças-padrão, base de raciocínio prático e o analisador semântico denominado *Semantic Inferentialist Analyser* (SIA).

3.1.1 Bases Semânticas do SIM

As seguintes definições serão utilizadas para explicar as bases semânticas [Pin10]:

- C é o conjunto dos conceitos de um língua natural:

  $C = \{c_1, c_2, ..., c_n | c_i$ é um conceito de uma língua natural $\}$.

- R_c é o conteúdo inferencial dos conceitos de uma língua natural:

  $R_c \subset CxC = \{r_{c1}, r_{c2}, ..., r_{cm} | r_{cj}$ é uma relação inferencial entre quaisquer dois conceitos $c_i, c_k \in C\}$.

- P é o conjunto das sentenças-padrão:

  $P = \{p_1, p_2, ..., p_q | p_i$ é uma sentença-padrão $\}$.

- T_p é o conjunto das partes nominal, verbal e complementar de todas as sentenças-padrão em P : $T_p = \{t_{p1}, t_{p2}, ..., t_{ps} | t_{pi}$ é uma das partes (nominal, verbal ou complementar) de uma sentença-padrão qualquer $p_j \in C\}$.

- R_p é o conteúdo inferencial das sentenças-padrão:

  $R_p \subset T_p xC = \{r_{p1}, r_{p2}, ..., r_{pt} | r_{ci}$ é uma relação inferencial entre uma parte de uma sentença-padrão $t_{pj} \in T_p$ e um conceito $c_k \in C\}$.

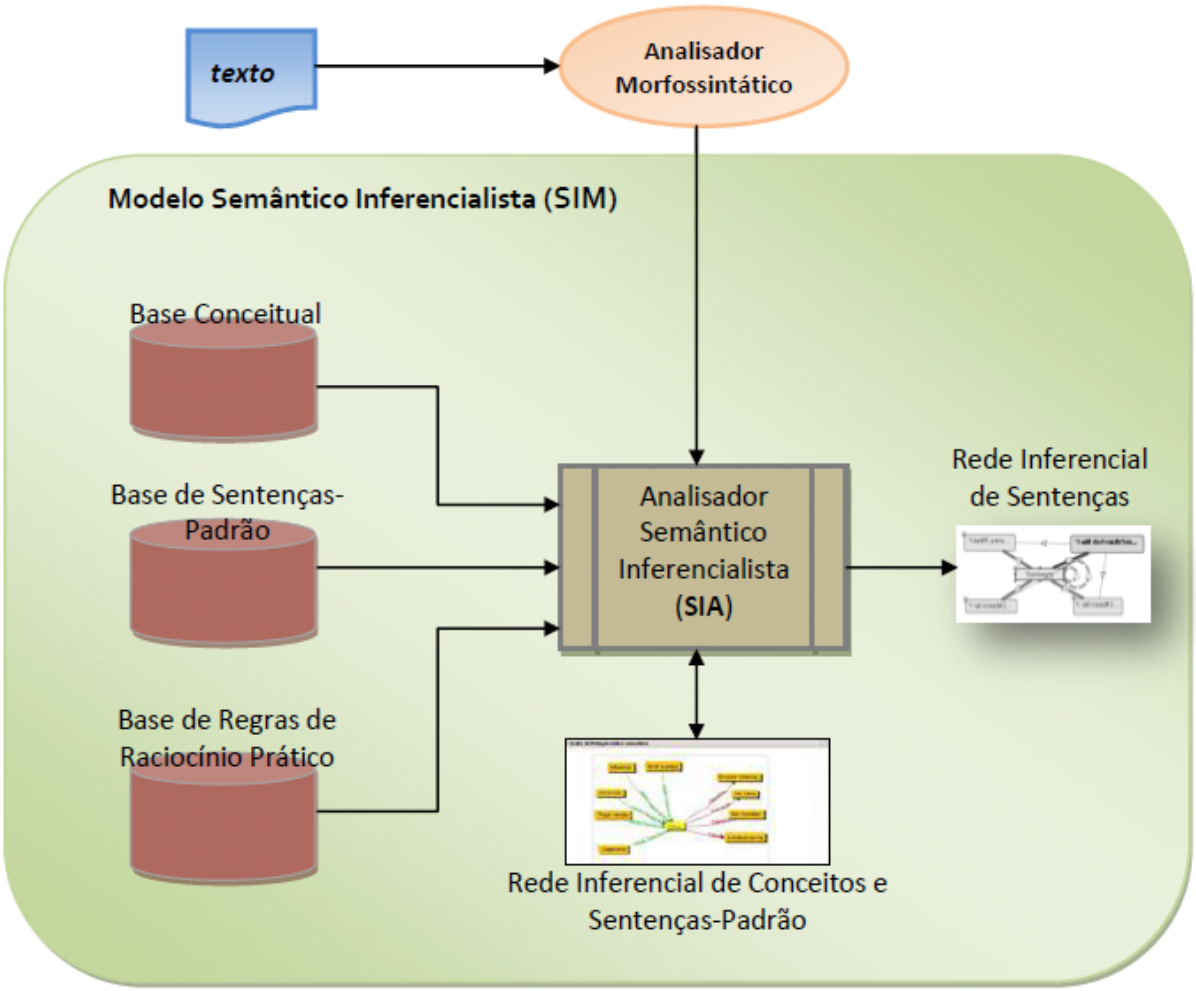


Figura 3.1: modelo de arquitetura do SIM

- N é o conjunto dos tipos de relações semânticas entre dois conceitos quaisquer $c_i, c_k \in C$:

  $N = \{nome\_relacao_i | nome\_relacao_i$ é o nome de uma relação semântica $\}$.

Base Conceitual: armazena o conteúdo de conceitos em língua natural, que são definidos em uma determinada comunidade ou área de conhecimento. O conteúdo de um conceito são as potenciais inferências em que ele pode participar e o que determina estes relacionamentos inferenciais são as premissas e conclusões que relacionam um conceito a outro. É utilizado nesta base um grafo direcionado $G_c(C, R_c)$, $C =$ (os vértices do grafo), $R_c =$ conjunto de arestas rotuladas (vide tabela 3.1) por uma variável que indica o tipo de pré ou pós condição que relaciona um conceito a outro em C. Cada relação inferencial $r_{cj} \in R_c$ é representada por uma tupla $(nome\_relacao, c_i, c_k, tipo)$, onde:

- $nome\_relacao \in N$
- c_i e $c_k \in C$

- tipo = “Pre” ou “Pos”

A rede inferencial de um conceito particular c_i é representada pelo subgrafo $G_{ci}(C_i, R_{ci})\, de\, G_c$, onde:

- C_i é o subconjunto de C formado pela união do conceito c_i com todos os conceitos com os quais c_i se relaciona (elementos do conjunto imagem do conjunto R_{ci}): $C_i \subset C = \{c_i\} \bigcup \{c_j | c_j \in I(R_{ci})\}$
- R_{ci} é o subconjunto de R_c que contém as relações inferenciais em que o conceito $c_i \in C$ é o conceito domínio da relação: $R_{ci} \in R_c = \{r_{cj} | D(r_{cj}) = \{c_i\}\}$.

A Figura 3.2 representa como o conceito “CRIME” é armazenado no SIM. A figura ilustra as seguintes relações:

- (capazDe, “crime”,”ter vítima”, “Pre”)
- (capazDeReceberAcao, “crime”,”evitar por polícia”, “Pre”)
- (primeiroSubEventoDe, “crime”,”escolher a vítima”, “Pre”)
- (usadoPara, “crime”,”vingança”, “Pre”)
- (capazDeReceberAcao, “crime”,”cometer com arma”, “Pre”)
- (capazDe, “crime”,”envolver violência”, “Pre”)
- (éUm, “crime”,”violação da lei”, “Pre”)
- (motivacaoDe, “crime”,”vingança”, “Pre”)
- (efeitoDe, “crime”,”culpa”, “Pos”)
- (efeitoDesejavelDe, “crime”,”julgamento”, “Pos”)

Base de sentenças-padrão: contém sentenças genéricas que obedecem a uma dada estrutura sintática e que funcionam como *templates*, cujos *slots* podem ser preenchidos com termos em linguagem natural e também contém as relações da sentença com outros conceitos em forma de pré e pós condições. Uma sentença-padrão segue certa estrutura de sentença, por exemplo do tipo “X ser assassinar por Y” que segue a estrutura <sentença> ::= <SN> <SV> <SP>, com algumas partes variáveis a serem instanciadas (*slots*) por elementos da base conceitual

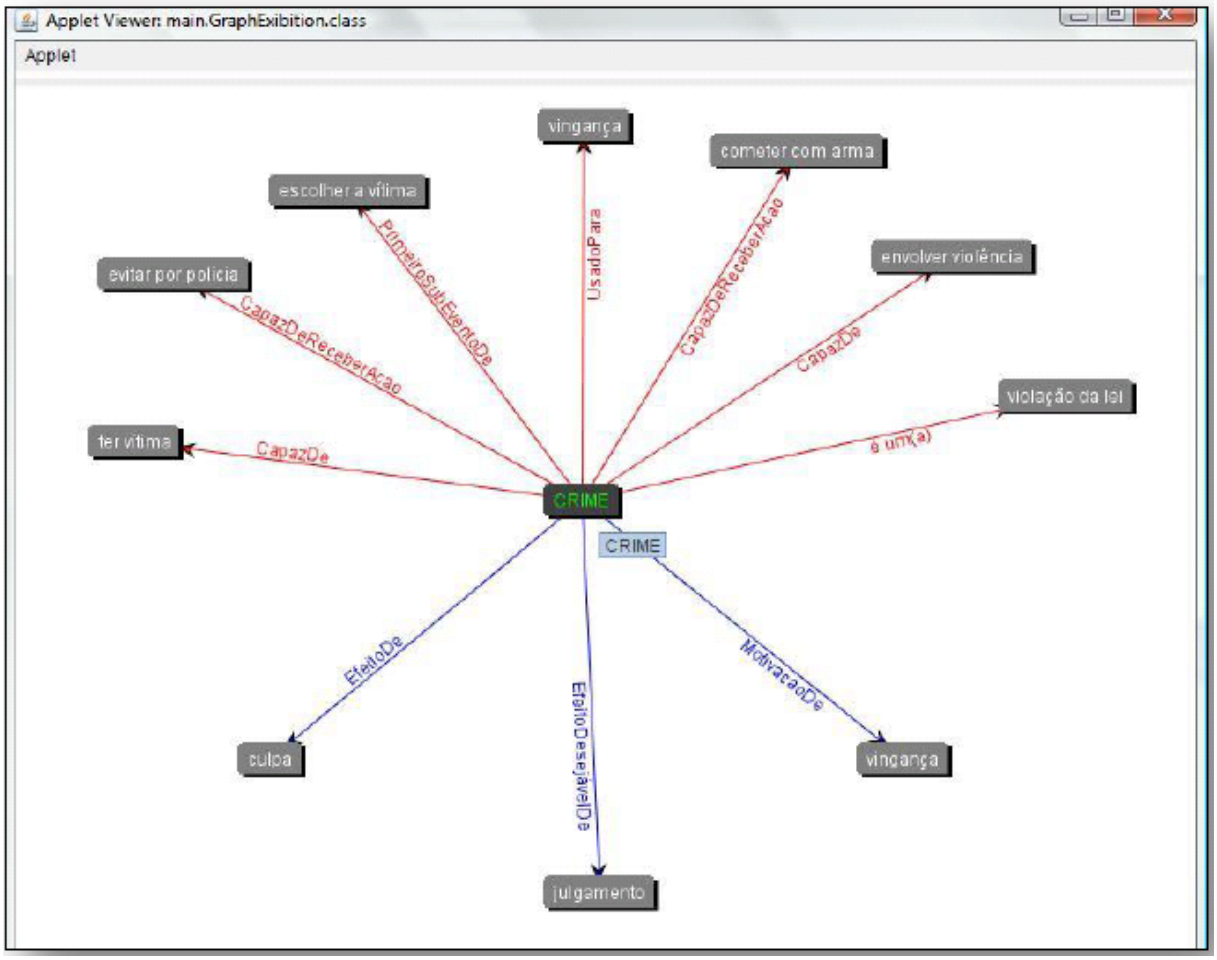


Figura 3.2: Grafo do Conceito “CRIME” no SIM

ou por outros elementos subsentenciais (nomes próprios, artigos, preposições, conjunções etc). Neste exemplo, o sintagma nominal (SN), representado por X, poderia ser preenchido por “uma mulher” e o sintagma preposicional (SP), representado por Y, poderia ser complementado por “seu amante”. Têm-se então a sentença “uma mulher ser assassinar por seu amante”, gerada a partir do padrão “X ser assassinar por Y”. A relevância da base de sentenças-padrão para análise semântica se dá porque parte do que se pode inferir ao se ler uma sentença não advém direta e unicamente, pelo menos de forma eficiente, do conteúdo dos conceitos da sentença, mas dos conteúdos dos conceitos combinados e articulados sob determinada estrutura de sentença. Por exemplo, uma pessoa que lê a sentença “uma mulher foi assassinada por seu amante” é capaz de responder quem foi a vítima (“uma mulher”) e quem foi o assassino (“seu amante”). Um mecanismo de raciocínio para gerar estas conclusões poderia até raciocinar sobre o conteúdo do conceito “assassinar” - precondições de uso “existir um assassino” e “existir uma vítima” - porém, identificá-las de forma direta e eficiente na sentença exigiria conhecimento de como articular este conteúdo e o elemento estruturador “por”. Este conhecimento é justamente o que a base de sentenças-padrão provê: expressar conteúdo inferencial (premissas e conclusões)

Rótulos	Tradução Possível
X DefinedAs Y	X é definido como Y
X PartOf Y	X é parte de Y
X UsedFor Y	X é usado para Y
X CapableOf Y	X é capaz de Y
X DesirousEffectOf Y	X possui o efeito desejado Y
X EffectOf Y	X possui o efeito Y
X MotivationOf Y	X é motivação para Y
X SubEventOf Y	X é um sub-evento de Y
X CapableOfReceivingAction Y	X é capaz de receber a ação Y
X FirstSubEventOf Y	X é o primeiro sub-evento de Y
X LastSubEventOf Y	X é o último sub-evento de Y
X PreRequisiteEventOf Y	X é um evento pré-requisito de Y
X DesireOf Y	X é desejo de Y
X LocationOf Y	X está localizado em Y
X PropertyOf Y	X é uma propriedade de Y
X IsA Y	X é um Y
X MadeOF Y	X é feito de Y

Tabela 3.1: Tabela dos rótulos das arestas que representam relações entre os conceitos X e Y

das sentenças-padrão. Este conteúdo inferencial consiste de conhecimento que não pode, pelo menos de forma eficiente, ser inferido do conteúdo dos conceitos.

A Base de Sentenças-Padrão é representada em um grafo direcionado $G_p(V, R_p)$, onde:

- $V = D(R_p) \bigcup I(R_p)$

  Conjunto não vazio formado pela união de sentenças-padrão p_j (conjunto domínio do conjunto R_p) e conceitos c_i (conjunto imagem do conjunto Rp).

- R_p

  Conjunto não vazio de relações inferenciais r_{pj} das sentenças-padrão.

Cada relação inferencial $r_{pj} \in R_p$ é representada por uma tupla $(nome\_relacao, t_{pi}, c_k, tipo)$, onde:

- $nome_relacao \in N$
- $t_{pi} \in T_p$
- $c_k \in C$
- tipo = “Pre” ou “Pos”

A rede inferencial de uma sentença-padrão particular p_i é representada pelo subgrafo $G_{pi}(V_i, R_{pi}) de G_p$, onde:

- $V_i = D(R_{pi}) \bigcup I(R_{pi})$

  Conjunto não vazio formado pela união das partes da sentença-padrão p_i (conjunto domínio do conjunto R_{pi}) e conceitos c_i (conjunto imagem do conjunto R_{pi}).

- R_{pi}

  Subconjunto de R_p que contém as relações inferenciais r_{pj} cujo elemento domínio é uma das partes da sentença-padrão p_i : $R_{pi} \subset R_p = \{r_{pj} | D(r_{pj}) = \{t_{pi}\}, t_{pi} \in T_p$ é uma das partes da sentença-padrão $p_i\}$.

A Figura 3.3 ilustra o grafo $G_{assassinar}$ que representa a rede inferencial da sentença-padrão $assassinar =$ " $< X >< serassassinar >< por >< Y >''$, e seu conteúdo inferencial expresso nas relações:

- (éUm, sn(assassinar),"pessoa", "Pre")
- (éUm, sn(assassinar),"vítima", "Pos")
- (éUm, sp(assassinar),"assassino", "Pos")

Base de regras de raciocínio prático: são regras que combinam o conteúdo dos conceitos enunciados com outros conceitos enunciados e da base conceitual, com o objetivo de possibilitar a expressão de conhecimento prático existente na comunidade linguística. Estas combinações visam aprimorar a relação do raciocínio semântico que está sendo automatizado com as regras de senso comum utilizadas por nós no dia-a-dia. Brandom [Bra00] relaciona regras deste tipo com conceitos como "preferência", "compromisso", "obrigação". Por exemplo: "João é um policial". O elemento subsentencial de s, c = "policial" possui a pós-condição temCompromisso("policial", "estar fardado"), em outras palavras, "policiais têm o compromisso de estarem fardados". Então conclui-se que "João deve está fardado".

As Regras para Raciocínio Prático formam um conjunto de cláusulas Horn da forma $(A_1 \wedge A_2 \wedge ... \wedge A_n \rightarrow B)$, onde:

- A_i pode ser:
  - uma sentença s_i

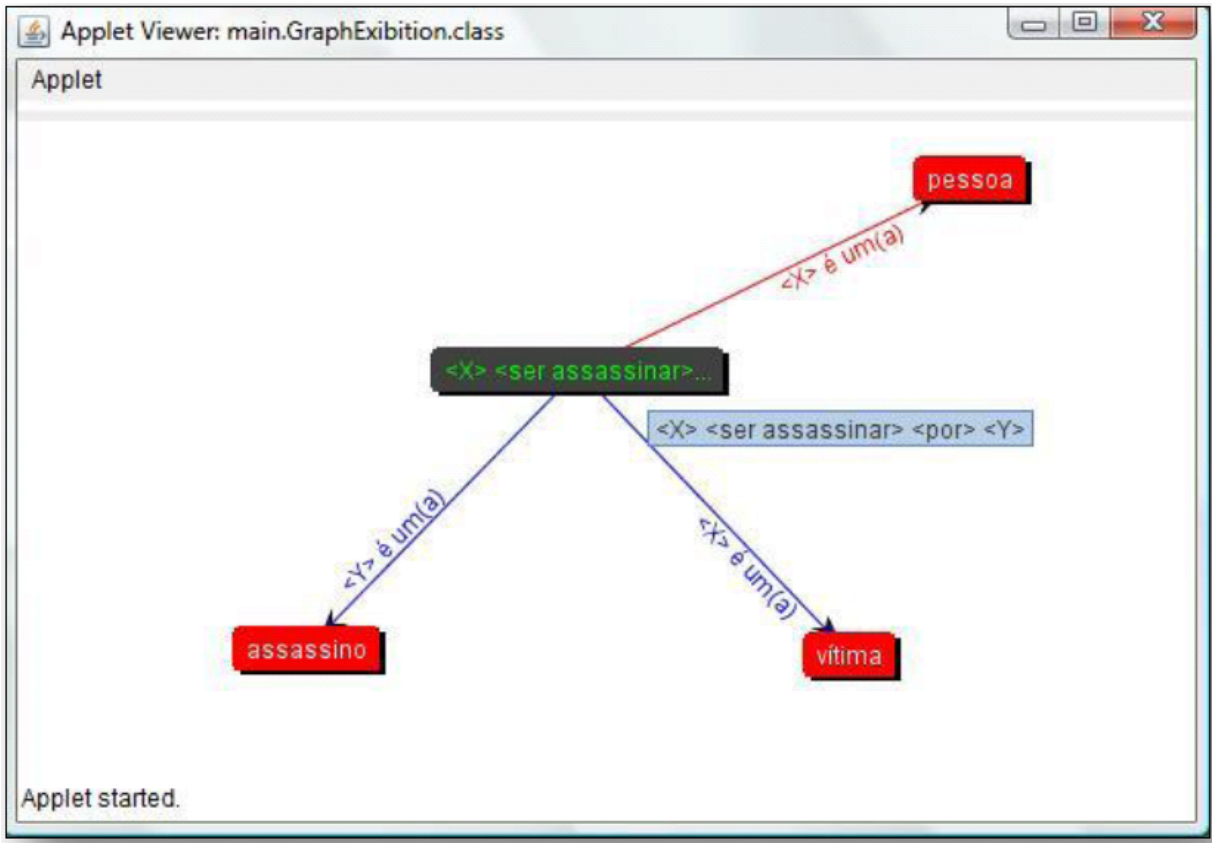


Figura 3.3: Grafo da sentença-padrão "<X> <ser assassinar> <por> <Y>" no SIM

  - uma relação inferencial $rc_j \in R_c$
  - uma relação inferencial $rp_j \in R_p$

- B é uma relação inferencial binária entre duas sentenças s_i e s_k, da forma $(tipo, s_i, s_k)$
  - tipo é o tipo de relação inferencial entre as sentenças s_i, s_k : uma precondição ("Pre") ou uma pós-condição ("Pos").

Este modelo pode ser aplicado a qualquer domínio de informações através da capacidade de extensão inerente ao SIM. Uma descrição detalhada de como as bases do SIM foram construídas para a língua portuguesa pode ser encontrada em *InferenceNet.Br*(as bases semânticas do SIM para língua portuguesa) [PPFF10].

3.1.2 SIA

Para calcular o significado de uma expressão linguística no SIM é necessário descobrir e combinar a contribuição semântica de cada conceito utilizado nas sentenças a partir do conteúdo inferencial dos conceitos articulados dentro das sentenças. Para esta finalidade Pinheiro et AL

[Pin9a] desenvolveram um analisador semântico denominado *Semantic Inferentialist Analyser* (SIA), que raciocina sobre o conteúdo inferencial dos conceitos e sobre os padrões de sentença. Ele se propõe a ser um mecanismo de raciocínio generalizável que analisa de forma holística o conteúdo inferencial de conceitos e sentenças dentro da prática linguística [Pin9a]. O significado das palavras como em um dicionário é insuficiente para o entendimento da comunicação, além disto, é necessário identificar onde os significados são aplicáveis. O SIA reconhece que os conceitos interagem entre si e utiliza uma medida de relacionamento inferencial para auxiliar na descoberta desta interação na medida em que define quais pré ou pós condições são relevantes para a construção do significado da sentença.

O analisador semântico é o responsável pelo entendimento de uma sentença em linguagem natural. Ele implementa um mecanismo de inferência sobre G_c e G_s (bases conceituais e de sentenças-padrão) que gera uma rede inferencial de premissas e conclusões de sentenças. A rede inferencial de sentenças é definida por um grafo direcionado $G_n(V,E)$, onde $V =$ conjunto de sentenças S_i (vértices do grafo), $E =$ conjunto de arestas rotuladas que indicam o tipo de pré ou pós condição que relaciona uma sentença com outra. Como nos outros grafos há duas funções s e t que associam uma aresta a sua origem e seu destino respectivamente [PPFN09].

O SIA segue os seguintes passos [Figura 3.4]:

1. Passo 1: o texto de entrada é recebido já analisado morfossintaticamente e as suas sentenças constituintes são combinadas com as sentenças contidas na base de sentenças-padrão existentes no SIM (Match([sentenças])). São então gerados grafos G'_s a partir das associações das sentenças originais com as sentenças-padrão s_i da base de sentenças-padrão onde, $E' =$ conjunto de arestas e onde $s(e) = s_i$ e $V' = \{s_i\} \bigcup \{t(e), \forall\, e \in E'\}$.

2. Passo 2: são selecionados conceitos da base conceitual que provavelmente são os conceitos utilizados nas sentenças (Select_Concept([sentenças])). Para cada conceito é gerado o grafo G'_c que representa o seu conteúdo inferencial.

3. Passo 3: dentre os conceitos prováveis são definidos os conceitos que realmente estão presentes nas sentenças através dos seus conteúdos inferenciais (podem existir conceitos diferentes, porém homônimos) e as suas contribuições semânticas para com a sentença a que pertencem (Define_Concept([sentenças])). Tal contribuição é definida como um sub-grafo de G'_c onde $c \in G'_c$, utilizando-se uma medida de relacionamento inferencial.

4. Passo 4: as partes variáveis das sentenças-padrão em G'_s são instanciadas por elementos presentes nas sentenças originais (conceitos e elementos de ligação) (InstanceSentence([sentenças])).

5. Passo 5: a rede inferencial G_N^S é gerada com pré e pós condições de cada sentença s instanciada a partir das contribuições semânticas dos conceitos e das sentenças-padrão expressas em G_c' e G_s'. Pode-se também filtrar o resultado com base em algum objetivo de extração de informação (Update([sentenças],[objetivos])).

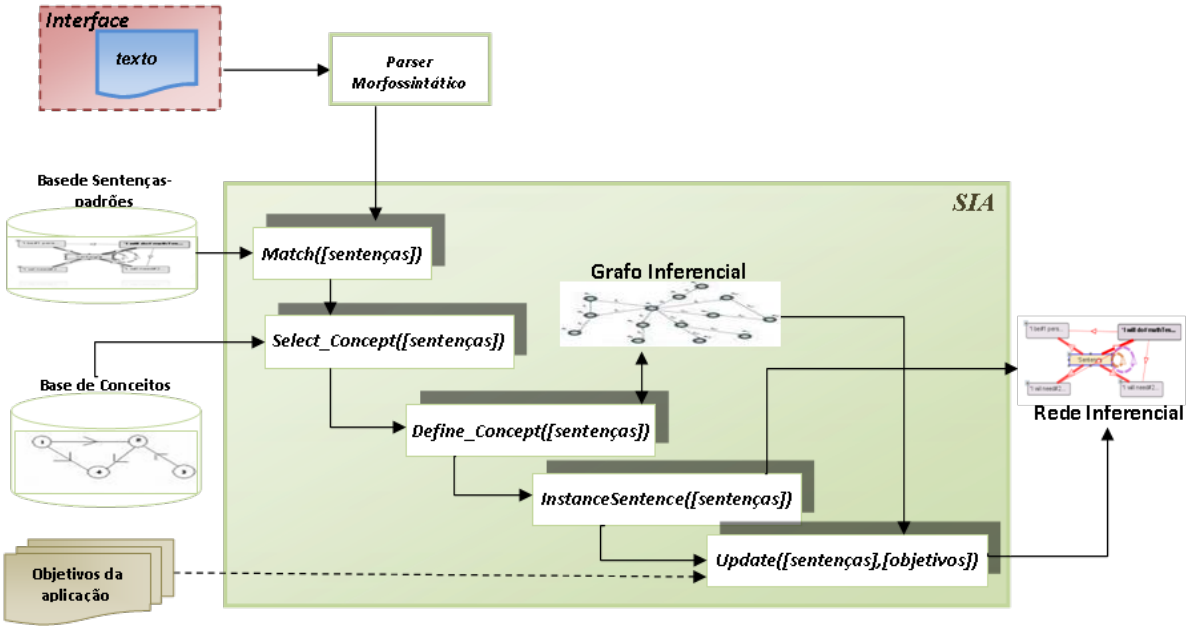


Figura 3.4: Modelo da execução do SIA

O raciocínio do SIA ao gerar a rede inferencial $G_N^S(V,E)$ é baseado na transcrição para a linguagem natural feita por Dummett a partir dos padrões dos conectivos lógicos definidos por Gentzen [PPFN09]. A definição de um conceito está na especificação de suas regras de introdução e eliminação. As regras de introdução são as precondições de uso de um conceito, as condições suficientes para a utilização de tal conceito. As regras de eliminação são as pós-condições de uso de um conceito, as condições necessárias para seu uso. O SIA implementa estas ideias em regras da seguinte forma:

1. Regras de introdução: $(I-c)$ define que o conceito c pode ser utilizado em uma sentença-padrão s se as precondições de c $(pre(c,p))$ forem satisfeitas no contexto linguístico.

$$\frac{pre(c,p),s}{s'(c)}(I-c)$$

2. Regras de eliminação: $(E-c)$ define que após o uso de um conceito c em uma sentença-padrão s', as pós-condições de c $(post(c,p))$ são satisfeitas no contexto linguístico.

$$\frac{post(c,p),s'(c)}{s'(c|p)}(E-c)$$

Para a composição de regras desta forma na rede inferencial G_N^S o SIA admite que se um conceito c é utilizado na sentença s' então as precondições do conceito foram satisfeitas. Tais precondições também são premissas da sentença s'. Por exemplo: o conceito $c =$ "porta" sendo utilizado em $s' =$ "João bateu na porta", uma premissa do conceito $c =$ "porta" é "éUm(porta, sólido)" que pode ser admitida como satisfeita já que o conceito foi utilizado na sentença s'. Uma precondição da sentença s' gerada no grafo G_N^S é: pre("porta éUm sólido", "João bateu na porta"). A mesma ideia se dá quando um conceito possui pós-condições e é utilizado em uma sentença s'. O conceito $c =$ "comer" possui a conclusão efeitoDe("comer", "ganhar energia"). A sentença $s' =$ "João comeu uma maçã" utiliza o conceito $c =$ "comer", logo uma pós-condição da sentença s' será criada no grafo G_N^S na forma post("João comeu uma maçã", "João ganhou energia").

O SIA consegue gerar as premissas e conclusões das sentenças através da combinação do conteúdo inferencial dos conceitos e do uso das sentenças-padrão expressos nos grafos G'_c e G'_s. As inferências são realizadas através do conteúdo inferencial dos conceitos e das sentenças e se tornam explícitas no grafo G_N^S servindo do base para respostas a questionamentos, argumentações, refutações e extração de informações. Como as bases do SIM se propõem à expressão de conhecimento dos mais variados domínios, sendo também capaz de expressar senso comum e conhecimento pragmático de linguagem natural, o SIA torna-se uma ferramenta poderosa capaz de gerar inferências sobre todo esse conteúdo.

A medida de relacionamento inferencial utilizada no SIA é baseada nas teorias semânticas em que o próprio SIM se fundamenta. A relação semântica entre conceitos não pode ser destituída dos seus usos em sentenças e deve basear-se no conteúdo inferencial compartilhado entre os conceitos. Um conceito utilizado em uma sentença está mais próximo inferencialmente de outro conceito quanto mais relacionadas estiverem suas premissas e conclusões. Com isso Pinheiro et AL [Pin9a] apresentam três heurísticas para definir as relações existentes entre premissas (ou conclusões) de conceitos distintos e possibilitar o cômputo da proximidade inferencial entre conceitos. É atribuído um peso para cada heurística $w_k (k = 1,2,3)$, que é definido por um parâmetro.

1. O conceito c_1 está relacionado com c_2 quando uma premissa ou conclusão de c_1 está diretamente ligada a c_2. Por exemplo, $c_1 =$ "vinho" e $c_2 =$ "álcool". Precondição de $c_1 =$ "contém(vinho, álcool)", $c_1 =$ "correr" e $c_2 =$ "cansar". Pós-condição de $c_1 =$ "efeitoDe(correr, cansar)".

2. O conceito c_1 está relacionado com c_2 quando possuem premissas ou conclusões de mesma relação com o conceito c_3. Por exemplo, $c_1 =$ "bebida", $c_2 =$ "comida" e $c_3 =$

"festa". Precondição de c_1 = "localizadoEm(bebida, festa)", precondição de c_2 = "localizadoEm(comida, festa)". c_1 = "faca", c_2 = "revólver" e c_3 = "ferir". Pós-condição de c_1 = "serCapazDe(faca, ferir)" e "serCapazDe(revólver, ferir)".

3. O conceito c_1 está relacionado com c_2 quando possuem premissas ou conclusões distintas que se referem ao mesmo conceito c_3. Por exemplo, c_1 = "faca", c_2 = "dedo" e c_3 = "ferir". Precondição de c_1 = "serCapazDe(faca, ferir)" e pós-condição de c_2 = "sofrerAção(dedo, ferir)".

A medida de relacionamento inferencial entre dois conceitos c_1 e c_2, $\theta(c_1, c_2)$, é calculada pelo somatório do produto dos pesos w_k pelas forças $\varphi(r_i)$ das pré e pós condições dos conceitos c_1 e c_2 que estão relacionados em uma das três formas mencionadas acima.

$$\theta(c_1, c_2) = \sum_k \sum_i \varphi(r_i) \times w_k$$

1. c_1 e c_2 são conceitos da base conceitual (vértices de G_{c_1} e G_{c_2})
2. r_i é a relação inferencial do conceito c_i com outros conceitos que combinaram em uma das três formas descritas acima (arestas de G_{c_i})
3. $\varphi(r_i)$ é a força da relação r_i definida semelhantemente a ConceptNet [LS04] como: $\varphi(r_i) = \log_2(f_i + 0.5(\sharp r_i + 1) \times wr_{rel}$
   (a) f_i é o número de vezes que a relação r_i aparece nas bases do SIM
   (b) $\sharp r_i$ é o número de vezes que a relação r_i foi inferida a partir de outras relações da ConceptNet
   (c) wr_{rel} é o peso do tipo da relação r_i (definido por parâmetro).
4. w_k é o peso da forma k de proximidade inferencial $(k = 1, 2, 3)$ (definido por parâmetro).

A medida de relacionamento inferencial é utilizada no SIA para:

1. desambiguação de conceitos homônimos
2. definição da contribuição semântica de um conceito c para com a sentença s'.
3. descarte de pré e pós condições dos conceitos que são irrelevantes para o significado da sentença.
4. seleção de premissas e conclusões na geração da rede inferencial da sentença (G_N^S) de acordo com os objetivos da aplicação.

3.2 Aplicação em WikiCrimesIE

O SIM está sendo utilizado em uma aplicação real de extração de informações. A aplicação WikiCrimesIE (WikiCrimes Information Extractor) [Pin9a] extrai informações sobre crimes a partir de notícias policiais publicadas em jornais da *web*, em língua portuguesa. Informações como tipo do crime, tipo de arma utilizada e motivo do crime exigem mais do que as abordagens usuais oferecem pois estas informações geralmente estão implícitas no texto.

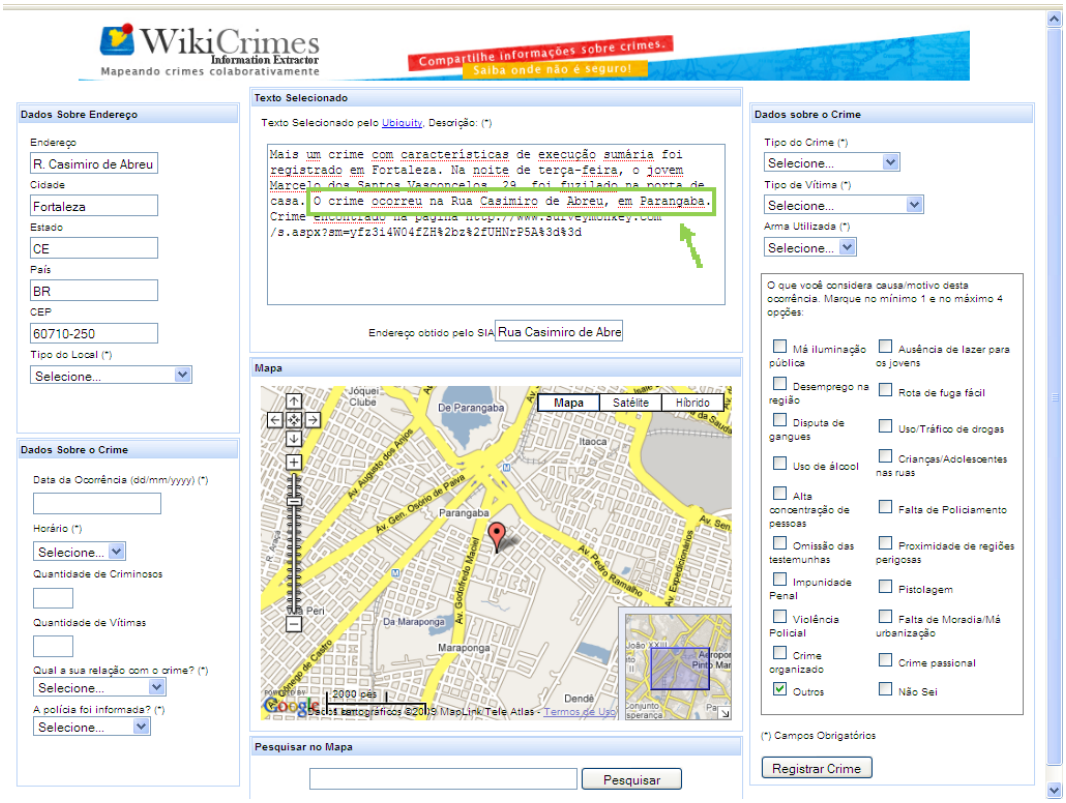


Figura 3.5: Utilização do SIM no WikiCrimesIE

A interface do *WikiCrimesIE* é apresentada na figura 3.5. Um texto previamente capturado de uma página da *Web* foi analisado pelo sistema com o objetivo de encontrar as informações sobre "local do crime". Na figura 3 vê-se que o local do crime "Rua Casimiro de Abreu, Parangaba" foi extraído da sentença marcada na figura S'="O crime ocorreu na Rua Casimiro de Abreu, em Parangaba".

No exemplo o sistema *WikiCrimesIE* foi executado com o objetivo de encontrar informações sobre "crime" e "local". Os seguintes passos foram observados:

- Foi identificado a sentença-padrão utilizada S = "X ocorrer em Y".
- Foram selecionados os possíveis conceitos da base conceitual utilizados em S' ("crime",

”ocorrer”).

- O sistema definiu os conceitos usados na sentença S’ (“crime”,”ocorrer”). O grafo inferencial foi gerado a partir dos conteúdos inferenciais dos conceitos e sentença-padrão encontrados. Por exemplo, a pós-condição de S: r_{pos} = (ehUm,Y,local(X)), foi gerada no grafo.

- Foi instanciada a sentença-padrão S com os elementos subsentenciais da sentença original S’: X=”o crime” e Y=”em a Rua_Casimiro_de_Abreu, em Parangaba”.

- A rede inferencial de S’ foi gerada com suas premissas e conclusões, a partir do grafo inferencial dos conceitos “crime”, “ocorrer” e da sentença-padrão S, filtradas pelos objetivos acima descritos.

  No exemplo, foi gerada a conclusão (ehUm,”em a Rua_Casimiro_de_Abreu, em Parangaba”,local(“o crime”)) como resposta ao objetivo, pois os conceitos “crime” e “local” da conclusão são melhor relacionados inferencialmente (pela medida de relacionamento inferencial do SIA) aos conceitos do objetivo proposto, do que os conceitos de outras premissas e conclusões.

4 Resolução de anáforas pronominais

Identificado como um processo integrante da interpretação do texto, a resolução de anáfora é essencial para a correta compreensão do texto. Isto se dá porque a anáfora é um recurso amplamente utilizado tanto na língua falada como na língua escrita. O processo de resolução inicia-se a partir do momento que o leitor começa a leitura do texto. A cada sintagma lido, o leitor percebe sua posição no texto e o identifica como saliente ou não, como um termo importante ou menos relevante dada a temática e sequência de raciocínio do texto. A partir de então alguns sintagmas vão ocupando local de destaque na memória do leitor. Ao se deparar com uma anáfora, o leitor identifica que o conceito que está sendo utilizado é, na verdade, o mesmo conceito relativo ao sintagma que ele guardou na memória em uma posição de destaque. Assim, as novas informações recebidas são relacionadas ao mesmo conceito descrito anteriormente.

Uma anáfora é toda forma de acrescentar informações sobre um conceito já existente no texto utilizando-se de sintagmas diferentes mas, que são identificados como relacionados ao mesmo conceito. As formas mais comuns de sintagmas utilizados pela anáfora são os sinônimos, a hiponímia e os pronomes. Neste trabalho focaremos no uso de pronomes pessoais para criação de anáforas, denominadas anáforas pronominais.

4.1 Pronomes pessoais

Um pronome é tradicionalmente conhecido como um termo que substitui um nome. Porém a tradição da cultura greco-romana carrega consigo um grande equívoco. O que conhecemos hoje como pronomes não são necessariamente substitutos, e quando o são podem substituir outros termos além de nomes. Por outro lado, há termos que substituem nomes e não são pronomes. Muitos linguistas recusam a definição “palavra que substitui o nome” e consideram que vários pronomes na prática são pro-adjetivos, pro-verbos e até pro-frases sem que com isso se obtenha uma agramaticalidade, ou seja, uma sentença que não é reconhecida como pertencente a uma determinada língua [Mon94]. Exemplos:

Quando "me" perguntam qual é minha religião, "eu" digo que é a cristã.

Ao tentar substituir os pronomes ("me" e "eu") por nomes (por exemplo nomes próprios) tem-se uma agramaticalidade.

Deveria sabê-"lo", tantas foram as vezes que eu li.

O clítico "lo" não se refere a apenas um nome ou expressão, mas a todo um trecho do discurso proferido até então.

Mesmo na hipótese que o sindicato pode recorrer, deverá fazê-"lo" através de advogado.

O mesmo clítico "lo" se refere aqui a um verbo, mas não o substitui propriamente já que o verbo "fazer" é que tem esta função.

Eu, geralmente no "ocaso", eu procuro estar conversando com alguém, para não ver o "crepúsculo".

O vocábulo "crepúsculo" substitui "ocaso", porém não se trata de um pronome.

A relação se estabelece entre o pronome e algo que está em mente no momento em que o pronome é enunciado. Essa relação tende a ser mais semântica do que sintática. Segundo Carrasco [*apud* Monteiro, 1994, p.30] a função de substituição dos pronomes podem ocorrer entre o pronome e todo o predicado da oração anterior, sendo assim, o pronome não substitui meramente um item lexical específico, mas todo um trecho do discurso.

Pronome é então um elemento linguístico que obrigatoriamente remete à algo na mensagem a qual ele pertence. Com isso entende-se que a definição de pronome como substituto de um nome é, em muitos casos, inadequada para definir este tipo de vocábulo. Porém é mais viável utilizar os preceitos já definidos pela tradição gramatical com a ressalva de que o significado etimológico deve ser deixado um pouco de lado e deve-se chamar a atenção às diferenças entre os vários pronomes existentes e seus empregos diferenciados no discurso [Mon94].

Segundo Monteiro [Mon94] existem três hipóteses para o significado pronominal. Uma hipótese é que certos vocábulos não possuam significado, incluindo neste grupo vários elementos relacionais (conjunções, preposições, etc.) e elementos substitutivos. A ideia de que um pronome seja vazio de significado é um tanto repulsiva quando se observa a prática linguística. O pronome pessoal "eu" significa a pessoa que está enunciando o discurso, sendo assim, ele não é vazio de significado. Por outro lado "ele" pode se referir a qualquer objeto, pessoa ou até outro predicado do discurso. Outra hipótese remete à teoria do signo formulada por *Peirce* diferenciando símbolos (nomes) e índices (pronomes) e entendendo que é necessária pelo menos uma análise sintática onde os significados só se definem no enunciado. Uma terceira hipótese coloca

o problema da significação pronominal no âmbito da pragmática, alegando que certos elementos linguísticos só são bem definidos se forem considerados os fatores situacionais e sociais que condicionam o ato de fala.

Há duas características dos pronomes que auxiliam no entendimento do significado pronominal: o uso dêitico e o uso correferencial. O uso dêitico, *a priori*, se dá quando um pronome se refere a um elemento existente fora do texto que está implícito e não é referenciado no texto por nenhum nome. No uso correferencial o pronome se refere a um elemento existente no texto. Mais adiante o significado de uso dêitico e correferencial será redefinido. Os exemplos a seguir apresentam o pronome "eu" em um uso estritamente dêitico e o pronome "ele" que pode ser entendido como dêitico ou correferencial.

"Eu" sou o dono do carro. (uso dêitico)

Apontou para a terceira cadeira e disse: "Ele" é o assassino! (uso dêitico)

"O homem" fugiu, mas depois "ele" foi capturado e preso. (uso correferencial)

Etimologicamente "dêitico" ou "dêixis" significa "apontar", "indicar". Esta indicação é relativa ao discurso e muda de acordo com a perspectiva adotada. Toda vez que o pronome "eu" é enunciado ele está relativo à pessoa que o enuncia. O uso dêitico de um pronome se dá quando este indica um objeto real do universo (lembre-se que o universo do discurso pode ser imaginário e os objetos reais deste universo pode conter seres imaginários e etc.).

"o referente de um dêitico é um lugar vazio que pode ser ocupado por todos os particulares capazes de estabelecer com o ato de fala a relação significada pelo dêitico em questão" [Lah79]

O objeto denotado pelo uso dêitico de um pronome se define face às relações que vigoram no momento do ato comunicativo. Uma vez que estas relações podem mudar, o objeto denotado também muda, porém o significado dêitico é fixo e consiste na indicação precisa das relações e circunstâncias que apontam para o objeto. O significado dêitico não consiste no objeto em si, mas na função que aponta para o objeto. Tal função é dependente das circunstâncias que envolvem o ato de fala incluindo fatores sociais e pessoais e pode ser entendida como um meio capaz de apontar para o objeto sem possuir todas as características formais de uma função.

Observe os seguintes exemplos:

"Eu" não sei tudo o que "tu" sabes.

"O barbeiro" trabalha bem quando "ele" tem uma navalha afiada.

Os referentes dos pronomes "eu" e "tu" só são conhecidos quando enunciados, já o pronome "ele" se refere a "barbeiro". Porém todos, "eu","tu" e "ele" possuem algum tipo de função

dêitica. Visto que o uso correferencial também se enquadra em um função dêitica redefiniremos que no primeiro exemplo temos um uso "dêitico situacional" e no segundo um uso "dêitico textual". O uso "correferencial", denominado a partir de agora também de "dêitico textual", remete a duas entidades que se referem a um mesmo objeto real do universo e que está descrito no discurso. Portanto o uso correferencial de um pronome também é dêitico porém, aponta para um elemento existente no próprio texto.

Os pronomes não possuem um significado determinado por eles mesmos. Pelo contrário, o seu significado é exatamente de que ele depende de uma referência que deve ser resolvida em cada discurso em que aparece. Esta resolução é guiada por traços morfossintáticos como gênero e número.

Os exemplos a seguir ilustram o uso de variáveis semânticas que auxiliam no entendimento dos pronomes.

"Você" está cansada.

"João" brigou com Arnold Schwarzenegger. "Ele" "se" machucou bastante.

"Ele" é o cara que "eu" conheci na festa.

"Qualquer" aluno gosta quando "seu" trabalho é julgado melhor que o trabalho dos outros.

Nos exemplos acima os pronomes pessoais "eu" e "ele" são variáveis semânticas cujos valores serão atribuídos em função da dependência no contexto linguístico ou extralinguístico. Já os pronomes reflexivos "se" e possessivo "seu" são também variáveis semânticas e seus valores covariam com o valor dos seus antecedentes, sejam os nomes próprios "João" e "Arnold Schwarzenegger" ou o sintagma quantificado "qualquer aluno" [Mul01]. No último caso, se "Jorge é aluno", então "Jorge gosta quando o trabalho de Jorge é julgado melhor que o trabalho dos outros".

Em seu uso anafórico um pronome possui sua referência no próprio contexto linguístico, seja em sentenças anteriores ou posteriores. Estudos em Teoria Gramatical apontam para o fato de que esta distinção é na verdade um caso específico do mesmo fenômeno, de alguma maneira ambos se referem a uma entidade determinada pelo contexto que está altamente saliente no momento da atribuição de valores ao pronome [Mul01].

Por exemplo, o pronome "ele" é uma variável, pois podem ser atribuídos valores diferentes ao pronome em contextos diferentes, logo:

$ele^{Jorge} = Jorge$ e $ele^{Carlos} = Carlos$ então $ele^{Jorge} \neq ele^{Carlos}$

(ele^{Jorge} é o pronome "ele" no contexto onde lhe é atribuído o valor "Jorge" e ele^{Carlos} é o pronome "ele" no contexto onde lhe é atribuído o valor "Carlos").

As relações anafóricas são guiadas por limites estruturais. Como exemplo tem-se os pronomes pessoais e reflexivos, onde no contexto em que cabe a ocorrência de um, o outro é gramaticalmente proibido. Poder-se-ia até, a princípio, afirmar que um pronome reflexivo tem um antecedente dentro da oração em que ele aparece e o antecedente possui posição superior (por exemplo, a de sujeito), enquanto que o pronome pessoal tem um antecedente em uma oração distinta.

Em um contexto onde o pronome se refere a "Zelda":

- *Zelda se adora.*
- *Zelda adora ela.**
- *Zelda acha que Carlos se adora.**
- *Zelda acha que Carlos adora ela.*

Em um contexto onde o pronome não se refere a "Zelda".

- *Zelda se adora.**
- *Zelda adora ela.*
- *Zelda acha que Carlos se adora.*
- *Zelda acha que Carlos adora ela.**

Onde as sentenças marcadas com * se referem a sentenças agramaticais.

Dessa forma vê-se que a relação de correferência é determinada pelo contexto, já a relação de variável ligada é restrita pela estrutura da sentença. A relação de um sintagma quantificado e um pronome depende de que o sintagma *c-comande* o pronome. C-comando é uma relação de superioridade estrutural. Por exemplo, um sujeito c-comanda todos os sintagmas dentro da sua oração. Esta relação só ocorre dentro de uma sentença, fazendo da sentença o escopo do sintagma quantificado [Mul01].

Um fato que ocorre fora desses padrões é o visto no seguinte exemplo:

Apenas um congressista admira Kennedy. Ele é muito jovem.

Um sintagma quantificado deveria ter escopo apenas dentro da oração. Quando opta-se por tratar este sintagma quantificado "Apenas um congressista" como um sintagma referencial, surgem problemas com outras características dos sintagmas referenciais como nos exemplos a seguir:

João veio ontem de manhã.=> João veio ontem.

Apenas um congressista veio ontem de manhã. =/> Apenas um congressista veio ontem.

João está nesta sala e João está na sala ao lado. (contradição)

Apenas um congressista está nesta sala e apenas um congressista está na sala ao lado.

Conclui-se então que este sintagma não pode ser igualado a um sintagma referencial. Por outro lado, pode-se admitir que alguns sintagmas quantificados tenham seu escopo para fora da oração. Porém neste caso pode-se obter uma interpretação incorreta da relação de ligação de variável, concluindo que "apenas um congressista" ao mesmo tempo "admira Kennedy" e "é jovem". Apesar de parecer uma interpretação razoável, ela admite que dois "congressistas" "admirem Kennedy" e apenas um seja "jovem".

Evans [apud Muller, 2001] propõe que nesta situação há um novo tipo de pronome que se refere a "o congressista que admira Kennedy" do tipo E-type. Este tipo de pronome é criado a partir da sentença anterior e é equivalente a um sintagma nominal definido.

Apenas um congressista admira Kennedy. O congressista que admira Kennedy é muito jovem.

Tem-se então que um sintagma quantificado pode servir de antecedente para um pronome referencial, porém carregando consigo um predicado que o define (ou restringe). Existir um predicado que define o sintagma quantificado parece ser uma exigência para a existência do pronome E-type, como no exemplo a seguir:

Nenhum menino que repetiu de ano foi convidado para a festa de Maria. Eles (os meninos que repetiram de ano) reclamaram.

Este pronome E-type é, portanto, uma descrição definida que traz consigo um predicado definido pelo contexto. Este predicado pode aparecer explícito ou implícito nas orações anteriores.

4.2 Anáforas

A função básica dos termos dêiticos (situacional e textual ou correferencial), como os pronomes, é o de identificar o referente. Sendo assim, eles remetem a uma fonte que lhes preenchem o significado. Esta fonte pode ser a situação em que o enunciado é proferido ou o próprio enunciado. Se a fonte for o próprio enunciado então temos o fenômeno denominado de anáfora. O mecanismo da anáfora envolve transferência de noções essencialmente dêiticas de espaço e localização, porém relativas à existência textual e não à existência real. O referente pode até não estar presente no texto, mas deve ser inferido por meio de um termo antecedente que o introduz ou identifica. Assim, uma anáfora também pode ser definida como um dêitico textual [Lyo77].

Segundo Teyssier [*apud* Monteiro, 1994, p.53] as diferenças entre dêitico situacional e anáfora podem ser entendidas pela fonte de informações e pelo tipo de relação com o referente. O vocábulo dêitico situacional busca informações no contexto linguístico e extralinguístico estando ligado ao referente de forma direta, indicando-o. Já o vocábulo anafórico tem sua fonte de informações no interior do discurso e está ligado ao referente por substituição do termo anafórico por um sintagma nominal previamente existente. Para Lyons [Lyo77] o vocábulo anafórico é um tipo de vocábulo dêitico restrito a informações textuais.

Diferenças tão sutis causam constantemente o emprego de vocábulos dêiticos como anafóricos e vice-versa, sendo possível até o uso de vocábulos com valor ambíguo ou de ambos ao mesmo tempo. Exemplos destes usos são vistos a seguir:

Nunca havia visto o mar até o ano passado. Aquilo é simplesmente maravilhoso.

O termo “aquilo” se refere tanto a “o mar” (anáfora) quanto à visão do mar que o locutor desfrutou (dêitico situacional).

Quem assistir ao filme dirá que já viu aquilo.

O valor é ambíguo pois “aquilo” pode estar relacionado a “o filme” (anáfora) ou a alguma situação apontada pelo locutor (dêitico situacional), mas não aos dois ao mesmo tempo.

Um grupo de palavras construídas em volta de um substantivo é denominado um sintagma nominal. Quando o sintagma nominal é precedido por um artigo definido então chama-se sintagma nominal definido [dF05]. Este tipo de grupo em particular é bastante utilizado para referenciar um conceito que está sendo repetido durante um discurso. Os pronomes, em geral, são utilizados para substituírem os sintagmas nominais.

Uma anáfora, então, é um termo que não introduz um novo conceito no discurso, mas se refere a um conceito antigo disponibilizando informações adicionais acerca do mesmo. Podemos

encontrar anáforas sob várias formas, entre elas temos: [dF05]

- Anáfora nominal: formado por um sintagma nominal que se refere a outro sintagma nominal anterior. Pode possuir uma relação de igualdade entre os sintagmas (sinonímia) ou uma relação de especialização (hiponímia). Exemplo: *José comprou um "cachorro". O "animal" faz muito barulho.* onde o termo "animal" pode ser substituído por "cachorro".

- Anáfora pronominal: formado por um pronome que substitui um sintagma nominal anterior. A relação é sempre de sinonímia. Exemplo: *José comprou "o cachorro". "Ele" faz muito barulho.* [Cha07], onde o termo "Ele" pode ser substituído por "o cachorro".

- Anáfora conceitual, anáfora pronominal genérica, anáfora esquemática ou anáfora indireta: Semelhante aos anteriores, porém este não possui concordância de gênero ou número entre a anáfora e o antecedente anafórico (termo ao qual a anáfora se refere). Isto se dá porque não há uma relação sintática entre os termos, mas somente uma relação semântica, de significado. Exemplo: *O crime organizado tem uma estrutura de poder e muito dinheiro. "Eles" têm armamento de última geração.* [dS04]. Aqui o termo "Eles" não possui nenhum termo anterior sintaticamente adequado que o substitua, porém semanticamente o conceito "armamento" tem relação com "crime organizado" fazendo com que o pronome "Ele" possa se referir a "crime organizado" e, neste caso, a "os integrantes do crime organizado".

- Catáfora: Espécie de anáfora onde o antecedente anafórico aparece no texto após a anáfora. Exemplo: *O pássaro seguia-"o" pelo caminho, reparou "o moço".* O pronome oblíquo "o" se refere a um termo posterior "o moço".

- Elipse: Neste caso de anáfora, o termo que representa a anáfora não aparece explicitado no texto, ao invés disto, fica subentendido que está oculto uma anáfora por causa da referenciação a um termo anterior. Exemplo *"João" saiu. Porém "ϕ" volta logo.* onde entende-se que "ϕ" é o local onde está oculto o termo "João".

Paduceva [*apud* Monteiro, 1994, p.54] afirma que a anáfora também consiste na relação entre dois termos que remetem ao mesmo referente. Portanto esta relação não se dá necessariamente com o uso de um pronome substitutivo. A sentença abaixo exemplifica o caso onde um sintagma nominal substitui outro sintagma nominal.

"O avião" começou a cair quando o motor da "aeronave" explodiu.

A seguir alguns exemplos ilustram a complexidade da tarefa de descobrir a correferenciação.

Exemplo 1:

O rapaz que caiu era o garçom do restaurante.

A correferência se dá entre "o rapaz que caiu" e "o garçom do restaurante". Mas não existe nenhum traço gramatical que indique isto. Pois, mantendo a estrutura sintática, têm-se:

O rapaz que caiu conhecia o garçom do restaurante.

Neste caso não existe mais correferência. Pode-se concluir então que, o mecanismo utilizado pelo usuário da língua que é capaz de criar ou identificar relações correferenciais não é meramente sintático. *"Percebe-se que informações estritamente morfossintáticas são insuficientes para clarificar as correferências. As correferências e as anáforas, por consequências, são fenômenos com características semânticas e assim devem ser entendidos."* [Mon94].

Exemplo 2:

Encontrei uns amigos e eles me falaram de você.

Encontrei uns amigos e uns amigos me falaram de você.

Voltando-se para a questão da substituição no uso da anáfora como correferência, em vários casos a substituição por termos correferentes altera o significado da sentença. Neste caso a substituição do termo anafórico pelo seu antecedente influenciou uma mudança no significado da sentença. Na primeira sentença entende-se que "os amigos encontrados" foram os mesmos que "me falaram de você". Na segunda sentença "os amigos encontrados" não são necessariamente os mesmos que "me falaram de você".

Exemplo 3:

O helicóptero tem sido muito utilizado e em breve será o transporte mais viável.

Muitas vezes o professor se dá conta que ele só exigia o processo mental de memória do aluno.

Além da dificuldade em estabelecer uma relação de substituição sem alteração de significado, também ocorre o fato de que existem certas restrições para a ocorrência de uma anáfora. Os termos correferenciais do exemplo 3 possuem uma boa conexão. Porém ao tratarmos como correferenciais supomos que eles remetem ao mesmo referente, então seria o caso de trocarmos as ocorrências de ambos:

O transporte tem sido muito utilizado e em breve será o helicóptero mais viável.

Muitas vezes ele se dá conta que o professor só exigia o processo mental de memória do

aluno.

Na primeira sentença alterada perde-se a conexão entre os termos e a sentença perde o significado. Na segunda alteração a anáfora inicial é desfeita e é criada uma relação do pronome a um termo anterior à sentença. Isto se explica porque o termo antecedente deve ser mais restrito que o termo consequente para se estabelecer uma anáfora. O fenômeno linguístico da anáfora deve ser explicado por uma teoria que relacione o texto e o contexto em função dos dados referenciais que são introduzidos no discurso. Deste modo, cabe à anáfora repetir uma identificação a algum objeto já introduzido anteriormente através de uma nova expressão, o que ocorre sem que haja necessariamente um casamento perfeito entre os referentes do antecedente e do termo anafórico. Assim, a anáfora pode não ser um caso específico da correferencialidade apesar de estar bem próxima a ela como no exemplo a seguir:

Um aluno que tenha escrito sua "tese" em agosto será considerado mais aplicado do que um que "a" tenha feito em novembro.

Este caso de anáfora não nos remete a correferência, pois o termo "tese" e "a" não se referem a um mesmo objeto. Porém mesmo não havendo uma correferência nota-se que os termos se referem a objetos de mesma natureza ou mesmo tipo e isto basta para a existência da anáfora.

Dentre as formas de anáforas apresentadas, a anáfora pronominal é a mais utilizada, provavelmente porque ela evita redundância, a qual ocorre quando a anáfora é estabelecida pela repetição do mesmo vocábulo ou por sinonímia ou hiponímia, por exemplo.

Para a gramática gerativa, a anáfora pronominal é um processo onde um elemento linguístico atua sobre outro, sendo ambos correferenciais e transformando o segundo em pronome. Para a teoria transformacionalista, a anáfora pronominal consiste de transformações que envolvem a omissão e pronominalização de um item lexical, que na estrutura profunda é idêntico e correferencial ao seu antecedente.

Outra classificação define que a anáfora pode ser regressiva se o pronome aparece antes do termo que ele substitui ou pode ser progressiva aparecendo primeiro o termo que o pronome substitui, denominado termo antecedente, e depois o pronome, denominado termo anafórico, ou simplesmente antecedente e anáfora respectivamente. A anáfora progressiva é bem mais frequente na prática linguística que a anáfora regressiva. [Mon94]

Exemplo de anáfora progressiva:

O menino era lindo. Ele é a cara da mãe.

Exemplo de anáfora regressiva:

Segundo ele mesmo, Carlos não irá trabalhar amanhã.

A certeza de que ele é abençoado por deus é a base do casamento.

Outra denominação proposta por Halliday e Hasan [HH76] renomeia os termos vistos até aqui. O mecanismo de coesão textual da teoria da referência que define a relação de um termo com um objeto da situação discursiva é denominado por "relação situacional" ou "exófora". Por outro lado, ao mesmo tipo de mecanismo utilizado para definir a relação de um termo com outros presentes no discurso que apontam para o mesmo referente denominou-se "referência intratextual" ou "endófora". A endófora por sua vez se divide em "anáfora" (anáfora progressiva) e "catáfora" (anáfora regressiva). Estes mecanismos são importantes para a obtenção de coerência e coesão do texto, pois eliminam redundâncias e tornam o texto menos suscetível a ambiguidades, orientando os interlocutores quanto aos esquemas correlacionais em que os termos presentes no texto estão envolvidos.

Outro fator importante para o entendimento da anáfora é que nem sempre o termo anafórico se realiza no discurso, embora esteja presente mentalmente. Os exemplos a seguir ilustram o caso onde a anáfora existe mas está oculta:

O fundo da piscina deu defeito e tiveram que esvaziar ϕ.

Nós temos roupas sociais, eu mando fazer ϕ *no alfaiate.*

Eu pego as revistas atuais e leio ϕ.

O homem é um caminhante, ϕ *é um viandante.*

Onde ϕ representa o local onde o termo anafórico foi apagado.

Estas construções, comuns em nossa língua, ultrapassam a pronominalização de um termo e chegam ao seu apagamento. O fenômeno de apagamento não se limita a algumas funções sintáticas, podendo ocorrer inclusive com o sujeito [Mon94].

A utilização de um pronome anafórico ou opção pelo seu apagamento não é algo banal. Embora em muitos casos ambos funcionem perfeitamente, em outros casos um ou o outro causa ambiguidade no entendimento da sentença. Como vê-se nos exemplos a seguir:

O advogado é que vai dizer ao seu constituinte até onde ele poderá chegar.

O advogado é que vai dizer ao seu constituinte até onde poderá chegar.

No primeiro exemplo o pronome "ele" pode se referir tanto a "advogado" como a "consti-

tuinte" introduzindo uma ambiguidade. A opção pelo seu apagamento no segundo exemplo dá uma visão mais clara de que o pronome "ele" apagado se refere ao advogado.

Abaixo são citados ainda outros exemplos sobre o uso do pronome ou a opção pelo seu apagamento:

O direito está inserido na própria realidade social, porque ele é fruto da interação social.

O direito está inserido na própria realidade social, porque é fruto da interação social.

O direito está inserido na própria realidade social, porque ela é fruto da interação social.

A primeira sentença ilustra o caso onde o pronome "ele" sujeito da oração subordinada "porque ele é fruto da interação social" referencia o sujeito da oração principal "o direito". Pode-se ver na segunda sentença que só é possível o apagamento do sujeito na oração subordinada "ele" porque o sujeito desta é correferencial ao da principal. Se não for correferencial ao sujeito da oração principal então deve ocorrer a pronominalização, o que ocorre na terceira sentença. Se o pronome "ela" da terceira sentença for apagado, ele deixa de referir-se a "realidade social" e passa a referir "o direito".

Enfim, Monteiro e Levinson apontam para a necessidade das construções anafóricas de serem interpretadas menos em função de regras sintáticas e mais em função de fatores discursivos e pragmáticos[Mon94], [Lev87].

4.3 Resolução de anáforas

O processamento de linguagem natural é historicamente dividido em níveis. Estes níveis vão avançando a partir de partes atômicas até incluírem todo o contexto em que um determinado ato de fala está inserido [CK92].

O primeiro nível, denominado de morfológico, analisa os termos de uma sentença individualmente e atribui ao termo sua pertinência a determinada classe gramatical de palavras. As classes podem ser abertas incluindo os substantivos, verbos, adjetivos e advérbios, sendo constituídas por uma quantidade possivelmente ilimitada de elementos e sofrendo frequentes mudanças com a prática linguística, ou podem ser classes fechadas como pronomes, preposições e conjunções, formadas por uma quantidade pequena de palavras e que muito dificilmente sofrem modificações. O trabalho neste nível de processamento está em identificar a classe gramatical da palavra.

O segundo nível do processamento é o sintático no qual a relação entre as classes de palavras

existentes em uma sentença e a ordem em que são utilizadas são analisadas para verificar se tal sentença está bem formada, segundo a gramática da língua, e qual a função sintática das partes da sentença. Analisadores morfossintáticos já são capazes de atribuir a um termo, além da classe gramatical, variáveis como gênero, grau, número, modo e funções como: sujeito, verbo, predicado, objeto direto, complemento nominal e etc. Estes analisadores são muito importantes para o entendimento de textos em linguagem natural e constituem a base para se compreender um texto.

Os próximos níveis de processamento são o nível semântico e o nível pragmático. O nível semântico trata de atribuir significado à sentença e o nível pragmático trata de sua função em um discurso. Teóricos pragmáticos como o segundo Wittgenstein e Searle, criticam a separação destes níveis, visto que o significado deve ser entendido como parte integrante da função da sentença em um discurso, não sendo possível atribuir significado sem levar em consideração o contexto e os propósitos com os quais uma sentença é proferida [Pin9a]. A maioria dos processos semânticos realizados atualmente atribuem um significado literal às partes de uma sentença em questão, sem admitirem a existência de usos particulares e fazendo com que o significado da sentença seja a junção dos significados das suas partes. Por outro lado, o processamento no nível pragmático busca obter o significado "não literal", tal como o homem entende ao ouvir uma sentença, fazendo o foco de sua análise ser o discurso e não as partes das sentenças.

No nível do processamento pragmático a tarefa de resolução de anáfora assume um papel importante na obtenção de elementos contextuais, no entendimento das relações entre as expressões e no descobrimento de conteúdo implícito [de 04]. Sem a resolução de anáfora, uma sentença onde ocorre uma anáfora não possui relação com a sentença onde ocorre o antecedente anafórico.

As ideias que norteiam os trabalhos atuais em resolução de anáfora tiveram início com a teoria de Grosz em 1977 que identificou um foco global, relacionado com os sintagmas nominais definidos e um foco local, relacionado com os pronomes e elipses. Ele descreve um mecanismo de focalização que utiliza uma base de dados semântica e identifica as entidades mais salientes no texto limitando as possibilidades de antecedentes de uma anáfora [dF05]. Este trabalho influenciou vários outros, nomeadamente a Teoria do Foco e a Teoria da Centragem.

A Teoria do Foco propõe uma separação entre encontrar os possíveis antecedentes de uma anáfora e validá-los. Baseado no foco, que é a entidade mais saliente em uma sentença, faz-se uma restrição no domínio de busca pelo antecedente anafórico. Dois focos são utilizados, um local e um global, que são guiados pela informação temática, informações gramaticais e sobre quais as entidades mais salientes na frase anterior. Melhorias propõem o uso de listas de focos

possibilitando a resolução de anáforas com seus antecedentes em uma mesma frase [dF05].

A teoria da Centragem lista todas as entidades existentes nas frases e lista todas as entidades em foco em cada frase. A partir destas listas as entidades são ordenadas pelas suas funções sintáticas, estabelecendo a preferência que cada entidade tem de ser o antecedente de uma anáfora. Utiliza-se uma ordem baseada em relações gramaticais sub-categorizadas pelo verbo principal, quase sempre coincidindo com a ordem em que aparecem no texto. Melhorias propõem outras formas de ordenação baseadas nas funções das entidades [dF05].

Para o entendimento da relação anafórica existem abordagens puramente sintáticas, abordagens semânticas e abordagens que fazem incursões pelo domínio da pragmática.

Ao definir anáfora como "dois termos que se associam numa estrutura frasal em função de certas regras" torna-se possível formular regras sintáticas que auxiliem na resolução do processo anafórico em questão [Grinder & Postal *apud* Monteiro, 1994, p.80]. Porém segundo Jackendorff [*apud* Monteiro, 1994, p.80] a relação entre os termos se dá pelo significado deles e, portanto, necessita de uma teoria da correferência, tornando o processo todo fundamentado na semântica. A natureza semântica do processo da anáfora se evidencia no fato de frequentemente o termo anaforizado remeter a outro que não está presente na mesma oração e até fora dos limites do período.

"A anáfora não se limita ao plano da oração ou da frase. Com frequência ela o ultrapassa. E um anafórico pode muito bem estabelecer uma conexão semântica entre duas frases que não se ligam sintaticamente."[Tesniere *apud* Monteiro, 1994, p.80]

Paduceva [*apud* Monteiro, 1994, p.81] descreve que as anáforas possuem um duplo comportamento:

1. Sintático, quando o antecedente é substituído por um pronome ou por seu núcleo.
2. Semântico, quando o termo anafórico é complementar ao antecedente, acrescentando informação sobre ele. Neste caso há uma proximidade entre o significado do antecedente e da sentença a que o pronome pertence.

Teixeira propõe 4 níveis de interpretação para a anáfora [Tei77]:

1. Oracional: o antecedente e o anafórico estão presentes na mesma oração.
2. Periodal: o antecedente e o anafórico se encontram em orações distintas, mas dentro do mesmo período composto por coordenação ou subordinação.

3. Textual: o antecedente está distante do anafórico, em outro período.

4. Pragmático: o antecedente está implícito e é inferido pela situação discursiva.

Sag e Hankamer esquematizam da seguinte forma [SH77]:

1. Superficial: o antecedente está presente no texto e controla o anafórico sintaticamente se relacionando com ele por uma identidade referencial.

2. Profunda: o antecedente não está explícito no texto e controla o anafórico pragmaticamente se relacionando com o anafórico através do contexto situacional ou discursivo.

Em face aos problemas encontrados com a ausência do antecedente, Levinson [Lev87], propõe que a anáfora seja um tema de ordem semântica ou, mais ainda, que deve ser resolvida no âmbito pragmático da linguagem. Para ele os processos anafóricos não se limitam ao uso de regras sintáticas, pelo contrário, eles levam em consideração outros princípios, como os que regem a conversação. No âmago da anáfora está o conceito de referenciação, sendo impossível descartar os dados extralinguísticos que introduzem informações complementares à compreensão do discurso.

A anáfora possui função de manter a coerência e coesão do discurso, eliminando ou substituindo nomes em função da economia de expressão e da clareza evitando repetições desnecessárias [Mon94]. A conexão criada pela anáfora entre o antecedente e o anafórico é determinada por particularidades na estrutura sintática e semântica do discurso e há regras que auxiliam na resolução do processo por parte do receptor. Portanto é previsível que as implicações da anáfora sejam esclarecidas remetendo-se ao próprio discurso e sua análise quanto ao que está explícito (sintático e semântico) e implícito (semântico e pragmático).

Pode-se afirmar que existe uma vasta gama de técnicas e recursos que foram propostos para a resolução de anáforas. Nesta, incluem-se aprendizado de máquina supervisionado e não supervisionado, *corpus* anotado, método estatístico, árvore de decisão, características específicas do domínio, características lexicais, sintáticas, morfológicas, semânticas e posicionais, redes semânticas, heurística, estrutura sintática com pesos e implementação multi-agentes. Algumas abordagens se restringem a classificar os termos como anafóricos ou não, ou seja, se são conceitos novos no discurso ou são referências a conceitos anteriormente proferidos. Poucos trabalhos são encontrados baseados na língua portuguesa. Os últimos trabalhos da área apontam para uma tendência de utilização de informações semânticas, mostrando que melhores resultados podem ser obtidos desta maneira [dS08] [dRAdOCT$^+$05] [Cha07] [Coe06] [CCV05] [GGV03] [Lef01] [Par97] [RPFV01] [San00].

A seguir serão explanados alguns trabalhos dentre os mais influentes na área de resolução de anáforas.

4.3.1 Lapin & Leass

O algoritmo de Lapin & Leass, ou simplesmente RAP (*Resolution of Anaphora Procedure*), originalmente apresentando em [Lea94] e continuamente utilizado e atualizado em trabalhos como [Coe06], processa os casos anafóricos pronominais de terceira pessoa onde os antecedentes aparecem na mesma sentença ou em sentenças distintas. O algoritmo é composto fundamentalmente por:

- Restrições intra-sentenciais sintáticas que restringem o domínio dos sintagmas nominais candidatos a antecedentes anafóricos através da utilização de dados sintáticos.
- Restrições morfológicas aplicadas sobre os sintagmas nominais quanto a concordância com o pronome em gênero, número e pessoa.
- Algoritmo de ligação que identifica os candidatos a antecedentes de um pronome reflexivo dentro de uma mesma sentença.
- Função que atribui as características preferenciais de cada sintagma nominal como paralelismo sintático com o pronome, papel gramatical e outros.
- Função de decisão que seleciona o antecedente apropriado a partir do domínio de sintagmas nominais candidatos a antecedentes.

É calculado o valor de saliência de cada candidato a antecedente de pronomes não reflexivos ou recíprocos [Coe06]. O valor de saliência de um sintagma nominal é formado pelo somatório de todos os fatores de saliência relacionados a ele e representa o quanto o sintagma nominal está evidenciado na sentença e sua probabilidade de ser referenciado posteriormente por um pronome. Para o fator de saliência é utilizado um sistema de pesos baseados em informações sintáticas.

Os pronomes reflexivos ou recíprocos são tratados pelo algoritmo de ligação, outros pronomes de terceira pessoa são tratados pelo filtro sintático que identifica candidatos a antecedente que não admitem correferência e elimina-os. Os candidatos que restarem passam pelo cálculo de saliência que lhes atribuem valores. O candidato de maior valor de saliência é escolhido. Em caso de empate, o mais próximo do pronome é escolhido. Os valores iniciais de saliência dos sintagmas nominais estão definidos na tabela 4.1 e serão explicados a seguir.

Tipo	Fator de saliência
Sentença Atual	100
Sujeito	80
Construção existencial	70
Objeto direto	50
Objeto indireto	40
Ênfase não adverbial	50
Sintagma nominal não contido	80
Paralelismo sintático	35
Sintagma posterior	-175

Tabela 4.1: Valores iniciais dos fatores de saliência

- Sentença Atual : 100

  A atribuição do fator "Sentença Atual" é aplicada a cada sintagma nominal durante o processamento da sentença à que ele pertence.

- Sujeito : 80

  O fator "Sujeito" é aplicado quando o sintagma faz parte do sujeito da oração:

  "A cadeira" é bonita.

- Construção Existencial : 70

  É atribuído ao sintagma nominal próximo a um verbo existencial como "existir", "haver" e que se relaciona com este verbo.

  Havia "um bêbado" no hospital.

- Objeto Direto : 50

  Fator de saliência atribuído ao sintagma nominal que possui a característica sintática de objeto direto.

  A mulher quebrou "a perna" na escada.

- Objeto Indireto : 40

  Atribuído ao sintagma nominal que possui função sintática de objeto indireto.

  A mulher quebrou a perna "na escada".

- Ênfase não Adverbial : 50

  Este fator é atribuído ao sintagma que não se encontra em uma locução adverbial preposicionada demarcada.

  Pessoas morrem por causa do "álcool".

O sintagma “álcool” receberá este fator de saliência, porque mesmo contido em uma locução adverbial preposicionada “por causa do” ela não está demarcada.

Em frente ao “banco”, João foi assaltado.

O sintagma “banco” está em uma locução adverbial preposicionada demarcada e não recebe este fator.

- Sintagma Nominal não Contido : 80

  O fator é atribuído a entidades que não estão contidas em sintagmas nominais. O fator é atribuído somente aos sintagmas nominais mais externos.

  “O ““computador” de bordo” do “carro”” está quebrado.

  Encontramos os sintagmas “O computador”, “O computador de bordo”, “carro” mas somente “O computador de bordo do carro” receberá este fator de saliência.

- Paralelismo Sintático : 35

  Os sintagmas que possuem a mesma função sintática que a anáfora pronominal a ser resolvida recebem este fator.

  “Pedro” matou João. Ele é um assassino.

  “Pedro” recebe o fator paralelismo sintático porque é o sujeito da oração, assim como o pronome “Ele”.

- Sintagma Posterior : -175

  Quando um sintagma aparece após a anáfora pronominal em questão ele recebe este fator de penalidade.

  Ele maltratava o “cachorro”.

Todos os fatores de saliência são atribuídos integralmente aos sintagmas de uma sentença no momento em que a sentença é processada. A cada nova sentença, os sintagmas das sentenças anteriores têm os seus valores reduzidos pela metade. Os valores são atribuídos aos sintagmas nominais para cada pronome analisado e são cumulativos, ou seja, um mesmo sintagma pode receber fatores de saliências diversos enquanto se enquadrar no domínio de cada fator. Quando um outro pronome começa a ser processado os valores de todos os sintagmas são reiniciados. O pronome resolvido é tratado como um sintagma nominal e passa a somar pontos para o seu antecedente anafórico.

O algoritmo é executado no texto da forma ilustrada no exemplo a seguir:

Pedro comprou o carro. Ele estava feliz.

A primeira sentença é analisada e encontram-se os sintagmas nominais "Pedro" e "carro". O sintagma nominal "Pedro" recebe os seguintes fatores de saliência: sentença atual (100), sujeito (80), sintagma nominal não contido (80), ênfase não adverbial (50) (total 310). O sintagma nominal "carro" recebe os seguintes fatores de saliência: sentença atual (100), objeto direto (50), sintagma nominal não contido (80), ênfase não adverbial (50) (total 280) [Tabela 4.2].

Tipo de saliência	**Pedro**	**carro**
Sentença Atual	100	100
Sujeito	80	0
Objeto direto	0	50
Sintagma nominal não contido	80	80
Ênfase não adverbial	50	50
Total Atual	310	280

Tabela 4.2: Valores dos fatores de saliência após a primeira sentença

Passa-se então para a próxima sentença e os valores de saliência são reduzidos pela metade ("Pedro" 155) ("carro" 140). O pronome "Ele" é encontrado e "Pedro" recebe outro fator de saliência, pois "Pedro" e "Ele" são sujeitos das orações: paralelismo sintático (35). Assim, têm-se "Pedro" com valor de saliência igual a "190" e "carro" com valor de saliência igual a "140". Ambos os sintagmas concordam em gênero e número com o pronome "Ele", porém "Pedro" é escolhido como antecedente por possuir maior valor de saliência [Tabela 4.3].

Tipo de saliência	**Pedro**	**carro**
Saliência anterior / 2	155	140
Paralelismo Sintático	35	0
Total	190	140

Tabela 4.3: Valores dos fatores de saliência após a segunda sentença

4.3.2 Mitkov

A abordagem proposta por Mitkov [MS98] é do tipo *knowledge-poor pronoun resolution*. Abordagens deste tipo utilizam pouco conhecimento externo ao texto. A vantagem em se utilizar pouco conhecimento é ter um sistema de resolução menos dependente de conhecimento sintático, semântico e inferencial (*real-world knowledge*) [San00]. Este tipo de abordagem possui implementação barata e leve sendo apropriada para funcionamento em *corpora*, onde há milhares e até milhões de sentenças que precisam ser analisadas e o tempo de análise é um ponto importante para a aplicação.

O que foi observado por Mitkov [MS98] é que alguns indicadores de preferência foram eficientes em determinar os antecedentes anafóricos. Ele estudou um *corpus* de manuais técnicos de computadores anotado manualmente por especialistas humanos em busca de uma maneira prática para a resolução de anáforas pronominais evitando análises sintáticas ou semânticas. Mitkov elaborou indicadores de antecedentes (*antecedent indicators*), além da concordância de gênero e número, que podem ser utilizados para calcular a preferência que um sintagma nominal possui quando candidato a antecedente de um pronome, podendo assim chegar até a escolha do termo adequado.

Os indicadores de antecedentes anafóricos são baseados em fatores textuais e foram observados empiricamente [Cha07]. São eles:

- Primeiro sintagma nominal da sentença: o primeiro sintagma nominal de uma sentença é pontuado positivamente. O tema de uma sentença é exposto no início da sentença e a chance desse início ser referenciado posteriormente é grande.

- Verbos de indicação: certos verbos observados por Mitkov possuem uma característica intrínseca de propiciarem ao sintagma nominal que aparece exatamente antes desses verbos maior probabilidade de serem um antecedente anafórico. Por exemplo os verbos: analisar, acessar, apresentar, checar, considerar, cobrir, definir, descrever, desenvolver, discutir, examinar, exibir, explorar, identificar, investigar, ilustrar, revisar, sintetizar e sumarizar.

- Reiteração lexical: o fato de um termo aparecer mais vezes dentro de um mesmo parágrafo aumenta a sua chance de ser um antecedente anafórico.

- Título da seção: se um texto é dividido em seções, a ocorrência do sintagma no título seção aumenta a probabilidade de ser ele o antecedente pronominal de algum pronome dentro da seção.

- Padrão de colocação: este indicador dá preferência aos sintagmas que possuem mesmo padrão de colocação que o pronome em questão dentro das suas respectivas frases. Os padrões são: <SN ou pronome + verbo>, <verbo + SN ou pronome>, caso o verbo seja "ser" ou "estar" o padrão é <SN ou pronome + verbo + adjetivo ou particípio>.

- Preferência por alguns termos: os sintagmas nominais que representam conceitos dentro do tema central do texto possuem maior probabilidade de serem antecedentes anafóricos que os outros sintagmas nominais.

- Distância referencial: quanto mais próxima a frase que contém o sintagma nominal candidato a antecedente anafórico estiver da frase que contém o pronome mais chance ele, o

sintagma nominal, tem de ser o antecedente do pronome.

- Sintagma nominal preposicional: um sintagma nominal preposicional é iniciado por uma preposição e possui menos probabilidade de ser referenciado por um pronome posteriormente.

- Sintagma nominal indefinido: um sintagma nominal indefinido é iniciado por um artigo indefinido e possui chances mínimas de ser um antecedente anafórico para um pronome.

O processo para a resolução dos pronomes anafóricos, proposto por Mitkov, segue os seguintes passos: Tomaremos como exemplo as sentenças: *O dono da casa dormia no quarto. O assaltante escalou o muro do jardim. Ele acabou ficando preso na grade da janela.*.

- O texto é dividido em sentenças e os sintagmas nominais precisam ser identificados e marcados devidamente. Os sintagmas nominais que aparecem após o pronome ou discordem dele em gênero ou número são descartados.

  - Anáfora pronominal: “Ele” (singular e masculino).
  - Sintagmas nominais da sentença 1: “dono”, “casa”, “quarto”.
  - Sintagmas nominais da sentença 2: “assaltante”, “muro”, “jardim”.
  - Sintagmas nominais da sentença 3: “grade”, “janela”.

  Os sintagmas da sentença 3 serão excluídos por aparecerem após a anáfora e o sintagma nominal “casa” será excluído por ser feminino. Aqui tem-se uma limitação da abordagem pois há a ocorrência de ambos os casos na língua portuguesa como nos exemplos a seguir:

  O marido mentia para ela descaradamente. A mulher não aguentava mais.

  A quadrilha já existia há anos. Eles eram ladrões profissionais.

- É definido no processo de resolução uma janela textual de duas sentenças anteriores ao pronome. Isto quer dizer que no processo de resolução só serão considerados candidatos a antecedentes anafóricos os sintagmas nominais que estiverem presentes em uma das duas sentenças anteriores ao pronome.

  - Sintagmas nominais candidatos a antecedente anafórico: “dono”, “quarto”, “assaltante”, “muro”, “jardim”.

Apesar de ser incomum, um pronome pode se referir a um sintagma nominal em qualquer posição no texto, como no exemplo a seguir, onde o antecedente anafórico está três sentenças antes do pronome:

"O doutor" divulgou o resultado de suas pesquisas. Houve um alvoroço tremendo. A imprensa cogita um futuro prêmio nobel. Segundo "ele", os resultados ainda podem melhorar.

Os autores argumentam que limitações por regras impeditivas como estas são benéficas por duas razões:(i)qualquer modelo é incapaz de atender a todas possibilidades de referenciação de um pronome e precisa se concentrar nas mais frequentes; (ii)como não há, até então, uma abordagem que cubra os usos menos comuns dos pronomes, o fato do algoritmo sempre errar em alguma construção anafórico apenas reduz o percentual de acerto do algoritmo.

- São atribuídos pesos para cada indicador de antecedente anafórico, conforme a sua importância empírica em determinar a probabilidade de um sintagma nominal ser um antecedente anafórico.
  - padrão de colocação: +10.
  - distância referencial: -10 por frase.
  - sintagma nominal preposicionado:-10.
  - preferência por alguns termos (relação com o tema assalto):+10.
- Após os filtros aplicados e o domínio restringido, os indicadores são aplicados sobre os sintagmas nominais restantes e somados os pesos de cada indicador. O sintagma nominal que adquirir maior valor na soma dos pesos dos indicadores é escolhido como o antecedente anafórico.
  - "dono":distância referencial (2 frases) $\Rightarrow -20 = -20$.
  - "quarto":sintagma nominal preposicionado, distância referencial (2 frases) $\Rightarrow -10 - 20 = -30$.
  - "assaltante": preferência por alguns termos (relação com o tema assalto), padrão de colocação (SN + verbo), distância referencial (1 frases) $\Rightarrow 10 + 10 - 10 = 10$.
  - "muro":distância referencial (1 frases) $\Rightarrow -10$.
  - "jardim":sintagma nominal preposicionado, distância referencial (1 frases) $\Rightarrow -10 - 10 = -20$.

O sintagma nominal de maior pontuação foi "assaltante" (10) e por isso foi escolhido como o antecedente anafórico de "Ele" pelo algoritmo.

4.3.3 Centering

Centering é um conjunto de regras e restrições que determinam as relações entre o assunto tratado no texto e as escolhas linguísticas feitas pelo autor para expor a sua sequência de ideias. O algoritmo desenvolvido na implementação do *Centering* foi originalmente apresentado nos trabalhos de Brennan [BFP87] e reformulado em Grosz [GJW95].

É detectado o foco do discurso através das relações entre as entidades de sentenças distintas com a finalidade de associá-los ao pronome. Estas relações são classificadas como *continuing*, *retaining*, *shifting* e *shifting-1*. O modelo consiste de 3 conjuntos C_f, C_p e C_b, cujos elementos são os sintagmas nominais das sentenças e uma sequência de sentenças U_1,U_2,...,U_n [ACC$^+$04]. A intenção é determinar o foco central do discurso com a finalidade de encontrar os antecedentes dos pronomes.

A cada sentença U_n está associado um conjunto $C_f(U_n)$ (*foward looking center*) ao qual pertencem todos os elementos de U_n que podem ser foco na sentença posterior, os sintagmas nominais. O primeiro elemento desse conjunto é denominado $C_p(U_n)$ (*preferred center*) convencionalmente o mais provável de ser o foco da próxima sentença. O $C_b(U_n)$ (*backward looking center*) corresponde ao $C_p(U_{n-1})$, ou seja, geralmente ao primeiro elemento do C_f da sentença anterior, uma entidade que foi anteriormente introduzida, ainda permanece no contexto e é o provável foco.

A implementação do *Centering* conta ainda com um conjunto de regras e restrições que governam o fluxo de ideias apresentadas pelo texto. As restrições são 3:(i)Existe apenas um valor de $C_b(U_n)$; (ii)Todas as entidades pertencentes a C_f são enunciadas em U_n; (iii)$C_b(U_n)$ é o elemento mais importante em $C_f(U_{n-1})$. As regras são 2:(i)Se algum elemento de $C_f(U_{n-1})$ é pronome, então $C_b(U_n)$ e $C_p(U_{n-1})$ também são;(ii)a ordem de preferência entre as transações é $Continuing \succ Retaining \succ Shifting-1 \succ Shifting$.

Exemplo:

O réu conduzia um Alfa Romeu. Ele trafegava acima da velocidade permitida. O radar registrou que ele estava a 183km/h.

U_1 = O réu conduzia um Alfa Romeu.

C_f = réu, Alfa Romeu

C_b =

C_p = réu

U_2 = Ele trafegava acima da velocidade permitida.

C_f = Ele, velocidade

C_b = réu

C_p = Ele

U_3 = O radar registrou que ele estava a 183km/h.

C_f = radar, ele

C_b = Ele

C_p = ele

As transações *Continuing*, *Retaining*, *Shifting* e *Shifting-1* são determinadas pelos seguintes fatores:

- O foco de uma sentença $C_b(U_n)$ é ou não igual ao foco da sentença anterior $C_b(U_{n-1})$.
- O foco de uma sentença $C_b(U_n)$ é ou não igual ao provável elemento central da próxima sentença $C_p(U_n)$.

Essas regras formam as quatro possíveis relações entre as sentenças num texto coerente. A noção de coerência é que todas as proposições se referem a um elemento central que não é modificado e não são introduzidas novas entidades no texto (*Continuing*). Caso novas entidades sejam introduzidas, o foco da sentença anterior é conservado (*Retaining*). Ao pretender mudar o foco do discurso, a intenção é de mantê-lo nas próximas sentenças (*Shifting-1*) ou abandoná-lo (*Shifting*), segundo a tabela 4.4. A ordem de preferência comum é *Continuing* > *Retaining* > *Shifting-1* > *Shifting*.

Transações	$C_b(U_n) = C_b(U_{n-1})$	$C_b(U_n) \neq C_b(U_{n-1})$
$C_b(U_n) = C_p(U_n)$	Continuing	Retaining
$C_b(U_n) \neq C_p(U_n)$	Shifting-1	Shifting

Tabela 4.4: Possíveis transações entre o foco das sentenças

Para a execução do algoritmo nas sentenças: *O réu conduzia um Alfa Romeu. Ele trafegava acima da velocidade permitida. O radar registrou que ele estava a 183km/h.*, temos que:

Na primeira fase todos os pronomes pessoais são identificados como anáforas, e todos os sintagmas nominais são candidatos à solução. A janela de retomada de um termo pela anáfora pode ser convencionada em torno de 4 sentenças.

- anáforas: “Ele”, “ele”
- candidatos: “réu”, “Alfa Romeu”, “velocidade”, “radar”

Cada elemento de C_f é constituído por um par (anáfora, possível solução) e passa por uma checagem de gênero e número (pois devem concordar entre si). São montadas estruturas denominadas “Item” para as sentenças. Um “Item” é formado por um valor de C_b e um conjunto C_f das várias possibilidades de pares (anáfora, possível solução) para cada sentença.

- Itens da sentença 1: C_b=, C_f= ,“réu”, ,“Alfa Romeu”
- Itens da sentença 2: C_b=“réu”, C_f=“Ele”,“réu”, “Ele”,“Alfa Romeu”, “Ele”,“velocidade”
- Itens da sentença 3: C_b=“Ele”, C_f=“ele”,“réu”, “ele”,“Alfa Romeu”, “ele”,“Ele”, “ele”,“radar”

Na segunda fase tenta-se eliminar “Itens” que não apresentam soluções corretas da seguinte maneira: o elemento de $C_b(U_n)$ não pode ser diferente do elemento mais relevante de $C_f(U_{n-1})$ e a mesma anáfora não pode apontar para dois grupos nominais diferentes. Sempre que há repetição de uma anáfora no mesmo “Item” é exigido que haja repetição da possível solução, caso contrário o “Item” é eliminado.

- No exemplo que está sendo analisado, $C_b(U_n)$ foi previamente igualado a $C_f(U_{n-1})$ para evitar “Itens” desnecessários e não há ocorrência de duas anáforas em uma mesma sentença, logo não há repetição de anáforas em um mesmo “Item”.

Na terceira e última fase são eliminados os “Itens” vazios e que possuem anáforas repetidas. Os “Itens” são então classificados quanto a transação entre sentenças e ordenados quanto a sua preferência: 1- *Continuing*, 2- *Retaining*, 3- *Shifting-1*, 4- *Shifting*.

- Itens da sentença 1:
- Itens da sentença 2: C_b=“réu”, C_f=“Ele”,“réu”, “Ele”,“Alfa Romeu”, “Ele”,“velocidade”
- Itens da sentença 3: C_b=“Ele”, C_f=“ele”,“réu”, “ele”,“Alfa Romeu”, “ele”,“Ele”, “ele”,“radar”

Os "Itens" que apresentam transação *Continuing* são: C_b="réu" C_f="Ele","réu" e C_b="Ele", C_f="ele","Ele". Como há transação *Continuing* para resolução dos dois pronomes não há necessidade de classificar mais "Itens". O antecedente escolhido para o pronome "Ele" é "réu" e para o pronome "ele" é "Ele", ou seja, também "réu".

Os resultados desta abordagem para resolução de anáfora em textos em português apresentam percentagem de sucesso em 51% dos casos [ACC+04]. Este resultado está abaixo do encontrado nos trabalhos em língua espanhola, onde o algoritmo apresentou resultados acima de 75% [MBMA+99].

4.3.4 Algoritmo de Leffa, semântica sem conhecimento de mundo

Leffa [Lef03] apresenta uma abordagem pobre em conhecimento de mundo com a utilização de restrições sintáticas e semânticas que tentam resolver o problema da existência de dois ou mais candidatos legítimos para o antecedente do pronome. O resultado é baseado na ocorrência do pronome *they* em 1400 exemplos extraídos de um conjunto de textos composto por 10.000.000 de palavras em língua inglesa. As restrições baseiam-se nas funções sintáticas e traços semânticos dos candidatos a antecedente dos pronomes (anáforas) analisados.

Os seguintes exemplos mostram as restrições sintáticas e semânticas envolvidas com o processo de resolução anafórica e como se dá esse processo.

Exemplo 1: simplicidade

Casas são compradas porque elas oferecem conforto.

O termo "elas" não possui relação com um conceito próprio, mas está relacionado com outro termo. Pelas restrições sintáticas de gênero e número "elas" é um termo no feminino e plural, logo o termo "casas" se encaixa neste perfil, sendo o melhor candidato (e único).

Exemplo 2: 2 candidatos

Casa são compradas por pessoas porque elas oferecem conforto.

"casas" e "pessoas" obedecem as restrições sintáticas de feminino e plural. Uma opção é utilizar o paralelismo sintático, ou seja, o fato de "casas" e "elas" serem sujeito das orações e "pessoas" ser um objeto do verbo principal, colocando "casas" como a melhor opção para a resolução da anáfora.

Exemplo 3: simples mudança

Casas são compradas por pessoas porque elas desejam conforto.

o termo “oferecem” foi substituído por “desejam” alterando o antecedente do termo anafórico. Neste caso a prioridade dada a “casas” pelo paralelismo sintático é falsa, pois viola uma restrição semântica de que “casas” não desejam coisas. Deve-se então descartar o paralelismo sintático e utilizar-se da restrição semântica descrita acima, ou seja, o fato de “elas” ser o sujeito do verbo “desejar” que é uma propriedade de coisas que são “+ANIMADAS”, o antecedente anafórico deve ter a característica de ser “+ANIMADO”. Com isso voltamos a ter como melhor opção o termo “pessoas”.

Mais exemplos :

As concessionárias vendem carros para empresas porque elas oferecem garantia de longo prazo.

As concessionárias vendem carros para empresas porque elas possuem modelos a prova de bala.

As concessionárias vendem carros para empresas porque elas oferecem muito dinheiro.

Os exemplos acima podem sugerir que são facilmente resolvidos através de restrições semânticas baseadas em conhecimento de mundo. Por exemplo, dinheiro é pago do comprador para o vendedor, carros são entregues do vendedor para o comprador, carros podem ser a prova de balas, concessionárias oferecem garantia para o objeto que vendem e empresas podem ser muito ricas.

O problema é que restrições baseadas em conhecimento de mundo e suas combinações exponenciais podem explodir em custos computacionais. Leffa [Lef03] mostra que a solução ideal para resolução de anáfora deve encontrar-se entre a simplicidade das restrições sintáticas e a complexidade do conhecimento de mundo. Quase todos os dados necessários para a resolução de anáfora em um determinado texto precisam estar disponíveis na superfície do texto. Estes dados estão na forma de concordância em gênero e número, paralelismo sintático, repetição léxica e proximidade do antecedente e podem ser rapidamente verificadas por algoritmos. Portanto, prefere-se juntar a esses métodos soluções pobres em conhecimento de mundo, as quais utilizam metodologias dirigidas por *corpus* linguísticos ou modelos probabilísticos e estatísticos.

O processo se inicia com os termos classificados em classes e subclasses (substantivos, adjetivos, verbos..., verbo transitivo, complemento nominal...) e atribuindo-se algumas características sintáticas e semânticas (gênero, número, +humano, +animado...).

Exemplo:

Turistas preferem uma grande casa sobre a colina.

turistas[sujeito, masculino, plural, +animado, nominativo]

uma grande casa sobre a colina [objeto, feminino, singular, -animado, acusativo]

Leffa investiga o pronome pessoal "they" (eles/elas). A questão explorada é se é ou não possível resolver as anáforas sem recorrer a conhecimento de mundo, em outras palavras, o quanto a anáfora é dependente de restrições sintáticas e semânticas. A aplicação prática é a tradução do pronome "they" para línguas como francês, espanhol ou português onde há mais de uma opção de tradução dependendo do antecedente anafórico.

A metodologia utilizada por Leffa [Lef03] envolve o uso de 1400 ocorrências de "they" em um *corpus* com 10 milhões de palavras. Após encontrada a anáfora no trecho selecionado os possíveis antecedentes são identificados e classificados pelas suas funções sintáticas. O antecedente deve ser capaz de executar a mesma função sintática da anáfora sem quebrar as restrições sintáticas e semânticas envolvidas.

O algoritmo é o seguinte [Figura 4.1]:

- passo 1 : procurar por um nome plural até 80 palavras a esquerda de "they". Se um nome for encontrado, pula para o passo 2. Senão pula para o passo 4.
- passo 2 : o nome encontrado possui a mesma função sintática que "they"? "sim" pula para o passo 3, "não" volta para o passo 1.
- passo 3 : o nome encontrado pode substituir "they" sem causar uma anomalia semântica? "sim" pula para o passo 7, "não" volte para o passo 1.
- passo 4 : procurar por um nome até 80 palavras a esquerda de "they". Se um nome for encontrado, pula para o passo 5, senão pula para o passo 6.
- passo 5 : o nome encontrado pode substituir "they" sem causar uma anomalia semântica? "sim" pula para o passo 7, "não" volte para o passo 4.
- passo 6 : nenhuma solução encontrada. Adote uma resposta padrão e pule para o passo 7
- passo 7 : fim do procedimento.

Como vê-se no algoritmo, as restrições sintáticas como paralelismo sintático não credencia um possível antecedente a ser considerado o antecedente da anáfora, esta tarefa é sempre concluída com base nas restrições semânticas. Temos então duas fases de teste. A primeira testa os possíveis antecedentes com prioridade para a proximidade, com base nas restrições sintáticas e

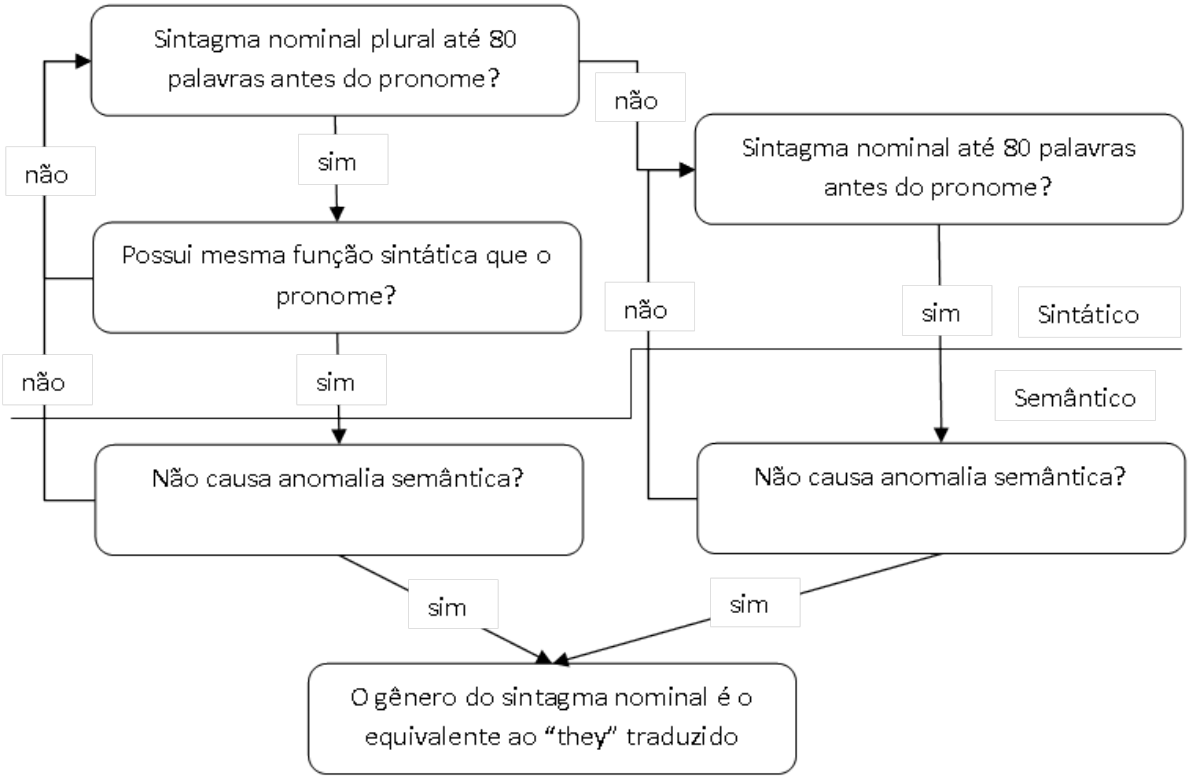


Figura 4.1: Fluxograma do algoritmo de Leffa

semânticas, se nenhum nome passar por todos os testes, então os nomes são testados só pelas restrições semânticas.

Um exemplo de quando o procedimento falha:

...Perseu casa-se com Andrômeda, e eles "they" posteriormente tornam-se rei e rainha...

o procedimento não consegue entender que o sujeito é composto e os núcleos estão separados na oração anterior.

Leffa destina-se a responder 3 questões:

- Quantos casos são resolvidos corretamente utilizando-se como restrição sintática apenas o paralelismo sintático, considerando apenas os casos em que as restrições semânticas não afetam o resultado?

  86%

- Quantos casos são resolvidos corretamente utilizando-se apenas as restrições semânticas?

  12%

- Quantos casos são resolvidos corretamente?

  98%

O paralelismo sintático mostrou-se o fator mais forte de resolução da anáfora. Analisando a literatura pertinente observou-se que entre os sistemas que usam restrições sintáticas e semânticas combinadas com abordagens estatísticas este é o nível de acerto mais elevado. Considerando a simplicidade do algoritmo a taxa de acertos é muito boa. Uma possível explicação é que é mais fácil encontrar o antecedente do pronome "they", do que o antecedente de outros pronomes, porque sintagmas nominais no plural são menos frequentes, diminuindo a quantidade de elementos no conjunto de candidatos a antecedentes anafóricos. Também deve ser levado em consideração que no teste do algoritmo o resultado é avaliado como sucesso se a categoria de gênero for atribuída corretamente ao pronome "they" e não se o sintagma nominal é realmente o antecedente anafórico.

O antecedente anafórico tende a ser o foco do discurso, e o foco do discurso tende a ser o sujeito das orações facilitando o encontro do antecedente anafórico de "they".

O paralelismo sintático (sujeito que se refere a outro sujeito) tende a ser muito forte na resolução da anáfora. Mesmo assim, o antecedente de "they" pode ser encontrado exercendo qualquer função sintática em outra oração. Leffa ainda argumenta que o nível de acerto elevado também se dá pela ordem em que o algoritmo procura e testa os possíveis antecedentes anafórico, privilegiando a proximidade com a anáfora. A forma como o algoritmo dá mais crédito as restrições semânticas também é um fator importante.

A aplicação de restrições semânticas pode ser confusa em orações na forma inversa (forma passiva). Isto porque um mesmo verbo pode ser +animado e também -animado dependendo da forma em que ele é aplicado.

Leffa ainda relata que na tentativa de reescrever o algoritmo para resolver os 2% dos casos em que ele falhou acabou em diminuir ainda mais o nível de acerto do algoritmo final [Lef03].

A resolução de anáfora através de restrições sintáticas e semânticas, sem recorrer ao uso de enciclopédias ou conhecimento de mundo tem o lado bom de alcançar altos níveis de acerto, bem próximo ao de seres humanos executando a mesma função. A parte ruim é que alguns erros advindos da abordagem adotada podem parecer ridículos da perspectiva de conhecimento de mundo baseado em senso comum e intuição humana. Leffa argumenta que a resposta aos problemas encontrados nesta abordagem tendem para a ideia de que existem muito mais dados analisados para a resolução de anáfora do que as informações contidas na superfície do texto [Lef03]. Conhecimento de mundo parece ser a fonte mais valiosa de todas, por outro lado,

em processamento de linguagem natural haverá apenas uma transferência do problema para um nível mais alto de abstração sem com isso resolvê-lo. Senso comum, intuição, variáveis sócio-culturais e outros componentes do conhecimento de mundo são muito evasivos e vagos para serem tratados adequadamente pela linguística computacional [Lef03].

4.4 Comparação entre as abordagens explanadas

Cada abordagem descrita acima possui características próprias relativas a como o fenômeno da anáfora é visto. Uma comparação foi realizada neste trabalho para possibilitar a análise dos pontos positivos e negativos de cada abordagem de acordo com as peculiaridades de cada uma. Foram definidos 5 fatores importantes para a avaliação de um algoritmo de resolução de anáforas:

1. Identificar o *corpus* sobre o qual o algoritmo foi testado e seus resultados foram obtidos. As características de um *corpus* podem influenciar na resolução de anáfora. Textos mais cultos, mais rebuscados ou até em forma de poesia dificultam o processo de resolução de anáfora. Conter somente textos específicos de um determinado domínio do conhecimento demonstra uma certa especificidade do algoritmo para um determinado domínio de conhecimento. Por outro lado, um *corpus* bem abrangente, contendo textos variados é capaz de informar sobre o comportamento do algoritmo em casos gerais.

2. O objetivo da aplicação ao utilizar o algoritmo e como se dá o caso de sucesso. Alguns algoritmos podem interpretar os casos de sucesso de maneira diferenciada. Isto se dá pela variedade de utilização do Processamento de Linguagem Natural. Os resultados devem ser comparados observando-se adequadamente a que objetivo eles se referem.

3. A taxa de acerto que o algoritmo alcançou durante os testes é um item quantitativo observado para a avaliação do algoritmo. A taxa de acerto é a quantidade de casos em que o algoritmo obteve sucesso em seu objetivo dividido pela quantidade de casos analisados.

4. As vantagens do algoritmo. Que pontos podem ser considerados positivos no algoritmo e que o diferenciam dos outros.

5. As desvantagens do algoritmo. Quais características podem ser observadas como negativas no algoritmo.

Algoritmo de Lappin e Leass utiliza um sistema de pesos atribuídos de acordo com a estrutura sintática da sentença. Apenas conhecimento sintático é utilizado para a resolução de

anáforas [Coe06].

- *corpus* testado: Os *corpora* foram todos anotados pelo parser morfossintático PALAVRAS. Foi utilizado um *corpus* literário, jornalístico e jurídico.
- objetivo da aplicação: Encontrar o antecedente anafórico de pronomes pessoais e pronomes recíprocos ou reflexivos. A resolução é considerada correta caso o sintagma apresentado pelo algoritmo seja o sintagma anotado no *corpus* como antecedente anafórico ou está contido nele.
- taxa de acerto: Foram um total de 1218 pronomes analisados e 428 resolvidos corretamente. Uma taxa de acerto de 35,14%.
- vantagens: O fato do algoritmo utilizar somente conhecimento sintático é uma vantagem do ponto de vista de eficiência computacional. Outra vantagem do algoritmo é poder ser utilizado para a resolução de pronomes pessoais e pronomes recíprocos ou reflexivos, englobando maior variedade de pronomes.
- desvantagens: Não considera o fator semântico da relação anafórica. Taxa de acerto muito baixa no caso analisado.

Algoritmo de Mitkov utiliza conhecimento sobre o domínio das sentenças e pouco conhecimento sintático. Insere o conceito de indicadores de preferências textuais na seleção do antecedente anafórico.

- *corpus* testado: Testado em um *corpus* formado por manuais técnicos de computadores escrito em inglês e anotado manualmente em relação aos sintagmas nominais, anáforas pronominais, gênero e número.
- objetivo da aplicação: Identificar o antecedente anafórico de pronomes.
- taxa de acerto: De um total de 223 pronomes anafóricos marcados manualmente o algoritmo acertou a identificação do antecedente anafórico em 200 casos, fazendo uma taxa de acerto de 89,7%.
- vantagens: Algoritmo leve que não exige demasiado conhecimento sintático ou semântico e com boa taxa de sucesso.

- desvantagens: Muito influenciado pelo domínio literário do texto analisado. Manuais técnicos de computadores são textos muito formais que repetem a estrutura das sentenças e as relações anafóricas. Pela natureza probabilística da abordagem, qualquer texto menos formal e consequentemente que repete menos as mesmas estruturas das sentenças, produziria uma queda na taxa de acerto.

Algoritmo *Centering* trata a anáfora como elemento que mantém a coesão textual. O sintagma principal da sentença (foco) é tratado como um elemento que preferencialmente é referenciado na sentença seguinte.

- *corpus* testado: *corpus* jurídico analisado morfossintaticamente pelo parser PALAVRAS.
- objetivo da aplicação: Identificar e resolver anáforas pronominais.
- taxa de acerto: De um total de 302 anáforas anotadas no *corpus* foram identificadas corretamente 282 anáforas e resolvidas corretamente 154. Com uma taxa de acerto de 51%.
- vantagens: Utilização de regras gramaticais abrangentes. Taxa de acerto razoável mesmo em um *corpus* complexo como o utilizado.
- desvantagens: O algoritmo é baseado na manutenção da coesão textual, portanto textos menos cultos provavelmente terão uma taxa de acerto reduzida.

Algoritmo de Leffa utiliza pouco conhecimento sintático e uma função semântica de validação.

- *corpus* testado: *corpus* de 10.000.000 de palavras de texto expositivo em língua inglesa, anotado sintaticamente e semanticamente.
- objetivo da aplicação: Descobrir o gênero do antecedente anafórico para a correta tradução do pronome em inglês “they”.
- taxa de acerto: Em 1400 ocorrências do pronome “they” foi identificado o gênero correto do antecedente anafórico em 98% dos casos.
- vantagens: Algoritmo muito simples, utilizando apenas o paralelismo sintático e a ocorrência de anomalias semânticas. Ótima taxa de acerto.

- desvantagens: O objetivo da aplicação foi muito simplificado em relação a resolução de anáforas pronominais. Mesmo que o antecedente anafórico fosse escolhido erroneamente mas possuísse mesmo gênero do antecedente anafórico correto, o algoritmo obteria sucesso em seu objetivo.

Analisando as abordagens apresentadas, percebe-se que a utilização de conhecimento semântico pode melhorar a resolução de anáforas pronominais. O conhecimento semântico é capaz de acrescentar informações que podem distinguir entre sintagmas que possuem a mesma inclinação a serem escolhidos como antecedentes anafóricos, além disso é possível evitar que anomalias semânticas sejam formadas, como por exemplo um animal ser apontado como o antecedente anafórico de um pronome que está ligado a um verbo cuja ação é estritamente executada por humanos. Poucas abordagens atuais utilizam este tipo de conhecimento por não ter-se uma definição conclusiva sobre como o significado de um sintagma é atribuído. Uma abordagem muito dependente dos textos utilizados para teste possuem um melhor resultado, mas são dificilmente portáveis para outros domínios do conhecimento. Já uma abordagem mais genérica quanto ao domínio do conhecimento necessita tratar as anáforas de modo igualmente abrangente, evitando regras que beneficiam as formas mais frequentes de anáforas e descartando alguns tipos delas. Entende-se então que um processo de resolução de anáfora pronominal deve contemplar todas as formas em que tais anáforas possam aparecer e além disso contar com informações sintáticas e semânticas, não sendo atrelado ao domínio dos textos analisados, para se obter sucesso na escolha do antecedente anafórico.

5 *Uma nova abordagem para a resolução de anáforas pronominais*

O presente trabalho propõe a utilização do SIM e da medida de relacionamento inferencial do SIA como uma etapa da resolução de anáforas pronominais pessoais, nomeadamente os pronomes “ele”, “ela”, “o”, “a” e “lhe”. Os pronomes possessivos “dele” e “dela” são entendidos pelo analisador sintático utilizado como “de ele” e “de ela” respectivamente. Por este motivo, a abordagem utilizada também trata de pronomes possessivos destas formas. A escolha restrita na implementação pelos pronomes “ele” e “dele” se dá como uma avaliação inicial do funcionamento desta abordagem, sendo posteriormente estendida para uma variedade maior de pronomes de acordo com o sucesso da técnica. Toda anáfora tem uma ligação semântica que é identificada pelo leitor na compreensão do enunciado. Processamentos sintáticos são rápidos e diretos mas não conseguem capturar a ligação semântica existente entre a anáfora e seu antecedente anafórico. Acredita-se que a abordagem semântica inferencialista de Pinheiro et AL [PPFN09] consiga explicitar essa relação semântica entre a anáfora e seu antecedente.

O processo de resolução da anáfora trata de encontrar expressões (pronomes) que se referem à conceitos anteriores, acrescentam informações sobre eles ou os relacionam com novos conceitos e explicitar tal informação ou relacionamento. O entendimento da sentença que contém um pronome pessoal só se completa quando entendemos a quem ou a que o pronome está ligado semanticamente.

As técnicas atuais de resolução de anáforas, em geral, se baseiam em informações sintáticas e morfológicas, avançando muito pouco na obtenção de informações semânticas. Aqui serão utilizadas estas informações em conjunto com o próprio SIM que provê a expressão de conhecimento semântico, inferências e uma medida de relacionamento inferencial que são capazes de tornar explícitas as relações entre os conceitos existentes no texto. Também será abordada um tipo de anáfora que ainda não é bem resolvida nas abordagens atuais, a anáfora conceitual. Tais tipos de anáforas são caracterizadas pelo antecedente anafórico não concordar gramaticalmente com a anáfora em gênero e/ou número. Deste modo é difícil relacioná-los sem utilizar as infe-

rências baseadas no conteúdo semântico dos conceitos que formam as sentenças. Acredita-se que as inferências feitas pelo SIM possam relacionar bem este tipo de anáforas.

A contribuição deste trabalho é explorar uma nova forma de atribuição de significado denominada semântica inferencialista na tarefa de resolução de anáforas pronominais. Ao mesmo ponto em que entendemos que a ligação existente entre um pronome e o sintagma que está relacionado a ele através do processo da anáfora, se dá semanticamente, compreendemos que a interpretação semântica de um texto e de suas partes só é possível através do entendimento das relações existentes entre os conceitos articulados no texto. Sendo assim, o Semantic Inferentialism Analyser (SIA) [Pin9a] possui um cálculo de relacionamento inferencial entre conceitos que ajudará a atribuir um valor de relacionamento entre conceitos de acordo com o nível de envolvimento que um conceito possui em relação a outro e tal valor será utilizado para identificar o sintagma mais próximo dos conceitos envolvidos na sentença em que o pronome está inserido.

5.1 Corpus linguístico e analisador sintático

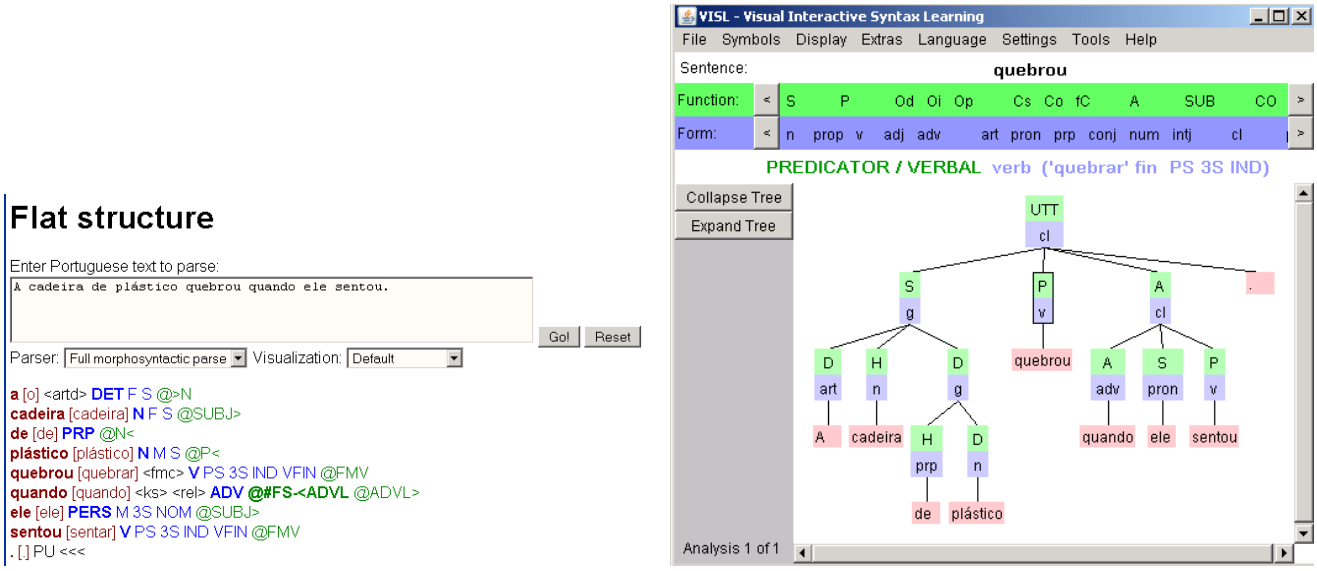
A Floresta Sintática é composta por um conjunto de textos (frases) que foram sintaticamente analisados e revistos intelectualmente, para servir em estudos variados além das pesquisas em que os seus criadores estão envolvidos. Os *corpora* que compõem a Floresta Sintática foram desenvolvidos no projeto VISL, um projeto de pesquisa e ensino da Universidade do Sul da Dinamarca com estudos na área de análise computacional automática iniciado em 1996. Utilizando o *parser* PALAVRAS construído para o português [Bic00], foram feitos ferramentas e banco de dados linguísticos para análise automática de textos. Um *corpus* tem a finalidade de representar uma língua através de um pequeno grupo de textos quando relacionado com toda a produção literária de uma língua, porém grande o suficiente para minimizar as características individuais dos autores. Assim, um *corpus* torna-se uma ferramenta fundamental para análises de textos em linguagem natural. Foi criado uma página na internet a partir da qual é possível entrar com um texto e verificar sua análise morfossintática bem como uma árvore representando a estrutura de dependências das frases. O VISL engloba atualmente 14 línguas, onde 6 delas possuem análise automática segundo o paradigma CG (*Constraint Grammar*) [ABHS01]. A aplicação disponível via *web* possui vários filtros notacionais, apresentando ao usuário uma interface capaz de colorir as palavras de acordo com sua classe morfossintática e criar árvores sintáticas gráficas com rótulos informando a forma e função sintática de cada palavra.

O projeto lançado pelo Ministério da Ciência e da Tecnologia “Processamento computa-

cional do português" também participou na construção da Floresta Sintática. Este projeto tem como um de seus objetivos a criação de recursos públicos, validados por linguistas, que possam ser utilizados na avaliação de analisadores sintáticos e outras ferramentas na área de processamento computacional do português [ABHS01].

A primeira etapa da plantação da Floresta Sintática consistiu da análise dos textos para o enriquecimento do parser *PALAVRAS* com novas palavras (cerca de 8000 a 9000) que não pertenciam ao seu dicionário e para uma revisão dos critérios automáticos de separação frásica. A segunda etapa englobou a criação automática e revisão manual dos *corpora* em formato CG. CG é uma gramática dependencial que marca as dependências com os símbolos ">" e "<" indicando a direção do núcleo sintático de que os constituintes possuem dependência.

A figura 5.1 apresenta "A cadeira de plástico quebrou quando ele sentou." analisada pelo PALAVRAS e o resultado da análise em dois formatos diferentes: o formato CG e a estrutura de dependência da frase.



(a) http://visl.sdu.dk/visl/pt/parsing/automatic/parse.php (b) http://visl.sdu.dk/visl/pt/parsing/automatic/trees.php

Figura 5.1: *Análise morfossintática do PALAVRAS*

As tabelas 5.1 e 5.2 exibem os *corpora* disponíveis para *download* na página do projeto VISL *http://www.linguateca.pt/Floresta/corpus.html♯download*. Tais tabelas informam para cada *corpus* a quantidade de palavras, a quantidade de frases, se o *corpus* foi revisado por especialistas humanos, se os textos incluem o português de Portugal, o gênero literário, o domínio de conhecimento expresso nos textos, se o registro dos textos se deu de modo formal ou informal, o modo de obtenção dos textos e a origem dos textos pertencentes ao *corpus*. Os *corpora* disponíveis são: "Floresta virgem", "Amazônia", "Bosque", "Selva literária", "Selva falada" e "Selva científica".

Corpus	Floresta virgem	Amazônia	Bosque
Palavras	1.640.000	4.580.000	186.000
Frases	96.000	275.000	9.368
Revisão	Não	Não	Integral
Variantes	PT / BR	BR	PT / BR
Gênero	Jornalístico	Opinião	Jornalístico
Domínio	Genérico	Cultura brasileira	Genérico
Registro	Formal	Formal e informal	Formal
Modo	Escrito	Escrito	Escrito
Origem	Jornais	Blog Overmundo	Jornais

Tabela 5.1: Tabela dos *corpora* produzidos na Floresta Sintática

Corpus	Selva literária	Selva falada	Selva científica
Palavras	105.000	170.000	125.000
Frases	7.900	14.000	6.200
Revisão	Parcial	Parcial	Parcial
Variantes	PT / BR	PT / BR	PT / BR
Gênero	Literário	Entrevistas / debates	Acadêmico / informativo
Domínio	Genérico	Biografia / política	Educação/ ciências
Registro	Formal	Formal e informal	Formal
Modo	Escrito	Falado	Escrito
Origem	Livros	Museu(PT,BR)/debates	Bibliotecas/wikipedia

Tabela 5.2: Tabela dos *corpora* produzidos na Floresta Sintática

Dada as características dos *corpora* [tabela 5.1 e 5.2], e diante da dificuldade da utilização total do *corpus* "Amazônia" foi escolhido para avaliação do presente trabalho um *corpus* formado por: trechos do *corpus Amazônia* pela sua ligação direta com a práxis da comunidade linguística brasileira, pelo seu conteúdo inteiramente nacional e geograficamente diversificado e pela sua quantidade maior de frases englobando diversos domínios do conhecimento; trechos do *corpus Bosque* pela inclusão de língua culta devido ao caráter jornalístico; trechos dos textos analisados pelo SIM em sua avaliação no sistema WIKICRIMESIE.

O *corpus Amazônia* foi extraído do sítio colaborativo "Overmundo" [Figura 5.2], que está disponínel para acesso em "www.overmundo.com.br". Este sítio tem como objetivo a expressão da produção cultural brasileira, contando com diversos autores em diferentes locais do Brasil. Foram extraídos todos os textos da seção "overblog" e todos os textos de não-ficção da seção "banco de cultura" disponíveis em 30 de setembro de 2008, somando um total de 4070 textos e 1303 autores. A capacidade de um *blog* em capturar a práxis linguística é notável, visto que as pessoas se sentem à vontade em exprimir seus pensamentos e opiniões sem serem censuradas ou reprimidas.

O *corpus Bosque* foi retirado de jornais digitais da Folha de São Paulo (CetemFolha) e de

jornais de Portugal (CetemPúblico).

As notícias de crimes utilizadas na avaliação do SIM foram retiradas de *sites* da *internet* e são notícias reais de crimes ocorridos no Brasil.

Os trechos utilizados para avaliação do algoritmo desenvolvido são compostos por sentenças onde há a ocorrência dos pronomes "ele" e "dele" e que se encontram no início do *corpus Amazônia*, do *corpus Bosque* e do grupo de textos que descrevem relatos de crimes denominado "Coleção Dourada" da avaliação do SIM no projeto WikiCrimeIE [PPFF10]. Os textos extraídos de dada um destes *corpora* foi analisado por um humano adulto e resolvida a anáfora pronominal existente em cada trecho anotando com a *tag "<anaf>"* onde estava o antecedente correto da anáfora. De maneira nenhuma a informação desta *tag* foi utilizada no processo de resolução descrito neste trabalho, a não ser para verificar automaticamente se a escolha do antecedente anafórico foi correta ou errada com a finalidade de se obter a taxa de acerto do algoritmo. Está disponível na internet um documento compartilhado com os textos utilizados neste trabalho: *https://docs.google.com/Doc?docid=0ARGXRjw55R-PZGRucGM5bWhfMTVoaHQ4OHFoYw&hl=en*

Figura 5.2: http://www.overmundo.com.br/overblog

5.2 Processo de resolução de anáforas utilizando o SIA

A abordagem proposta para o problema da resolução de anáforas pronominais pessoais de terceira pessoa é composta de duas fases de desenvolvimento. Na primeira fase serão criados critérios sintáticos, heurísticas sintáticas, que se baseiam, cada um, em um fundamento linguístico para encontrar o antecedente anafórico. Tais critérios visam retirar, do domínio dos sintagmas nominais candidatos a antecedente, os sintagmas que não se enquadram em certos preceitos gramaticais. Na segunda fase será criado um algoritmo que recebe apenas os candidatos que foram selecionados pelos critérios sintáticos e combina os critérios semânticos adicionando pesos em cada critério semântico e selecionando o candidato que adquirir maior pontuação geral.

5.2.1 Testes preliminares

A implementação e os testes preliminares dos critérios sintáticos utilizados para apontar o antecedente anafórico foram avaliados no *corpus* Bosque [ABHS01], onde foram selecionados apenas os textos onde ocorriam a presença dos pronomes "ele" e "dele", o que somou 88 textos. Um critério sintático não aponta necessariamente para somente um sintagma definindo-o como o antecedente anafórico. Pelo contrário, a utilidade de um critério está em pontuar positivamente (ou negativamente) alguns sintagmas apontando-os (quase sempre mais de um) como possíveis candidatos a antecedente anafórico ou diminuindo as chances que um sintagma tem de sê-lo. O algoritmo preliminar implementado e testado será utilizado no algoritmo final, compondo a parte sintática deste último. Os resultados apresentados são apenas resultados preliminares. Vê-se a seguir uma lista dos critérios implementados no algoritmo de teste deste trabalho:

- Concordância em gênero e número: Os candidatos que concordam em gênero e número com a anáfora pronominal a ser resolvida possuem grande chance de serem o antecedente. Os testes mostraram que 100% dos antecedentes do pronome concordavam em gênero e número, sendo este critério um fator que sempre abrange o candidato mais provável a antecedente (com algumas exceções que não estão presentes no *corpus* avaliado). Por outro lado, este critério é muito abrangente e em todos os casos foram indicados mais de um candidato, sendo portanto ineficaz sua utilização sem a combinação com outros critérios, ou seja, este critério não é definitório no sentido de escolher um antecedente anafórico único.

- Ocorrência anterior: É muito provável que o antecedente anafórico esteja entre os sintagmas nominais que ocorrem anteriormente ao pronome. Os testes mostraram que em

87,5% dos casos as anáforas pronominais ocorriam após o seu antecedente anafórico, o que nos indica que este critério é quase sempre respeitado, mas não pode ser um fator de corte definitivo de um candidato a antecedente. Este critério apesar de quase sempre ser obedecido pelo antecedente anafórico também é ineficaz quando utilizado individualmente por não apontar para um único sintagma.

- Proximidade com o verbo principal: Os termos que mais se aproximam do verbo principal na árvore estrutural da sentença analisada pelo PALAVRAS são termos muito salientes no texto. Os termos salientes são mais facilmente de serem referenciados após o seu uso, o que o coloca numa posição de forte candidato a antecedente pronominal [ACC$^+$04]. Para este critério foram utilizados os quatro sintagma mais próximos do verbo principal. Tal critério apontou corretamente o antecedente anafórico em 37,5% dos casos analisados, ou seja, em 37,5% dos casos analisados o antecedente anafórico estava entre os quarto sintagmas mais próximos do verbo principal da sentença.

- Proximidade do pronome na árvore: A árvore estrutural de uma sentença liga os termos da sentença pelas suas funções sintáticas. Estas ligações explicitam a forma como montamos uma sentença para raciocinar sobre ela. A proximidade do candidato a antecedente em relação ao pronome quando visto do ponto de vista da árvore estrutural da sentença revela a distância funcional entre eles. Para tal critério foram utilizados os quatro sintagmas mais próximos. A escolha pelos quatro candidatos mais próximo do pronome na árvore estrutural mostrou 43,2% de acerto., ou seja, em 43,2% dos casos o antecedente anafórico estava entre os quatro sintagmas mais próximos do pronome na árvore sintática.

- Proximidade do pronome no texto: A proximidade com o pronome é um fator empírico [San00] que ajuda na resolução de uma anáfora. Os sintagmas mais próximos do pronome possuem uma certa tendência de serem o antecedente anafórico correto por serem termos que ainda estão bem presentes na mente do leitor do texto no momento do entendimento da anáfora. Também neste critério foram utilizados os quatro sintagmas mais próximos do pronome. Os testes mostraram que 35,2% dos antecedentes anafóricos estavam entre os quatro sintagmas nominais mais próximos do pronome.

- Semelhança sintática: O candidato escolhido por esse critério são os que possuem a função sintática semelhante a do pronome em questão. Este fator é responsável por criar uma ligação de coesão no texto e permanência do foco textual [Coe06]. A utilização deste critério nos testes apontou que em 48,8% dos casos os antecedentes anafóricos respeitavam este critério, ou seja, em 48,8% dos casos o antecedente anafórico concordava com o pronome quanto as suas funções sintáticas no texto.

Vale salientar que os critérios utilizados não apontam para um único candidato, sendo inviável suas utilizações isoladamente para determinação do melhor candidato. Durante a avaliação dos critérios já descritos foi realizada uma tentativa ponderada de união dos critérios. Ao respeitar um critério utilizado o candidato era pontuado positivamente e após todos os critérios eles foram ordenados em ordem decrescente com relação ao somatório dos pontos que cada um recebia. O candidato que obteve maior somatório das pontuações era então apontado como o melhor candidato a antecedente do pronome. Tal combinação extritamente sintática mostrou que em 46,6% dos casos o candidato selecionado era realmente o antecedente anafórico correto do pronome.

Um critério adicionado após os testes preliminares e que está no cerne do presente trabalho é calcular a relação inferencial entre o pronome e os sintagmas nominais candidatos a antecedente do pronome. Como o pronome em si não possui valor inferencial, serão utilizados os termos diretamente ligados ao pronome dentro da sentença. Tais termos precisam estar mais inferencialmente relacionados ao sintagma para que este seja o antecedente anafórico. Algumas anáforas só são corretamente resolvidas quando tal critério é utilizado. Os exemplos a seguir ilustram estes casos:

João comprou um gato. Ele andava muito triste.

João comprou um gato. Ele miava muito triste.

O pai e o filho brincaram o dia inteiro. Ele acabou faltando ao trabalho.

O pai e o filho brincaram o dia inteiro. Ele acabou faltando à escola.

O que faz a mente humana ligar o pronome "Ele" do primeiro exemplo a "João" e o pronome "Ele" do terceiro exemplo a "O pai" são informações inferenciais que extrapolam informações sintáticas ou restrições semânticas baseadas em *corpus* anotados. Em ambos os casos tanto "João" quanto "gato", bem como, "O pai" e "o filho" são sintaticamente capazes de ocupar o lugar de antecedente anafórico, porém algo nos diz que somente um deles possui razão para ocupar tal lugar. Isto se dá porque "João", como ser humano, está ligado a estados emocionais como "tristeza" mais fortemente que "gato", como um animal doméstico, por outro lado, "o gato" está mais relacionado a "miava" do que "João". Da mesma maneira "O pai" está mais fortemente relacionado a "trabalho" do que "o filho", bem como "o filho" está mais relacionado a "escola" do que "O pai", visto que um pai trabalha e um filho estuda. Estas informações são impossíveis de serem capturadas por restrições sintáticas porque elas se estabelecem através do uso da linguagem pela comunidade linguística. Com isso, propomos o uso da medida de relacionamento inferencial do SIA como um dos critérios finais de seleção do melhor candidato

a antecedente anafórico.

Vê-se que, com a escolha de não colocar os critérios com um caráter impeditivo, é possível abranger uma quantidade maior de ocorrências de anáforas pronominais incomuns e possibilitando o relacionamento entre pronome e antecedente em um nível semântico-pragmático que supera todas as outras abordagens existentes. O exemplo a seguir demonstra essa superioridade:

O time jogou muito bem. Elas estavam todas entrosadas.

Sem concordância de número ou gênero, a anáfora se dá porque sabe-se que um "time" é formado por várias pessoas, conotando o sentido de plural, e é possível que "time" seja formado por pessoas do sexo feminino.

5.2.2 Algoritmo de resolução de anáfora utilizando o SIA

O algoritmo desenvolvido para a resolução de anáfora analisa todos os candidatos a antecedente anafórico do texto. Para o algoritmo, todo sintagma nominal é considerado um candidato a antecedente, inclusive os adjetivos com valor de substantivo como na sentença "O feio era mais inteligente.", onde o adjetivo "feio" assume valor de substantivo, sendo até mesmo o sujeito da oração, neste caso. Todos os sintagmas são então colocados em uma lista de candidatos a antecedente anafórico.

Partindo deste princípio o algoritmo analisará quais os critérios sintáticos e semânticos que cada candidato obedece. Os critérios sintáticos são baseados em características sintáticas que são apontadas em trabalhos anteriores como capazes de indicar o antecedente anafórico de um pronome. Esta indicação se dá de forma incerta e na maioria dos casos os critérios são obedecidos por mais de um candidato, tornando os critérios sintáticos, em geral, eficazes somente se utilizados em conjunto. Os critérios semânticos serão explicados mais adiante nesta seção.

Ao analisar se um candidato obedece a um determinado critério sintático o algoritmo atribui um valor para cada candidato. O algoritmo realiza os seguintes passos para cada critério: se o candidato obedece ao critério em questão, lhe é atribuído o valor multiplicado pelo peso correspondente ao critério atual. Após esta atribuição de valores os candidatos que obtiverem valor menor que o limite de corte do critério são eliminados da lista. Se após as eliminações a lista possuir menos que dois candidatos, então as eliminações são desfeitas. Após todos os critérios serem analisados a lista é então ordenada de forma decrescente de acordo com a pontuação geral de cada candidato e os quatro primeiros candidatos são apontados como os possíveis antecedentes anafóricos do pronome. O valor "quatro"para esta função de corte foi escolhido

analisando-se empiricamente os dados e chegando-se a conclusão de que os critérios sintáticos utilizados sempre colocavam entre os quatro candidatos mais bem colocados nas pontuações o antecedente anafórico. Esta função de corte faz-se necessária pelo fato do processamento dos passos seguintes serem mais demorados que os realizados até o momento.

Os critérios sintáticos, os pesos e os limites de corte utilizados são explicados a seguir:

1. Concordância em gênero e número: Para cada candidato que concorda em gênero com o pronome, ou seja, se ambos são masculinos ou ambos femininos, então o candidato recebe o valor +1. O mesmo ocorre para a concordância em número, se ambos, pronome e candidato, estão no singular ou se ambos estão no plural o candidato recebe o valor +1. Após este critério os candidatos podem obter os valores 0, 1 ou 2. Todo candidato que obteve valor menor que 2 neste critério é eliminado do processo de resolução da anáfora. Porém, como visto anteriormente, se ocorrer da lista de candidatos ficar com menos que 2 candidatos as eliminações são desfeitas. O fato de desfazer algumas eliminações garante o tratamento das anáforas conceituais, que embora possua uma frequência menor, é capaz de relacionar um sintagma a um pronome mesmo que ambos discordem em gênero e número. Porém nestes casos foi visto que não há mais do que um candidato que concorde em gênero e número com o pronome. O peso atribuído por este critério é igual a 10. O limite de corte é igual a 2. Valor do critério: $V = (x_1 + x_2) * 10$. Limite de corte: $x_1 + x_2 < 2$. Onde $x_1 = 1$ se concorda em gênero ou 0 caso contrário, $x_2 = 1$ se concorda em número ou 0 caso contrário e $V =$ pontuação atribuída ao candidato na avaliação deste critério.

2. Ocorrência anterior ao pronome: É então verificado a ocorrência do pronome no texto analisado. Se o candidato ocorrer no texto anteriormente ao pronome ele recebe o valor 1. Este critério atribui valores 0 ou 1 para cada candidato. É também um critério utilizado como fator de corte do candidato igual a 1. Atribuir-se-á um peso igual a 10 a este critério, mas não será impossibilitada a resolução de catáfora, caso onde o antecedente anafórico só ocorre após o pronome, pois nestes casos o comum é que não haja outros candidatos anteriores ao pronome, fazendo com que as eliminações sejam desfeitas por conta da lista de candidatos ficar menor que 2. Valor do critério: $V = (x) * 10$. Limite de corte: $x < 1$. Onde $x = 1$ se ocorre antes do pronome ou 0 caso contrário e $V =$ pontuação atribuída ao candidato na avaliação deste critério.

3. Proximidade no texto: Todo sintagma possui influência sobre os sintagmas próximos a ele e esta influência diminui com a distância. O proximidade se refere a influência que o sintagma exerce sobre seus vizinhos próximos no texto. Os sintagmas que ocorrem

próximos ao pronome no texto estão mais salientes na mente do leitor e são mais facilmente recuperados no momento da resolução mental da anáfora. Deste modo é comum que ao redigir um texto o autor posicione o antecedente anafórico próximo ao pronome para facilitar o entendimento do texto durante a resolução da anáfora pronominal. Aqui são atribuídos aos candidatos a antecedente anafórico o valor igual a sua distância ao pronome multiplicado por -1. A distância é calculada com base em quantas palavras estão entre o candidato e o pronome + 1. Por exemplo nas sentenças: *O mais lindo era o lago. Ele brilhava como a lua.* o sintagma "lago" possui distância igual a 1, enquanto que os sintagmas "lindo" e "lua" possuem distâncias igual a 4. O maior valor atribuído é igual a -1 e o menor é igual a quantidade de palavras nas sentenças analisadas multiplicado por -1. O peso atribuído a este critério é igual a 15 e o limite de corte é -50, ou seja, candidatos distante do pronome por mais de 50 palavras são eliminados. Valor do critério: $V = ((x+1)*-1)*15$. Limite de corte: $x > 50$. Onde $x =$ quantidade de palavras entre o candidato e o pronome e $V =$ pontuação atribuída ao candidato na avaliação deste critério.

4. Proximidade na árvore: Semelhantemente ao critério anterior, a proximidade é que está em questão. Porém ao invés de se prender a posição dos sintagmas no textos, este critério analisa a distância do ponto de vista da construção sintática das sentenças, analisando a árvore sintática gerada pelo PALAVRAS para cada sentença. Toda sentença possui como raiz o seu verbo principal e como nós os demais sintagmas. Estes por sua vez se dividem de acordo com a importância dos sintagmas (sujeitos, objetos e etc) ou a ocorrência de outros verbos (orações subordinadas, coordenadas e etc). A distância então é a quantidade de arestas percorridas para se chegar do candidato ao pronome. Quando o candidato não está na mesma sentença do pronome assume-se que as raízes das árvores estão todas ligadas por arestas. O maior valor atribuído por este critério é -1 e o menor é igual a profundidade da árvore sintática multiplicado por -1. O peso atribuído a este critério é igual a 35. Valor do critério: $V = (x*-1)*35$. Limite de corte: $x > 10$. Onde $x =$ quantidade de arestas do menor caminho entre o candidato e o pronome e $V =$ pontuação atribuída ao candidato na avaliação deste critério.

5. Altura na árvore: Baseando-se no foco local, entende-se que um sintagma que aparece ligado diretamente ao verbo principal na árvore sintática de uma sentença está mais evidente que um sintagma distante deste. Como a raiz da árvore é sempre o verbo principal quanto mais alto na árvore um sintagma estiver ele estará mais próximo do verbo principal e portanto será maior sua evidência no texto. Como a utilização de anáforas pronominais se propõe a ser um elemento que auxilie no entendimento do texto como um todo, é pos-

sível que o autor de um texto evidencie o antecedente anafórico inserindo-o próximo ou ligado ao verbo principal de uma sentença. Este critério atribui como valor a distância do candidato ao verbo principal da sentença em que ele está inserido multiplicado por -1, ou seja, a quantidade mínima de arestas entre o candidato e a raiz da árvore sintática * -1. O maior valor atribuído é igual a -1 e o menor é igual a profundidade da árvore sintática multiplicado por -1. O peso deste critério é igual a 15. Valor do critério: $V = (x * -1) * 10$. Limite de corte: $x > 10$. Onde $x =$ quantidade de arestas do menor caminho entre o candidato e o verbo principal da sentença e $V =$ pontuação atribuída ao candidato na avaliação deste critério.

6. Semelhança sintática: Alguns trabalhos indicam que para alcançar maior clareza no texto e melhor entendimento por parte dos leitores, procura-se utilizar o antecedente anafórico e o pronome com um mesmo papel sintático dentro das sentenças. Isto auxilia na interpretação da anáfora incluindo um ponto a mais de relacionamento entre o antecedente anafórico a a anáfora pronominal. Por exemplo: *O carro passou em alta velocidade. Ele estava com a porta aberta.* Principalmente na função de sujeito, este tipo de construção anafórica se torna muito clara no momento da resolução mental da anáfora. O critério utilizado aqui atribui o valor 1 se o pronome e o candidato a antecedente anafórico possuírem a mesma função sintática. O peso deste critério é igual a 20. Não há limite de corte. O maior valor atribuído é 1 e o menor é 0. Valor do critério: $V = (x) * 20$. Onde $x = 1$ se o candidato possui mesma função sintática que o pronome ou 0 caso contrário e $V =$ pontuação atribuída ao candidato na avaliação deste critério.

7. Restrição de Reinhart: Reinhart mostrou que alguns empecilhos sintáticos podem ajudar na escolha do antecedente anafórico [Rei83]. Foram implementados duas formas sintáticas que apontam para a não existência de relação anafórica entre um sintagma e um pronome incluindo a noção de Dominância, Dominância Imediata e C-Comando existente na árvore sintática de uma sentença. Dominância ocorre quando o caminho de um sintagma A até um sintagma B é sempre para baixo, ou seja, A é ancestral de B. Dominância Imediata ocorre quando A é pai de B diretamente. A relação de C-Comando entre sintagmas ocorre quando o nó que domina imediatamente o sintagma A também domina não imediatamente o sintagma B, assim temos que, A C-Comanda B. A restrição para ocorrência de anáfora se dá das seguintes formas: uma anáfora pronominal não C-Comanda o antecedente anafórico e o antecedente anafórico não domina a anáfora pronominal. Por exemplo na sentença *Ele era o dono da casa, quando o ladrão entrou.* O pronome “Ele” C-Comanda o sintagma “ladrão”, indicando que não há relação anafórico entre esses termos. Em *João bateu nele.* o sintagma “João” domina o pronome “nele”,

evidenciando que não há relação anafórico entre os termos citados. Quando um candidato se enquadra em uma destas formas lhe é atribuído o valor -1. O peso do critério é igual a 10 e não há um limite de corte. Valor do critério: $V = (x) * 10$. Onde $x =$ -1 se o pronome C-Comanda o candidato, -1 se o candidato domina o pronome ou 0 caso contrário e $V =$ pontuação atribuída ao candidato na avaliação deste critério.

8. Pontuação geral: A atribuição de pesos para cada critério se deu de forma empírica. De acordo com a quantidade de antecedentes anafóricos que é pontuado corretamente em cada critério. Os critérios que mais acertam atribuindo os maiores valores ao antecedente correto também são os mais vagos, que também atribuem os maiores valores a vários candidatos. A tabela 5.3 exibe os pesos e limites de corte de cada critério. Somente os candidatos que possuírem valor igual ou maior ao limite de corte de cada critério permanecem na lista de candidatos (excluindo-se os casos onde o tamanho da lista se torna menor que 2). A soma dos pesos de cada critério forma a pontuação geral do candidato. Os quatro candidatos que possuírem as maiores pontuações gerais são indicados pelo algoritmo como possíveis antecedentes anafóricos. As relações entre os pesos atribuídos aos critérios refletem o grau de certeza com que cada critério indica o antecedente anafórico, de um modo geral, em um *corpus* literário diversificado.

Critério	**Peso**	**Limite de Corte**
Concordância em Gênero e Número	10	< 2
Ocorrência Anterior	10	< 1
Proximidade no Texto	15	> 50
Proximidade na Árvore	35	> 10
Altura na Árvore	15	> 10
Semelhança Sintática	20	não há
Restrição de Reinhart	10	não há

Tabela 5.3: Tabela dos pesos dos critérios sintáticos

De acordo com os testes realizados a taxa de acertos da parte sintática do algoritmo em escolher quatro candidatos onde um deles é o antecedente anafórico correto nos 118 casos analisados é próxima a 100%. Após reduzir o domínio dos candidatos para quatro sintagmas chega-se a fase semântica do algoritmo. Esta fase é constituída por dois critérios semânticos. O primeiro deles utiliza o Cálculo de Relacionamento Inferencial do SIA [Pin9a] para atribuir um valor ao candidato. O segundo critério semântico atribui ao candidato um valor de acordo com a quantidade de relacionamentos que este possui que são semelhantes ao verbo ligado ao pronome. A seguir estes dois critérios são melhor explicados.

1. Relacionamento Inferencial: Este critério calcula o relacionamento inferencial entre sintagmas de acordo com o modelo proposto pelo SIA [Pin9a]. Para cada candidato é calculado o relacionamento deste com todos os sintagmas nominais existentes na sentença onde o pronome está inserido. Cada valor de relacionamento é normalizado entre 0 e 1 e então somados para formar o valor de relacionamento inferencial do candidato. Este critério utiliza peso 1 e não há limite de corte. Valor do critério: $V = (x) * 1$. Onde $x =$ somatório dos valores de relacionamentos inferenciais e $V =$ pontuação atribuída ao candidato na avaliação deste critério.

2. Relação com o Verbo: Neste critério o candidato é analisado de acordo com o verbo ligado ao pronome. De acordo com o SIM [PAP+08] os relacionamentos de precondição e pós-condição são representados por arestas cujos rótulos determinam o papel do relacionamento. O tipo da relação é quem caracteriza a relação inferencial como uma pré ou pós condição de um conceito. Uma relação inferencial é da forma "X papel Y" e explicita um relacionamento entre os conceitos X e Y, por exemplo "motor" parte de "carro", onde significa que o conceito "motor" faz parte do conceito "carro". Os papéis existentes no banco de dados do SIM até o momento são os seguintes [PPFF10]: *DefinedAs* (definido como), *PartOf* (parte de), *UsedFor* (usado para), *CapableOf* (capaz de), *DesirousEffectOf* (efeito desejado de), *EffectOf* (efeito de), *MotivationOf* (motivação de), *SubEventOf* (sub-evento de), *CapableOfReceivingAction* (capaz de receber ação), *FirstSubEventOf* (primeiro sub-evento de), *LastSubEventOf* (último sub-evento de), *PreRequisiteEventOf* (evento pré-requisito de), *DesireOf* (desejo de), *LocationOf* (localizado em), *PropertyOf* (propriedade de), *IsA* (é um) e *MadeOf* (feito de) como visto na tabela 5.5.

   Foram analisados a ocorrência de 82 verbos diferentes nos casos analisados. Cada verbo foi relacionado a um conjunto de papéis de acordo com as características dos sintagmas que são articulados por tal verbo. Por exemplo o verbo "ser" foi relacionado aos papéis *DefinedAs, PartOf, IsA, PropertyOf e MadeOf* visto que as sentenças onde ocorrem "pronomes" ligados ao verbo "ser" podem ser: "Ele é um estrategista.", "Ele é uma peça do xadrez.", "Ele é um esporte.", "Ele é quadrado." e "Ele é feito de madeira." respectivamente aos papéis enunciados.

   Para os testes realizados foram classificados os seguintes verbos: "negar", "ser", "transpor", "dar", "clicar", "afastar", "colocar", "sugerir", "ter", "falar", "seguir", "responder", "ficar", "morrer", "haver", "ajudar", "antipatizar", "anunciar", "assistir", "atrair", "avaliar", "chegar", "comemorar", "comprar", "conceder", "conseguir", "cuidar", "declinar", "dever", "dizer", "elaborar", "encaminhar", "entravar", "estar", "explicar", "faltar", "fazer", "filmar", "funcionar", "gostar", "imperar", "impor", "impulsionar", "ir", "jogar",

"votar", "viver", "levar", "ligar", "limitar", "marcar", "obrigar", "perder", "permanecer", "plantar", "poder", "precisar", "pretender", "prever", "proferir", "propor", "quebrar", "queimar", "reduzir", "representar", "respeitar", "respirar", "reunir", "saber", "sair", "sobreviver", "terminar", "tirar", "tornar", "tratar", "usar", "ver", "andar", "miar", "entrosar", "entrosar", "morar"

Com base nisto foram calculados quantos relacionamentos utilizando os papéis relacionados ao verbo ligado ao pronome os candidatos possuem. Esta quantidade é o valor atribuído ao candidato neste critério. Aqui também não há limite de corte e o peso utilizado por este critério é igual a 1000. Valor do critério: $V = (x) * 1000$. Onde x = quantidade de relacionamentos utilizando os papéis relacionados ao verbo ligado ao pronome que o candidato possui e $V =$ pontuação atribuída ao candidato na avaliação deste critério.

3. Pontuação Final: São então somadas as pontuações atribuídas pelos critérios semânticos visto na tabela 5.4 e a lista é ordenada de forma decrescente. O primeiro candidato da lista é apontado como o antecedente anafórico do pronome em questão. Nos 118 casos analisados o algoritmo obteve êxito em 91 casos, ou seja, uma taxa de acerto em torno de 77,1% em textos extraídos de domínios variados.

Critério	**Peso**
Calculo de Relacionamento Inferencial	1
Relação com o Verbo	1000

Tabela 5.4: Tabela dos pesos dos critérios semânticos

Um exemplo de um texto sendo processado pelo algoritmo desenvolvido aqui é visto a seguir.

Texto: *O mundo enfrenta uma grande crise. Os trabalhadores ganham pouco dinheiro. João é o jardineiro do rei e ele mora em um casebre.*

O algoritmo retira como candidato a antecedente da anáfora os seguintes sintagmas:

- mundo
- crise
- trabalhadores
- dinheiro
- João

Papel	Tradução Possível
X DefinedAs Y	X é definido como Y
X PartOf Y	X é parte de Y
X UsedFor Y	X é usado para Y
X CapableOf Y	X é capaz de Y
X DesirousEffectOf Y	X possui o efeito desejado Y
X EffectOf Y	X possui o efeito Y
X MotivationOf Y	X é motivação para Y
X SubEventOf Y	X é um sub-evento de Y
X CapableOfReceivingAction Y	X é capaz de receber a ação Y
X FirstSubEventOf Y	X é o primeiro sub-evento de Y
X LastSubEventOf Y	X é o último sub-evento de Y
X PreRequisiteEventOf Y	X é um evento pré-requisito de Y
X DesireOf Y	X é desejo de Y
X LocationOf Y	X está localizado em Y
X PropertyOf Y	X é uma propriedade de Y
X IsA Y	X é um Y
X MadeOF Y	X é feito de Y

Tabela 5.5: Tabela dos papéis presentes nos relacionamentos entre os conceitos X e Y

- jardineiro
- rei
- casebre

Cada critério sintático é aplicado aos candidatos. São atribuídos os valores de cada candidato em cada critério e calculado o total da pontuação de cada sintagma candidato a antecedente anafórico.

sintagma	**(1)*10**	**(2)*10**	**(3)*15**	**(4)*35**	**(5)*15**	**(6)*20**	**(7)*10**	**geral**
mundo	2	1	-19	-5	-3	2	0	-435
crise	1	-	-	-	-	-	-	-
trabalhadores	1	-	-	-	-	-	-	-
dinheiro	2	1	-9	-4	-2	0	0	-275
João	2	1	-7	-3	-1	2	0	-155
jardineiro	2	1	-4	-3	-1	0	0	-150
rei	2	1	-2	-5	-3	0	0	-220
casebre	2	0	-	-	-	-	-	-

Tabela 5.6: Tabela dos critérios sintáticos

Após a aplicação dos critérios sintáticos são selecionados os quatro candidatos com maior pontuação e inicia-se o processamento dos critérios semânticos. (1) Relação inferencial com os sintagmas ligados ao pronome e (2) relação com o verbo ligado ao pronome.

sintagma	**(1) * 1**	**(2) * 1000**	**geral**
João	65.115	623	623065
dinheiro	0.013	164	164000
rei	741.729	46	46741
jardineiro	313.015	37	37313

Tabela 5.7: Tabela dos critérios semânticos

De acordo com a tabela 5.7 o sintagma "João" é o mais relacionado ao pronome e portanto é o escolhido como o antecedente anafórico.

5.3 Resultados obtidos

A parte sintática do algoritmo obteve uma taxa de acerto semelhante às outras abordagens atuais analisadas. Quando utilizado para selecionar somente um sintagma candidato e apresentá-lo como o antecedente anafórico do pronome o algoritmo sintático obteve 47,5% de acerto.

Após a análise de algumas estratégias a que apresentou melhores resultados foi a de afrouxar o algoritmo sintático e selecionar os quatro candidatos mais prováveis e passar esse conjunto de quatro sintagmas para o algoritmo semântico selecionar o antecedente anafórico.

O algoritmo semântico possui dois critérios que tratam da relação entre os conceitos ligados ao pronome e o candidato a antecedente anafórico. Quando utilizado o algoritmo sintático e em seguida o algoritmo semântico a taxa de acerto subiu para 77,1%. Uma tentativa de teste utilizando somente a parte semântica do algoritmo não foi possível, pois o tempo gasto para realizar tais testes seria demasiadamente grande. Sem os critérios sintáticos, a quantidade de sintagmas candidatos a antecedente anafórico chega em média a 20 candidatos por sentença, número bem maior que os 4 selecionados pelo algoritmo sintático.

Uma comparação entre um algoritmo puramente sintático e o algoritmo sintático-semântico proposto neste trabalho pode ser visto na forma de uma pequena modificação no algoritmo desenvolvido aqui. A seguir é explanada um exemplo de resolução de anáfora onde será selecionado o candidato melhor colocado na fase sintática do algoritmo e o candidato selecionado ao final do algoritmo. Seja o seguinte texto:

Um anti-exemplo deste princípio pode ser visto em uma das cenas mais decisivas do filme "O Advogado do Diabo", quando o diabo apresenta, ao advogado, a opção deste deixar suas atribuições profissionais para cuidar da sua esposa que estava emocionalmente fragilizada. Talvez inspirado pelo imperativo "do it" se pediu, aguenta, ***ele*** *nega-se a abandonar seus*

deveres profissionais respondendo que pode dar conta de ambos. Desta forma, o diabo obtém assim permissão para atingir o seu lado mais frágil (simbolizado por sua esposa) e fazer a devastação moral.

A partir deste texto são extraídos os candidatos a antecedente da anáfora representada aqui pelo pronome pessoal **ele**. Os candidatos são nomeadamente: “anti-exemplo”, “princípio”, “cenas”, “filme”, “O=Advogado=do=Diabo”, “diabo”, “advogado”, “opção”, “atribuições”, “esposa”, “imperativo”, “deveres”, “conta”, “permissão”, “lado”, “esposa” e “devastação”.

Agora os critérios sintáticos serão aplicados aos candidatos. Após o primeiro critério, Concordância em gênero e número, e o segundo critério, Ocorrência anterior, temos os seguintes candidatos e seus valores obtidos:

- anti-exemplo : valor = 30
- princípio : valor = 30
- filme : valor = 30
- O=Advogado=do=Diabo : valor = 30
- diabo : valor = 30
- advogado : valor = 30
- imperativo : valor = 30

Todos os demais candidatos são eliminados nestes dois primeiros critérios. Os outros critérios sintáticos são: (3) Proximidade no texto, (4) Proximidade na árvore, (5) Altura na árvore, (6) Semelhança sintática e (7) Restrição de Reinhart.

sintagma	**3**	**4**	**5**	**6**	**7**	**geral**
anti-exemplo	<-50	-	-	-	-	-
princípio	-750	-210	-60	0	0	-990
filme	-555	<-10	-	-	-	-
Advogado=do=Diabo	-540	<-10	-	-	-	-
diabo	-495	-245	-75	20	0	-765
advogado	-435	-280	-90	0	0	-775
imperativo	-90	-35	-15	0	0	-110

Tabela 5.8: Tabela dos critérios sintáticos

Como visto na tabela 5.8, o algoritmo puramente sintático escolherá o sintagma “imperativo” e concluirá erroneamente a resolução da anáfora. Passaremos então a fase semântica

do algoritmo com os quatro candidatos com maior pontuação na fase sintática: “imperativo”, “diabo”, “advogado” e “princípio”.

Cada critério semântico também atribui aos candidatos valores de acordo com a relação deles com os sintagmas ligados ao pronome no primeiro critério e com o verbo ligado ao pronome no segundo critério [Tabela 5.9].

sintagma	**1**	**2**	**geral**
princípio	0	1	1000
diabo	0	0	0
advogado	0	12	12000
imperativo	0	0	0

Tabela 5.9: Tabela dos critérios semânticos

Agora vê-se que a fase semântica do algoritmo encontrou corretamente o antecedente anafórico do pronome “ele”, atribuindo ao sintagma “advogado” a relação anafórica com tal pronome.

A seguir vê-se ainda um outro exemplo da superioridade do algoritmo sintático-semântico:

O menino comprou um gato. Ele andava muito triste.

O menino comprou um gato. Ele miava muito triste.

Note que a única diferença entre as sentenças é o conteúdo do verbo na segunda frase. É exatamente este conteúdo que define a qual sintagma o pronome se refere. Um algoritmo puramente sintático não conseguirá distinguir entre as sentenças acima e fatalmente errará uma das resoluções de anáfora. O algoritmo desenvolvido neste trabalho, por outro lado, diferencia as sentenças porque o verbo “andar” está mais associado ao conceito “menino” enquanto que o verbo “miar” está mais associado ao conceito “gato” dentro da prática linguística resolvendo corretamente ambos os casos de anáfora.

Analisando que os textos utilizados nos testes foram retirados de *corpora* de domínios bem variados e temas próximos a dinâmica da comunicação humana o algoritmo obteve resultados melhores que os trabalhos relacionados tratados aqui. Aponta-se então que dada a novidade e carência de trabalhos em PLN que se utilizam de informações semânticas inferencialistas, este caminho de pesquisa pode trazer muitos frutos no entendimento da linguagem natural bem como na construção de sistemas que consigam extrair mais informações da linguagem natural de forma semântica e pragmática.

5.4 Análise dos resultados

O algoritmo desenvolvido com base em informações sintáticas obteve uma taxa de acerto sobre os textos utilizados em torno de 44,5%. Esta taxa está em conformidade com a aplicação dos trabalhos atuais em resolução de anáforas [Coe06] e [ACC+04]. Influenciados pela abordagem inferencialista em Processamento de Linguagem Natural de Pinheiro [PPFF10] foi desenvolvido então um algoritmo que recebe do algoritmo sintático os quatro candidatos mais prováveis de serem o antecedente anafórico do pronome e sobre eles atua mais dois critérios semânticos de escolha.

O primeiro critério semântico desenvolvido se refere ao Cálculo de Relacionamento Inferencial de Pinheiro e relaciona cada candidato a todos os sintagmas ligados ao pronome em questão. O resultado de cada relacionamento é somado e é atribuído este valor ao candidato. O segundo critério semântico desenvolvido relaciona os candidatos ao verbo a qual o pronome está diretamente relacionado. Os verbos foram descritos em função dos rótulos das relações existentes entre os conceitos. Cada rótulo indica uma forma em que um conceito pode ser utilizado. Dessa forma os rótulos relativos a um verbo formam um grupo de possibilidades de uso do verbo.

Quando comparamos as formas em que um conceito representado por um verbo pode ser utilizado com as formas em que um conceito representado por um sintagma nominal pode ser utilizado estamos nos referindo as possibilidades das formas em que o sintagma pode ser utilizado em conjunto com tal verbo. Se há uma quantidade maior de usos em comum entre um sintagma e um verbo, então entende-se que tal sintagma seja utilizado em conjunto com o verbo em questão. Na prática se o conceito representado pelo verbo “X” é descrito como “X Localizado em Y” então o verbo “X” tem como característica semântica que ele representa uma ação que está localizada em um certo local “Y”. Quando comparamos este conceito representado pelo verbo “X” com um conceito representado por um sintagma que possui várias relações rotuladas por “Localizado em” entende-se que as formas de uso do verbo estejam bem relacionadas as formas de uso do sintagma e por consequência o verbo está bem relacionado ao sintagma. Assim, foram contabilizados quantos relacionamentos através dos rótulos descritos para cada verbo os candidatos possuem. Quanto mais relacionamentos, mais próximo está o candidato do verbo e por consequência também do pronome.

Após a utilização destes dois critérios semânticos sobre os quatro candidatos selecionados na fase sintática é escolhido o candidato com maior pontuação recebida na fase semântica e este é apontado com o antecedente anafórico do pronome. A utilização do algoritmo sintático-

semântico obteve uma taxa de acerto de 77,1%.

Os textos utilizados para os testes foram obtidos dos testes do SIA [PPFF10] denominada "Coleção Dourada" e dos primeiros trechos dos *corpora* Bosque e Amazônia e estão disponível no endereço eletrônico:

https://docs.google.com/Doc?docid=0ARGXRjw55R-PZGRucGM5bWhfMTVoaHQ4OHFoYw&hl=en

Não foi possível, no tempo proposto para o fim do trabalho, comparar os resultados obtidos no algoritmo desenvolvido aqui com os resultados obtidos nos algoritmos apresentados em trabalhos paralelos diante de um mesmo conjunto de textos. Esta impossibilidade se deu pela indisponibilidade dos conjuntos de textos utilizados pelos outros trabalhos ou pela ausência de textos em português de domínio comum e que houvesse sido analisado pelo *parser* PALAVRAS no formato *CG-dependency*, que é o utilizado pelo algoritmo desenvolvido neste trabalho. Com isto tentou-se contornar este problema através da utilização de textos bem variados de domínios de conhecimentos amplos e que traduzissem a prática linguística de uma comunidade como textos de um *blog*, descrições de crimes e textos jornalísticos de diversas áreas.

Note que apesar dos conceitos e suas relações estarem representados no sistema, a atribuição de significado é sempre realizada através das inferências realizadas com estes conceitos e as bases de conhecimento armazenadas são extraídos diretamente da prática linguística.

De acordo com o uso dos conceitos e de como eles são articulados conhece-se o significado das sentenças e das palavras. Entende-se assim, que o significado é dado em função do uso das sentenças, de forma pragmática. Abre-se um novo caminho para a resolução de problemas na área de PLN através de informações semânticas inferencialistas.

Esta nova abordagem para a resolução de anáforas pronominais utilizando um algoritmo parte sintático e parte semântico mostra que a parte semântica é importante para obtenção de melhorias em relação aos resultados já alcançados.

6 Conclusão

Este trabalho mostra como ocorre uma relação anafórica entre um pronome pessoal e um sintagma nominal. Fica claro que as anáforas incluem aspectos muito importantes da linguagem natural, incluindo o fato delas serem entendidas semanticamente (e pragmaticamente) através do conhecimento acumulado que a mente humana possui. Como uma anáfora é, em essência, um elemento textual que relaciona conceitos, a resolução de uma anáfora está intrinsecamente ligada ao entendimento das relações entre os conceitos [Mon94].

No capítulo 1 entende-se que a linguagem pode ser entendida pragmaticamente. Brandom [Bra00] descreve uma forma de entender as relações entre os termos e o significado dos termos com base na estrutura global do discurso. Em outras palavras, o significado de cada parte do texto está ligada ao significado do texto como um todo. Não é possível construir o significado de uma sentença a partir da junção dos significados particulares de cada termo componente da sentença. Brandom argumenta que os conceitos só são entendidos a partir do todo, esboçando uma abordagem *top-down* que ganha força ao tratar os conceitos dentro de uma rede de conceitos interligados por premissas e conclusões. Entende-se também que o significado de um texto está ligado a uma rede semelhante a dos conceitos porém formada por sentenças que se interligam por arestas rotuladas de premissas e condições. O entendimento de um texto então se completa ao compreender quais são as premissas de cada sentença e quais as conclusões advindas da enunciação de uma sentença.

Como a atribuição de significado está atrelado à construção da rede inferencial dos conceitos articulados nas sentenças e sendo esta rede inferencial determinada pela prática linguística de uma determinada comunidade, pode-se então denominar tal significado de pragmático. Baseando-se nessa linguística pragmática Pinheiro [Pin9a] inicia a construção de um modelo baseado em redes inferenciais de conceitos e sentenças, o SIM (Semantic Inferencialism Model). O SIM admite que entender os conceitos só é possível quando primeiro se conhece como eles funcionam na prática. O modelo ainda descreve como representar os conceitos sem contudo representar o seu significado. Os conceitos são armazenados em linguagem natural e são interligados por precondições e pós-condições existentes na prática linguística e armazenados

em forma de um grafo. A mesma coisa acontece com as sentenças que são armazenadas como formas de relacionar conceitos em torno de um verbo, as sentenças-padrão são possíveis relações entre conceitos também armazenados em forma de um grafo. Pinheiro [PPFF10] então desenvolve um raciocinador inferencial capaz de tratar a linguagem natural comparando-a com as informações existentes nos grafos de conceitos e sentenças-padrão e extrair inferências contendo informações que estavam implícitas no texto original.

Este trabalho então focou-se em resolver anáforas pronominais utilizando o arcabouço teórico inferencialista advindo do SIM e SIA (Semantic Inferentialist Analyser) [PPFF10]. Ao descrever o funcionamento de uma anáfora notou-se que este fenômeno linguístico envolve a compreensão dos conceitos relativos aos sintagmas e verbos existentes nas sentenças. Uma anáfora pronominal é possuidora de uma função dêitica textual que remete ao próprio texto e indica a qual sintagma o pronome enunciado se refere. Um pronome não insere um novo conceito no texto, pelo contrário, se refere a um conceito já existente no texto adicionando informações sobre ele. Também foi visto que para se compreender corretamente uma sentença se faz necessário resolver as anáforas pronominais existentes para obter-se o conhecimento de qual conceito está por trás da anáfora. Foi então desenvolvido um algoritmo sintático que utiliza conhecimento adquirido de outras abordagens paralelas a este trabalho. Existe uma variedade de algoritmos que se propõem a resolver anáforas pronominais que em geral se utilizam apenas de informações sintáticas. Por um lado as abordagens puramente sintáticas ganham em velocidade e robusteza, mas por outro, não contemplam a existência de anáforas pronominais regidas por informações semânticas. Como visto, alguns linguistas como Monteiro [Mon94] e Levinson [Lev87] recomendam o entendimento das anáforas de um ponto de vista semântico e pragmático pois a função dêitica existentes nas anáforas se utilizam de informações deste tipo.

6.1 Trabalhos futuros

Como visto inicialmente, pretende-se integrar o projeto WIKICRIMESIE [Pin9a] com a resolução de anáfora incorporada ao seu algoritmo. De fato o algoritmo descrito neste trabalho já está inserido no sistema que realiza a análise do texto no WIKICRIMESIE, porém ainda não foi testado conjuntamente com ele. A resolução de anáfora pode trazer à tona novas relações entre sentenças e consequentemente novas inferências podem ser feitas adicionando informações úteis na interpretação do texto.

Ao invés de só analisar o relacionamento entre conceitos, pretende-se utilizar também o relacionamento entre as sentenças existentes e as possíveis sentenças formadas pela substituição

do pronome pelos sintagmas candidatos a antecedente anafórico.

O critério Relação com o Verbo não utiliza o fato de que as relações com o verbo pode se dar de forma ativa e passiva. Assim, poder-se-ia separar as semelhanças com o verbo de forma ativa e de forma passiva separadamente, especificando mais ainda quais sintagmas são relacionados a cada verbo encontrado. Os verbos que foram classificados através dos papéis dos relacionamentos entre conceitos seriam melhor caracterizados se eles estivessem inseridos na própria rede de conceitos. Porém, até o momento, a base de conceitos contêm apenas substantivos simples e compostos.

A melhoria das bases evita muitos relacionamentos inferenciais com resultado de valor 0.0, que foi frequentemente encontrado no critério Relacionamento Inferencial. Dentre os casos onde o algoritmo avaliado falhou em encontrar o antecedente anafórico correto 12 casos foram prejudicados pela ausência do conceito desejado na base de conceitos. No caso dos nomes próprios que, em geral eram substituídos pelos conceitos “pessoa”, “homem” ou “lugar”. Viu-se ainda um outro problema, ao substituir tais conceitos muitos nomes próprios se igualam e se torna impossível de avaliar qual é o mais próximo do pronome. Acerca deste problema, dos 118 casos analisados encontra-se 11 casos onde o antecedente anafórico não foi encontrado corretamente por ter-se igualado a um sintagma errado e ocasionar a impossibilidade de distinção entre ambos, acarretando na escolha final do sintagma errado.

Ainda há a possibilidade de trabalhar com relacionamentos negativos entre conceitos, ou seja, a possibilidade de casos onde dado que um conceito “X” esteja inserido no texto de uma determinada forma “Y”, é improvável ou desaconselhável a existência de um segundo conceito “Z” de uma determinada forma “K”. Também pode-se tratar o contexto onde a anáfora está inserida como uma espécie de memória temporária utilizando os elementos existentes neste contexto como as sentenças e as relações entre os sintagmas e os verbos para auxiliar no cálculo de relacionamento inferencial e no relacionamento com o verbo. A aprendizagem das bases através do uso do sistema (práxis linguística) trataria então de como transformar esta memória temporária em permanente e como filtrar quais informações receberiam este tratamento.

Referências Bibliográficas

[ABHS01] Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. Floresta sintá(c)tica: um "treebank" para o português. *APL*, 2001.

[ACC+04] Ana Aires, Jorge Coelho, Sandra Collovini, Paulo Quaresma, and Renata Vieira. Avaliação de centering em resolução pronominal da língua portuguesa. In *5th International Workshop on Linguistically Interpreted Corpora of the Iberamia*, pages 1–8, Puebla, México, November 2004.

[Ant96] Grigoris Antoniou. *Nonmonotonic Reasoning*. Artificial Intelligence. MIT Press, Cambridge, MA, USA, 1996.

[BFP87] Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proceedings of the 25th Conference on Association for Computational Linguistics*, pages 155–162, Stanford, California, 1987.

[Bic00] Eckhard Bick. *The Parsing System PALAVRAS: Automatic Gramatical Analysis of Porutugese in a Constraint Grammar Framework*. Aarhus University Press, 2000.

[Bra94] Robert Brandom. *Making it Explicit*. Harvard University Press, Cambridge, MA, USA, 1994.

[Bra00] Robert Brandom. *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, Cambridge, MA, USA, 2000.

[CCV05] Sandra Collovini, J. C. B. Coelho, and Renata Vieira. Classificação automática de expressões anafóricas em textos da língua portuguesa. *XXV Congresso da SBC*, 2005.

[Cha07] Amanda Rocha Chaves. A resolução de anáforas pronominais da língua portuguesa com base no algoritmo de mitkov. Master's thesis, Universidade Federal de São Carlos, 2007.

[CK92] Daniel Coulon and Daniel Kayser. *Informática e Linguagem Natural: uma Visão Geral dos Métodos de Interpretação de Textos Escritos*. IBICT, Brasília, DF, BR, 1992.

[Coe06] Thiago Thomes Coelho. Resolução de anáfora pronominal em português utilizando o algoritmo de lappin e leass. Master's thesis, Universidade Estadual de Campinas, 2006.

[de 04] Kátia Cristina Cavalcante de Soares. Uso de estratégias de compreensão de anáforas em artigos de opinião. Master's thesis, Universidade Federal do Ceará, 2004.

[dF05] Sérgio Antônio Andrade de Freitas. *Processamento Automatizado de Textos: Processamento de Anáforas*. PhD thesis, Universidade Federal do Espírito Santo, 2005.

[dRAdOCT+05] Isa Maria da Rosa Alves, Rove Luiza de Oliveira Chishman, Paulo Miguel Torres, Duarte Quaresma, and José Saias. Busca e extração de informações através de pergunta e resposta: uma nova concepção de web. *XXV Congresso da SBC*, 2005.

[Dre07] Hubert L. Dreyfus. Why heideggerian ai failed and how fixing it would require making it more heideggerian. *Artificial Intelligence*, (171):1137–1160, 2007.

[dS04] Adriana da Silva. *A Leitura e Compreensão da Anáfora Conceitual*. PhD thesis, Universidade Estadual de Campinas, 2004.

[dS08] José Guilherme Camargo de Souza. Resolução automática de correferência aplicada à língua portuguesa. Master's thesis, Universidade do Vale do Rio dos Sinos, 2008.

[Dum78] Michael Dummett. *Truth and Other Enigmas*. Harvard University Press, 1978.

[Fel98] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[GGV03] Caroline Gasperin, Rodrigo Goulart, and Renata Vieira. Uma ferramenta para resolução automática de correferência. *IV Encontro Nacional de Inteligência Artificial*, 2003.

[GJW95] B. J. Grosz, A. K. Joshi, and S. Weinstein. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 12:203–226, 1995.

[Hab04] Jürgen Habermas. *Verdade e Justificao: Ensaios Filosóficos*. Edies Loyola, São Paulo, SP, BR, 2004.

[Hei96] Martin Heidegger. *Being and Time, Translated by Joan Stambaugh*. State University of New York Press, 1996.

[HH76] M. A. K. Halliday and R. Hasan. Reference. *Cohesion in English*, pages 55–60, 1976.

[Lah79] Michel Lahud. *A Propósito da Noção de Dêixis*. Ática, São Paulo, SP, Brasil, 1979.

[Lea94] Herbert J Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20:535–561, 1994.

[Lef01] Vilson J. Leffa. A resolução da anáfora no processamento da língua natural. Technical report, Universidade Católica de Pelotas, 2001.

[Lef03] Vilson J. Leffa. Anaphora resolution without world knowledge. *D.E.L.T.A.*, 19:181–200, 2003.

[Len95] Douglas B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communication of the ACM*, 38:33–38, 1995.

[Lev87] S. Levinson. Pragmatics and the grammar of anaphora: a partial pragmatic reduction of binding and control phenomena. *Pragmatics*, 23:379–434, 1987.

[LS04] Hugo Liu and Push Singh. Commonsense reasoning in and over natural language. In *Knowledge-Based Intelligent Information and Engineering Systems*, Lecture Notes in Computer Science, pages 293–306. Springer Berlin / Heidelberg, 2004.

[Lyo77] John Lyons. Deixis, space and time. *Semantics*, 2:636–724, 1977.

[MBMA+99] Patricio Martínez-Barco, Rafael Muñoz, Saliha Azzam, Manuel Palomar, and Antonio Ferrández. Evaluation of pronoun resolution algorithm for spanish dialogues. In *In Proceedings of the Venezia*, 1999.

[MCa07] John MCarthy. From here to human-level ai. *Artificial Intelligence*, 171:1174–1182, 2007.

[Mon94] José Lemos Monteiro. *Pronomes Pessoais: Subsídios para uma Gramática do Português do Brasil*. Edições UFC, Fortaleza, CE, BR, 1994.

[MS98] Huslan Mitkov and Malgorzata Stys. Robust reference resolution with limited knowledge. volume 17, 1998.

[Mul01] Ana Muller. Anáfora pronominal. *Revista Letras*, 56:259–275, 2001.

[PAP+08] Vládia Célia Monteiro Pinheiro, Thiago Assunção, Tarcisio Pequeno, Vasco Furtado, and Emanoel Freitas. Sim: Um modelo semântico-inferencialista para sistemas de linguagem natural. *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2008), WebMedia*, 2008.

[Par97] Ivandré Paraboni. Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul, 1997.

[Pin9a] Vládia Célia Monteiro Pinheiro. Semantic inferentialism analyser: Um analisador semântico de sentenças em linguagem natural. *STIL*, 2009a.

[Pin10] Vládia Célia Monteiro Pinheiro. *SIM: Um Modelo Semântico Inferencialista para Expressão e Raciocínio em Sistemas de Linguagem Natural*. PhD thesis, Universidade Federal do Ceará, 2010.

[PPFF10] Vládia Pinheiro, Tarcisio Pequeno, Vasco Furtado, and Wellington Franco. Inferencenet.br: Expression of inferentialist semantic content of the portuguese language. *PROPOR*, 2010.

[PPFN09] Vladia Pinheiro, Tarcisio Pequeno, Vasco Furtado, and Douglas Nogueira. Information extraction from text based on semantic inferentialism. In *FQAS '09: Proceedings of the 8th International Conference on Flexible Query Answering Systems*, pages 333–344, Berlin, Heidelberg, 2009. Springer-Verlag.

[Rei83] Tanya Reinhart. Anaphora and semantic interpretation. *Croom Helm Ltd*, 1983.

[RPFV01] D. Rossi, C. Pinheiro, N. Feier, and R. Vieira. Resolução de correferência em textos da língua portuguesa. *Revista Eletrônica de Iniciação Científica*, 2001.

[San00] Victor Martins Sant'anna. Cálculo de referências anafóricas pronominais demonstrativas na língua portuguesa escrita. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul, 2000.

[Sel80] Wilfrid Sellars. *Inference and Meaning*. Pure Pragmatics and Possible Worlds. Ridgeview Publishing Co., Reseda, CA, USA, j. sicha edition, 1980.

[SH77] Ivan Sag and J. Hankamer. Syntactically versus pragmatically controlled anaphora. In *Studies in Language Variation: Semantics, Syntax, Phonology, Pragmatics, Social Situations, Ethnographic Approaches*, pages 120–155, Washington, 1977. Georgetown University Press.

[Tat06] Ian Tattersall. How we came to be human, becoming human: Evolution and the rise of intelligence. *Scientific American Special Edition*, 2006.

[Tei77] Raquel F. A. Teixeira. Pronomes pessoais sujeitos em português: uma abordagem gerativo-transformacional. Master's thesis, UNB, Brasília, 1977.