SÉRGIO ANTÔNIO ANDRADE DE FREITAS

INTERPRETAÇÃO AUTOMATIZADA DE TEXTOS: PROCESSAMENTO DE ANÁFORAS

VITÓRIA

2005

SÉRGIO ANTÔNIO ANDRADE DE FREITAS

INTERPRETAÇÃO AUTOMATIZADA DE TEXTOS: PROCESSAMENTO DE ANÁFORAS

Tese apresentada ao Programa de Pós-Graduação em Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do Grau de Doutor em Engenharia Elétrica, na área de concentração em Automação.
Orientador: Dr. Crediné da Silva Menezes
Co-orientador: Dr. José Gabriel Pereira Lopes

VITÓRIA

2005

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Central da Universidade Federal do Espírito Santo, ES, Brasil)

Freitas, Sérgio Antônio Andrade de, 1967-

F866p Processamento automatizado de textos : processamento de anáforas / Sérgio Antônio Andrade de Freitas. - 2005.
184 f. : il.

Orientador: Crediné da Silva Menezes.
Co-Orientador: José Gabriel Pereira Lopes.
Tese (doutorado) - Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Processamento da linguagem natural (Computação). 2. Processamento de textos (Computação).3. Anáfora (Linguística). 4 Algoritmos de computador. I. Menezes, Crediné da Silva. II. Lopes, José Gabriel Pereira. III. Universidade Federal do Espírito Santo. Centro Tecnológico. IV. Título.

CDU: 621.3

Tese de Doutorado sob o título *"INTERPRETAÇÃO AUTOMATIZADA DE TEXTOS: PROCESSAMENTO DE ANÁFORAS"*, defendida por SÉRGIO ANTÔNIO ANDRADE DE FREITAS e aprovada em 11 de 2005, em Vitória, Estado do Espírito Santo, pela banca examinadora constituída pelos doutores:

Prof. Dr. Crediné da Silva Menezes
Orientador

Prof. Dr. José Gabriel Pereira Lopes
Co-orientador
Universidade Nova de Lisboa, Portugal

Profa. Dra. Rosa Maria Vicari
Universidade Federal do Rio Grande do Sul

Prof. Dr. Berilhes Borges Garcia
Universidade Federal do Espírito Santo

Prof. Dr. Davidson Cury
Universidade Federal do Espírito Santo

Prof. Dr. Orivaldo de Lira Tavares
Universidade Federal do Espírito Santo

Dedido esta tese a Deus,
às minhas luzes Tê, Mariana e Lucas,
à minha mãe Gilda, à minha irmã Sandra
e ao meu pai Antônio.

Agradecimentos

Acima de tudo e todos agradeço a Deus pela minha existência e por ter permitido que eu aqui esteja. Neste momento em especial, agradeço-Lhe também por ter apontado caminhos para o conhecimento *permitindo* que eu vislumbrasse os rumos do meu desconhecimento. Saber é importante, porém sinto que nada sei após a conclusão desta tese. Quão ignorante fui: um dia achei que sabia! Obrigado Pai por esta oportunidade de aprender a me conhecer melhor. Este foi o meu grande aprendizado.

Se por um lado esta é uma declaração de ignorância, por outro posso dizer que hoje me sinto mais alegre e forte. Não pelo que penso que aprendi, mas por saber que com trabalho posso aprender o que não sei. A visão do desconhecido é um forte convite a lançar-me a mares que ainda não naveguei. Oxalá eu possa ter sabedoria para empregar estas migalhas em proveito real do próximo. Senão que rédeas me sejam colocadas.

Nesta caminhada foram tantas as pessoas e os espíritos que me ajudaram que dificilmente não esquecerei alguém. Cara(o) amigo(a), peço desculpas de antemão por esta atitude de esquecimento, mas não de ingratidão.

Por entre marés nunca posso deixar de lembrar que nos momentos mais críticos só havia um par de pegadas na areia. Eram da Tê que silenciosamente me carregou. Junto a ela vejo o meu anjo da guarda, de quem sinto os conselhos a cada instante. Muito obrigado, muito obrigado, muito obrigado aos dois. Minha luz Tê trouxe ao mundo duas luzes as quais marcam, cada dia mais forte, o ritmo da minha existência atual: minha filha Mariana e meu filho Lucas. As crianças são um reino à parte. Será porque esquecemos disto?

Olhando para minhas mãos, lembro que um dia elas foram pequenas e as sinto nas mãos de uma grande mulher. Oh minha mãe Gilda quão bom é estar em seu colo e sentir seu carinho. Aqui vejo minha irmã Sandra, meu grande exemplo de determinação e perseverança. Se eu tivesse aprendido uma pequena parte do que você já é com certeza eu já não estaria aqui. Também lembro de meu pai que me ensinou o trabalho. Obrigado.

Nos caminhos da tese, a orientação do Prof. Gabriel permitiu-me desvencilhar de muitos caminhos tortuosos. Seus exemplos de amizade, dedicação ao trabalho e humanismo,

demonstrados a cada instante, estão sempre vivos em minha mente. Como orientandos sempre precisamos de capitães e mais uma vez você demonstrou ser um velho marujo, experiente e paciente. Obrigado por tudo, de coração.

Obrigado também ao Prof. Crediné, o qual me orientou por rumos que desconhecia. Nossas conversas permitiram-me desenvolver a percepção do que é um doutorado e, de uma maneira mais abrangente, porque/como fazer ciência. Foram estas conversas que desataram grandes nós em mim.

Existem anjos sobre a terra e eu tive o privilégio de conhecer um deles. Não que seja um anjinho de asas, mas sim um com as mãos sujas de terra e o rosto suado do trabalho honesto. Este tipo de anjo não dá conselhos, ele os faz. Mas quando você cruza com um deles fica logo sabendo. Eles mudam a sua vida simplesmente pela forma de ser. Obrigado Rosa.

Na classe dos anjos terrenos estão também todos os(as) amigos(as), os(as) quais me ajudaram direta e/ou indiretamente. Alguns estão só de passagem por aqui, pois o que tem para aprender já sabem, mas mesmo assim se colocam na posição de aprendiz para poder melhor ensinar. Obrigado Ayrton. Outros são tão discretos que parecem estar recebendo ajuda quando na verdade estão ajudando. Obrigado Celso. Outros têm um grande coração e uma amizade sincera que são difíceis de encontrar. Eles também mudam a sua vida pelo que dizem. Obrigado Raul. Outros pela clareza do raciocínio obtido a partir de um trabalho árduo e constante de estudo, aliados a leveza de coração resultam num manancial de críticas construtivas. Obrigado Berilhes. Há os que traduzem a mais bela expressão do viver, da busca da felicidade e dos sonhos de harmonia consigo mesmo. Estes também mudam nossos rumos. Obrigado Flávio.

Nos meus momentos de *não felicidade* sempre lembrei do meu amigo Paulo Jorge. Um mestre na arte do pensamento. Foram tantos os exemplos dados que ainda hoje não consegui aplicar metade deles. Que Deus o ilumine a cada dia. E também a sua mãe Tereza e seu pai Antônio. A minha grande amiga Sílvia Enoque que em vida me deu lições imorredouras de persistência e luta. No mais além continuou a ser mais do que uma amiga, se tornou irmã e passou a ser meu segundo anjo da guarda. A ela muito devo. Que ela possa agora alçar vôos mais altos nas novas esferas de liberdade e vida. Sílvia, foram seus exemplos e a sua presença nestes últimos tempos que permitiram terminar esta tese. Obrigado de coração.

Também não posso deixar de agradecer a todos os amigos de Portugal. Ao Paulo Quaresma e a Cati pela amizade. Os momentos que passamos juntos são inesquecíveis.

Só isto já valeu a pena o doutorado. A Irene Pimenta pelos conselhos e conversas. Ao Nuno Marques pelos nossos longos debates sobre assuntos *não técnicos*. À D. Lídia, Seu Carlos, a Maria dos Anjos e famílias pela amizade na mais bela expressão da fraternidade portuguesa. Enfim a todos os amigos do Departamento de Informática da FCT/UNL.

À Manuela e ao João pela amizade, pelas visitas, conversas e reuniões. Que Deus lhes ilumine os caminhos tão sofridos. Aos amigos do Perdão e Caridade em Lisboa: Carlos Alberto, Raquel, Seu Licínio, Berta, Antonieta, Virgínia, D. Idalina, D. Odete, Bento, entre outros que tão bem nos acolheram e aos quais considero como uma grande família.

Aos meus tios José Justino e Hilda pelo apoio incondicional e pelos conselhos. A Tia Nida pela confiança que em mim depositou. Nunca vou esquecer. Sou-lhe muito grato.

Ao pessoal do Departamento de Informática da UFES onde sinto-me muito feliz de estar trabalhando. Ao PPGEE/UFES por ter me aceito como aluno de doutorado. Muito obrigado.

Aos amigos aqui em Vitória: André, Célia, Elaine, Gustavo, Fabíola entre outros. Obrigado pela torcida.

A todos, inclusive os que esqueci de citar, dedico esta passagem:

> "Deus consola os humildes e dá força aos aflitos que lha pedem. Seu poder cobre a Terra e, por toda a parte, junto de cada lágrima colocou ele um bálsamo que consola. A abnegação e o devotamento são uma prece continua e encerram um ensinamento profundo. A sabedoria humana reside nessas duas palavras. Possam todos os Espíritos sofredores compreender essa verdade, em vez de clamarem contra suas dores, contra os sofrimentos morais que neste mundo vos cabem em partilha. Tomai, pois, por divisa estas duas palavras: devotamento e abnegação, e sereis fortes, porque elas resumem todos os deveres que a caridade e a humildade vos impõe. O sentimento do dever cumprido vos dará repouso ao espírito e resignação. O coração bate então melhor, a alma se asserena e o corpo se forra aos desfalecimentos, por isso que o corpo tanto menos forte se sente, quanto mais profundamente golpeado é o espírito." O Espírito de Verdade. (Havre, 1863.) Retirado de: O Evangelho Segundo o Espiritismo, cap. VI, item 8.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Siglas

SND	Sintagma Nominal Definido
AND	Anáfora Nominal Definida
DRT	do inglês: Discourse Representation Theory
SN	Sintagma Nominal
DRS	do inglês: Discourse Representation Structure
ADS	Árvore de Derivação Sintática
END	Estrutura Nominal do Discurso
LR	Lista de entidades Relevantes
LEN	Lista de Entidades não Resolvidas
LER	Lista de Entidades Resolvidas
PI	Ponto de Interpretação na END
CWA	do inglês: Closed World Assumption

Resumo

Esta tese apresenta uma solução para a interpretação de anáforas nominais definidas. Considere o seguinte texto:

(1) a. Mariana comprou *um carro novo.*

b. **O motor** veio danificado.

A frase (1a) apresenta duas entidades: Mariana e um carro novo. Já a frase (1.2b) tem apenas uma entidade – o motor. No processo de interpretação, humano ou computacional, a utilização do artigo definido "o" é um indicativo de que a entidade já havia sido introduzida no discurso, i.e. apresenta um caráter anafórico. Resolver uma anáfora é, *a priori,* identificar a quem ou a que se refere esta anáfora. Mas no caso acima é mais do que isto: sem dúvida o motor existe no texto por causa da existência de um carro, porém a interpretação do motor deve ir além disto e identificar como este motor está ligado com aquele carro. Isto é uma anáfora nominal definida.

A interpretação das anáforas nominais definidas ou de qualquer fenômeno anafórico pode ser generalizada como um processo que atribui valores aos itens da seguinte equação:

$$\mathcal{R}(\mathcal{A}, \mathcal{T}) \tag{2}$$

onde: $\mathcal{A}$ denota a entidade introduzida pela interpretação fora de contexto de um pronome, de uma elipse ou de um sintagma nominal definido, $\mathcal{T}$ denota o seu antecedente e $\mathcal{R}$ é a relação existente entre $\mathcal{A}$ e $\mathcal{T}$. O processo de resolução da equação, que é propriamente o processo de resolução de anáforas, consiste em descobrir $\mathcal{T}$ e $\mathcal{R}$ dado $\mathcal{A}$.

Nesta tese é proposta uma metodologia computacional que interpreta as anáforas nominais definidas cuja relação $\mathcal{R}$ é uma dentre: *parte de, membro de, subcategorizado por* e *coreferência.* A obtenção das relações é feita por um conjunto de regras pragmáticas [Freitas, Lopes e Menezes 2004, Filho e Freitas 2003] (cap. 3). Caso seja constatado que $\mathcal{A}$ não seja anafórica então ela é acomodada no contexto.

A metodologia computacional é construída sobre um ambiente de programação em lógica [Damásio, Nejdl e Pereira 1994] que permite raciocinar abdutivamente [Kakas, Kowalski e Toni 1992] sobre a representação semântica do texto [Kamp e Reyle 1993]. A partir da interpretação das entidades é construída a **estrutura nominal do discurso** [Lopes e Freitas 1994] (cap. 4), a qual permite: (1) fazer o acompanhamento das entidades mais salientes em cada frase [Freitas e Lopes 1994], (2) limitar o universo de escolha de possíveis antecedentes[Freitas e Lopes 1996] e (3) prover um resumo das entidades do discurso.

O resultado é uma metodologia que permite, de forma integrada, resolver anáforas e elipses, sendo que a estrutura nominal do discurso pode ser usada na busca de informações.

Abstract

This thesis presents a solution to the interpretation of definite descriptions in Portuguese. For example, consider the following text:

(1) a. Mariana bought a new car.

b. The engine was damaged.

The sentence (1a) introduces two entities: Mariana and a car which is new. The sentence (1b) introduces only one entity – the engine. In a human or computer interpretation process, the use of the definite article "the" preceeding a noun indicates that the introduced entity was already present at the discourse, i.e., it is an anaphoric entity. The resolution of an anaphora is a reference problem, but in the example (1) there is another problem: although the car is the entity that gives context to the engine, we can not say that the engine is the car (as for a pronominal anaphora). It also must be determined *how* the engine is related to the car. This is a definite description problem.

The interpretation of any kind of anaphora can be represented by the following equation:

$$\mathcal{R}(\mathcal{A}, \mathcal{T}) \tag{2}$$

where $\mathcal{A}$ denotes an entity introduced by the context interpretation of a pronoun, an ellipsis or a definite noun phrase, $\mathcal{T}$ denotes its antecedent and $\mathcal{R}$ is the relation between $\mathcal{A}$ and $\mathcal{T}$. The equation's resolution process is summarized as: given $\mathcal{A}$ find $\mathcal{T}$ and $\mathcal{R}$.

This thesis proposes a methodology to the definite description interpretation that the relation $\mathcal{R}$ is of: *part of, member of, subcategorized by* and *corefers.* These relations are obtained by a set of pragmatic rules [Freitas, Lopes e Menezes 2004, Filho e Freitas 2003], which are here defined (chapter 3). Also if $\mathcal{A}$ is not anaphoric then it is acommodated in the discourse context.

The computational methodology is implemented in a logic programming system [Damásio, Nejdl e Pereira 1994] that permits an abductive reasoning [Kakas, Kowalski e Toni 1992] at the semantic representation of the discourse [Kamp e Reyle 1993]. The interpretation of the entities is the basis to the Discourse Nominal Structure [Lopes e Freitas 1994] (chapter 4), which allows: (1) to track the most salient entities at each sentence [Freitas e Lopes 1994], (2) to limit the number of possible antecedents [Freitas e Lopes 1996] and (3) to give a discourse entities summary.

The result is an integrated metodology to solve anaphors and ellipses. Finally, the Nominal Structure of the Discourse can help the search/index of digital documents.

1 *Introdução*

"*Não há fé inabalável senão aquela que pode encarar a razão face a face, em todas as épocas da Humanidade.*"
Allan Kardec

Este capítulo apresenta o tema e o objetivo desta tese: a interpretação automatizada de anáforas nominais definidas, uma descrição das contribuições inovadoras deste trabalho e, por fim, a organização da tese.

1.1 Tema e objetivo da tese

Considere o seguinte texto:

(1.1) a. *A Mariana* ganhou *um livro* do *Lucas.*

b. **Ela** gostou do **presente**.

O processo de interpretação da frase (1.1a) introduz quatro entidades[1]: dois indivíduos introduzidos pelos nomes próprios *Mariana e Lucas*[2], um objeto introduzido pelo substantivo indefinido *um livro* e um evento introduzido pelo verbo *ganhou.* A interpretação da frase seguinte (1.1b) apresenta um estado introduzido pelo verbo *gostou* e duas outras entidades cujas referências não podem ser determinadas senão pela interpretação contextual: o objeto, indivíduo ou evento referenciado pelo pronome ***ela*** e o objeto, indivíduo ou evento referenciado pelo Sintagma Nominal Definido (**SND**) **o presente**. Os SNDs são aqueles precedidos de um artigo definido: o, a, os e as.

Ela e **o presente** são evidências de dois fenômenos do discurso conhecidos na literatura por, respectivamente, **anáforas pronominais** e **anáforas nominais definidas (AND)**. Em ambos os casos, o processo de interpretação requer a identificação da *entidade previamente introduzida* que está sendo referenciada: **o antecedente** (e.g. *Mariana*, *Livro* e *Lucas*). Cabe salientar que pode haver mais de um antecedente para uma mesma entidade. Por fim, o processo de interpretação é conhecido por **resolução anafórica** e o material sintático que ativa este processo é denominado **expressão anafórica** (e.g. pronome ***ela*** e SND ***o presente).***

O processo de identificação do antecedente pode utilizar uma ou mais das seguintes informações sobre a expressão anafórica e os seus possíveis antecedentes: (1) informação morfológica: gênero, número, pessoa e morfema, (2) informação sintática: sujeito, objeto e objeto direto, (3) informação temática: agente, ator, paciente, tema e localização, (4) informação semântica: tipagem de argumentos dos verbos e (5) informação pragmática: foco de atenção, animacidade das entidades, estrutura de língua e conhecimento de senso comum.

[1] No transcurso desta tese os termos: **objeto**, **indivíduo** e **entidade** serão usados intercaladamente, mas com preferência para o último que é mais genérico. Em relação a outros tipos de entidades, tais como: eventos, estados e tempos de referência, quando necessários serão mencionados de forma explícita.

[2] A convenção desta tese é que, nos exemplos, as partes em negrito assinalam as anáforas e as partes em itálico assinalam seus possíveis antecedentes. Também, quando necessário, são utilizados índices subscritos.

No caso específico da frase (1.1b) a anáfora pronominal sinalizada pelo pronome **ela** (singular e feminino) tem como antecedente, na frase anterior, a entidade denotada por *Mariana* (singular e feminino), pois este é o único par que respeita a igualdade das informações morfológicas. Já o SND ***o presente,*** usando apenas a informação morfológica, tem três antedentes possíveis: o indivíduo *Lucas,* o objeto *livro* e o evento em que o Mariana ganha o livro do Lucas. Esta última opção não será mais considerada no decorrer deste trabalho, porque necessita de um tratamento temporal [Rodrigues e Lopes 1995, Rodrigues e Lopes 1994] o qual está fora do escopo desta tese. Em relação às outras duas opções, mais informação é necessária para indicar qual antecedente deve ser preferido[3]. Utilizando-se a informação sintática é possível estabelecer uma ordem de saliência para as entidades: $sujeito > objeto > objeto2$. O resultado é que a entidade *livro* está melhor classificada do que a entidade *Lucas*, sendo então escolhido como antecedente da entidade **presente**.

Assim pode-se estabelecer, simplificadamente, que resolver (ou interpretar) uma anáfora pronominal ou uma anáfora nominal definida é identificar o seu antecedente, e.g., $Ela = Mariana$ e $presente = livro$. Porém veja o exemplo seguinte:

(1.2) a. **Mariana** comprou *um carro novo.*

b. **O motor** veio danificado.

O processo de interpretação da frase (1.2a) introduz duas entidades[4]: Mariana e carro novo. Já a entidade introduzida pela interpretação do **motor** na frase (1.2b) apresenta um comportamento duplo: (1) é um objeto novo para o discurso (e.g. *um motor*) e (2) apesar de fazer referência a outra entidade, apresentando um caráter anafórico [Donnellan 1966], não pode ser identificada diretamente com seu antecedente (e.g. $motor = carro$).

O processo de resolução das ANDs tem a fase de identificação do antecedente em comum com o processo de resolução das anáforas pronominais. Entretanto, introduz uma fase suplementar onde é necessário encontrar uma relação entre a entidade introduzida pelo SND e o seu antecedente. No caso específico do texto (1.2) é plausível assumir que, no contexto dado pela frase (1.2a), a entidade *o motor* se liga ao seu antecedente *carro novo* através de uma relação estrutural *parte de.*

[3] A preferência é apenas a escolha de alguma entidade dentro de uma lista de possíveis antecedentes. Escolhido um elemento a lista não é descartada pois: (1) pode ter sido uma escolha incorreta e outro elemento pode vir a ser escolhido ou (2) pode ocorrer que realmente exista mais de uma solução possível.

[4] Relembrando que os eventos, estados etc, já não estão mais sendo considerados.

Deste modo o processo de interpretação, quer das anáforas pronominais e das elipses quer das ANDs, pode ser generalizado como um processo que atribui valores aos itens da seguinte equação:

$$\mathcal{R}(\mathcal{A}, \mathcal{T}) \tag{1.3}$$

onde: $\mathcal{A}$ denota a entidade introduzida pela interpretação fora de contexto de um pronome, de uma elipse ou de um SND, $\mathcal{T}$ denota o seu antecedente e $\mathcal{R}$ é a relação existente entre $\mathcal{A}$ e $\mathcal{T}$. Por fim, o processo de resolução da equação, que é propriamente o processo de resolução de anáforas, consiste em descobrir $\mathcal{T}$ e $\mathcal{R}$ dado $\mathcal{A}$.

Nesta tese é proposta uma metodologia computacional que interpreta anáforas nominais definidas cuja relação $\mathcal{R}$ é do tipo estrutural: ***parte de***, ***membro de*** *e* ***subcategorizado por***. Além destas relações, são também tratadas a relação de ***co-referência*** e a *pseudo relação*[5] ***acomodação***.

A resolução das anáforas deve fazer parte de qualquer sistema automatizado de interpretação de textos. Com a resolução das anáforas, a representação obtida torna-se mais coesa pois são estabelecidas ligações entre as entidades introduzidas em cada frase. Esta representação permite um raciocínio melhor sobre as informações extraídas dos textos, em especial na busca de informações (apêndice A).

O protótipo computacional é construído sobre um ambiente de programação em lógica [Damásio, Nejdl e Pereira 1994] que permite raciocinar abdutivamente [Kakas, Kowalski e Toni 1992] sobre a representação semântica do texto [Kamp e Reyle 1993]. A partir da interpretação das entidades é construída a **estrutura nominal do discurso** [Lopes e Freitas 1994], a qual permite: (1) fazer o acompanhamento das entidades mais salientes em cada frase [Freitas e Lopes 1994], (2) limitar o universo de escolha de possíveis antecedente [Freitas e Lopes 1996] e (3) prover um resumo das entidades do discurso. No processo de interpretação das SNDs, foi criado um conjunto de regras pragmáticas que permitem identificar a relação entre a expressão anafórica e um possível antecedente [Freitas, Lopes e Menezes 2004, Filho e Freitas 2003] ou mesmo identificar que a entidade deve ser acomodada [Freitas e Lopes 1996].

Para avaliar a proposta, foram feitos dois testes: (1) utilizou-se o corpus marcado da Folha de São Paulo [Paulo 2002] para avaliar as regras de obtenção das relações $\mathcal{R}$ e

[5] Como será visto no capítulo 3, a acomodação acontece quando apesar da entidade ter sido introduzida por um SND, ela não é anafórica e comporta-se como uma entidade nova para o discurso.

foram feitos testes de desempenho na avaliação da Estrutura Nominal do Discurso para a resolução das anáforas (localização do antecedente $\mathcal{T}$ e identificação da relação $\mathcal{R}$).

1.2 Principais contribuições desta tese

Esta tese propõe algumas contribuições com aplicação no processo de interpretação de textos. A seguir as principais contribuições são destacadas.

Criação da estrutura nominal do discurso A principal inovação proposta é a construção de uma estrutura específica para o tratamento das anáforas nominais definidas, anáforas pronominais e elipses. Esta estrutura, denominada *Estrutura Nominal do Discurso*, é construída fazendo o acompanhamento dos focos do discurso. Tem múltiplas funções, das quais destacam-se: (1) limita o espaço de procura de antecedentes e de relações para uma anáfora ou elipse, (2) promove o acompanhamento explícito das entidades do discurso, (3) é um histórico que permite ao processo de resolução rever, a qualquer momento, as interpretações anteriores e (4) tem um conjunto de regras de (re)construção que impõe restrições ao processo de interpretação. Como será visto, esta estrutura específica é muito potente no processo de resolução e tem claras vantagens em relação a outras estruturas [Polanyi, Berg e Ahn 2003, Seville e Ramsay 1999, Mann e Thompson 1987, Grosz e Sidner 1986].

Visão integrada do processo de resolução de anáforas pronominais e anáforas nominais definidas

Fundamentada na forma genérica de resolução de anáforas (equação 1.3), a proposta desta tese contempla a integração dos processos de resolução de anáforas pronominais e nominais definidas num só ambiente computacional. Outras propostas para a resolução de anáforas: Teoria do Foco [Sidner 1981], Teoria da Centragem [Grosz, Joshi e Weinstein 1995] e Carter [Carter 1987], implementam primordialmente processos de resolução de anáforas pronominais.

Critérios para determinação das relações funcionais Apesar de não serem novas as relações utilizadas no processo de resolução proposto (relações estruturais, relações de co-referência e acomodação), são novos os critérios para a sua determinação: informação semântica sobre as entidades mais salientes e informação pragmática sobre os tipos de entidades envolvidas na resolução (expressão anafórica e antecedente). Cabe ainda

ressaltar que, na metodologia proposta nesta tese, estes critérios são suficientes para determinar dinamicamente as relações durante o processo interpretação, o que permite uma maior versatilidade do sistema em domínios desconhecidos. Esta proposta difere de outras [Hahn, Strube e Markert 1996, Hahn e Strube 1996] onde, apesar do número expressivo, as relações não podem ser determinadas dinamicamente durante o processo de interpretação, impedindo a fácil adaptação a novos domínios. Por fim, apesar do número pequeno de relações utilizadas (quatro), elas são suficientes para dar uma ampla cobertura na interpretação de textos ([Allen 1995]).

Nova visão sobre o foco do discurso A noção de que o foco é a entidade mais saliente em determinado ponto do discurso também não é nova. A inovação é a divisão do foco em dois: um foco para as entidades referenciadas explicitamente no discurso através das anáforas pronominais e elipses – foco explícito –, e um foco para as entidades referenciadas implicitamente pelas anáforas nominais definidas – foco implícito. Estes dois focos, essenciais para o acompanhamento das entidades do discurso através da estrutura nominal, contribuem decisivamente para possibilitar uma maior cobertura dos fenômenos tratados por outras definições de foco [Strohner et al. 2000, Grosz, Joshi e Weinstein 1995, Gundel 1994, Sidner 1981].

1.3 Organização da tese

No capítulo 2 é desenvolvida a problemática imposta pela interpretação das anáforas e são apresentadas propostas de resolução das anáforas. Na seqüência é delimitado o subconjunto dos fenômenos que são tratados nesta tese.

No capítulo 3 é definida a representação semântica do discurso, a qual é baseada na Teoria de Representação do Discurso (**DRT**) [Kamp e Reyle 1993]. Em seguida são definidas as relações estruturais tratadas e como elas podem ser interpretadas a partir de informação léxica, dando origem a um conjunto de regras pragmáticas. Por fim, é apresentada a implementação dessas regras.

No capítulo 4 é apresentada a estrutura nominal, com a qual é possível hierarquizar as entidades veiculadas pelo texto (entidades salientes ou não). Esta estrutura é fundamental no processo de resolução de anáforas nominais definidas (além de ser usada para a resolução de anáforas pronominais). Ela limita o espaço de procura de possíveis antecedentes.

No capítulo 5 é apresentado o protótipo para a interpretação das anáforas. Aqui são integradas num único ambiente: a estrutura desenvolvida no capítulo 4 e a regras pragmáticas para a obtenção das relações apresentadas no capítulo 3, sendo apresentada a forma como os dois interagem entre si.

No capítulo 6 são apresentadas as considerações finais sobre o trabalho realizado e algumas propostas de trabalhos futuros.

Finalmente, antevendo a aplicação da resolução de anáforas e da estrutura nominal do discurso na busca de informação em documentos digitais, é apresentado no apêndice A a arquitetura de uma máquina de busca que utiliza as propostas desta tese.

2 *Interpretação de anáforas*

"*Navegar é preciso, viver não*"
Luís de Camões

Neste capítulo são apresentadas a problemática imposta pela interpretação das anáforas, em especial é introduzida a equação $\mathcal{R}(\mathcal{A}, \mathcal{T})$, e algumas propostas da literatura para a resolução de anáforas nominais e pronominais.

2.1 Introdução

Anáfora é um fenômeno lingüístico no qual uma informação anteriormente introduzida é referenciada posteriormente em outra frase[1], através do uso de uma expressão lingüística mais simples, tal como no texto:

(2.1) O *João* ama a *Maria*, mas **ela** não **o** ama.

onde o sintagma nominal *o João* da primeira frase é referenciado na segunda frase através do pronome *o* (mesmo número, gênero e pessoa). O mesmo acontece com o sintagma *a Maria* e o pronome pessoal *ela* (mesmo número, gênero e pessoa). Formalmente, a expressão lingüística que introduz a anáfora na frase é denominada **expressão anafórica**. A informação previamente introduzida, a que deve ser ligada à expressão anafórica, é denominada **antecedente** e o processo pelo qual é identificado o antecedente duma expressão anafórica é denominado **resolução anafórica** ou **resolução de anáforas**. Tanto a expressão anafórica quanto o antecedente são representados como referentes do discurso [Kamp e Reyle 1993].

Ainda no exemplo (2.1), só existe um antecedente possível para cada expressão anafórica. Considere agora o texto:

(2.2) O João comprou um cão. Ele ladra muito à noite.

Neste exemplo, considerando apenas o número, gênero e pessoa, existem dois antecedentes possíveis para a resolução do pronome *ele*: o sintagma nominal *o João*, que ocupa a posição de sujeito da frase, e o sintagma nominal indefinido *um cão*, que ocupa a posição de objeto da frase. Diante desta ambigüidade, faz-se necessário encontrar outras informações que permitam escolher entre os dois antecedentes possíveis. Entre as possíveis fontes de informações, destacam-se:

1. o conhecimento de senso comum induzindo um interlocutor a dizer que os homens em condições normais não ladram enquanto os cães o fazem.

2. o conhecimento *inconsciente* do transmissor de que a emissão de um texto descrevendo situações sobre um mesmo objeto, i.e, mantendo o mesmo *centro de atenção,* facilita a interpretação por parte do receptor [Beaver 2004,

[1] Este trabalho considera apenas as anáforas interfrases (expressão anafórica e antecedente que estão em frases distintas), desconsiderando o tratamento das anáforas intrafrases [Reinhart 1981, Langacker 1966].

> Kruijff-Korbayová e Steedman 2003]. Os centros de atenção são repetidos, preferencialmente, sob a forma de anáforas, sendo que o objeto sobre o qual centra-se o texto tende a estar na posição do sujeito a cada frase.

Com estas duas fontes de informação pode-se então resolver a anáfora introduzida pelo pronome *ele* na segunda frase de (2.2): quem ladra a noite é o cão. A pergunta que fica é se, em qualquer contexto para resolução de anáforas, é indispensável a coexistência destes dois tipos de informações.

Nesta tese a resposta é afirmativa, porém com diferentes graus de influência sobre a escolha do antecedente: em contextos onde não se conhece muito sobre o domínio no qual o texto discursa, há uma tendência para que a informação estrutural [Gundel, Hegarty e Borthen 2003] seja determinante na escolha dos antecedentes, enquanto em contextos onde existe um conhecimento razoável sobre o domínio, há uma tendência a considerar o conhecimento de senso comum como determinante. Em ambos os contextos, uma informação não anula a outra na escolha do antecedente [Beaver 2004].

A resolução das anáforas nominais definidas é um caso particular da resolução de anáforas onde não basta identificar o antecedente, é necessário também identificar a relação que liga este à expressão anafórica. Veja o exemplo:

(2.3) O João comprou um carro. O motor veio avariado.

No texto (2.3) o sintagma nominal definido *o motor* – identificado pelo uso do artigo definido – e portanto introduzindo uma (possível) anáfora, deverá ser resolvido não apenas pela identificação de um antecedente (O João ou o carro), mas também pela identificação da relação existente entre o antecedente e a expressão anafórica, no caso *o motor é parte do carro*. Isto pode ser resumido na seguinte equação [Fraurud 1990]:

$$\mathcal{R}(\mathcal{A}, \mathcal{T}) \tag{2.4}$$

onde: $\mathcal{A}$ denota uma entidade introduzida pela expressão anafórica via pronome, elipse ou sintagma nominal definido, $\mathcal{T}$ denota o seu antecedente e $\mathcal{R}$ denota a relação existente entre $\mathcal{A}$ e $\mathcal{T}$. Assim, resolver uma anáfora pronominal é identificar pelo menos um antecedente $\mathcal{T}$ para uma anáfora $\mathcal{A}$, sendo a única relação possível a de *co-referência* [Sidner 1981, Sidner 1979] onde o referente do discurso introduzido pela interpretação de $\mathcal{A}$ deve co-referenciar o referente do discurso introduzido por $\mathcal{T}$. Já nas anáforas nominais definidas é necessário ainda identificar a relação $\mathcal{R}$ existente em $\mathcal{A}$ e $\mathcal{T}$ sem a qual não é

possível obter uma interpretação plausível para a existência das entidades no contexto de interpretação (e.g. o motor e o carro) [Beaver 2004, Scliar-Cabral 2002].

A resolução de anáforas nominais definidas é o tema central desta tese. Para melhor examiná-las, a seguir é feita uma revisão de propostas para resolução de anáforas, destacando-se a Teoria do Foco [Sidner 1981, Sidner 1979], a Teoria da Centragem [Grosz, Joshi e Weinstein 1995, Grosz, Joshi e Weinstein 1983] e a Teoria de Representação do Discurso - DRT [Kamp e Reyle 1993].

Este capítulo está assim estruturado: na seção 2.2 são apresentados as classificações de uso dos Sintagmas Nominais Definidos existentes na literatura, na seção 2.3 são apresentadas as propostas para resolução de anáforas com maior destaque na literatura e, finalmente, na seção 2.4 é feita uma avaliação destas propostas.

2.2 Usos de um Sintagma Nominal Definido

Nesta seção são apresentados os trabalhos que classificam os usos de um sintagma nominal definido. Serão vistos os trabalhos de Hawkins [Hawkins 1978], Russel [Russell 1919] e Prince [Prince 1992, Prince 1981].

2.2.1 A definição de Hawkins

Hawkins [Hawkins 1978] descreve e estende os casos de utilização dos SNDs propostos por Christopherson (cf. [Hawkins 1978]), propondo que os SNDs podem ser definidos com base na necessidade da existência de um antecedente do discurso (uso anafórico ou associativo) ou não (uso contextual, usos não relacionais com modificadores explicativos e usos não explicativos). A seguir são apresentados os tipos de utilização dos SNDs.

2.2.1.1 Uso Anafórico

Quando um SND referencia uma entidade previamente introduzida no discurso, ele passa a ter um caráter anafórico e então denominado **Anáforas Nominais Definida** (AND). Hawkins considera que ambos, o SND e seu antecedente, evocam a mesma entidade no mundo real. Exemplos são:

(2.5) a. Ayrton comprou *um carro novo*. Fabiola não gostou *do carro*.

b. André só veste *camisetas largas*. *A roupa* não lhe cai bem.

c. Célia trabalhou toda a manhã *na tese. A monografia* estará pronta até dezembro.

Note que no exemplo (2.5a) a entidade "carro novo" vai ser posteriormente referenciada por uma entidade introduzida cuja cabeça lingüística é a mesma (carro). Tipicamente isto é conhecido por sinonímia. Já no exemplo (2.5b) a entidade "camisetas largas" vai servir de antecedente para o SND "a roupa" numa relação de hiponímia (o antecedente é mais específico que o SND). Por fim no exemplo (2.5a) "a tese" pode ser expressa de diversas formas e normalizada sob a forma de *monografia* (relação de sinonímia).

2.2.1.2 Uso anafórico associativo

Diferente do uso anafórico puro, onde a relação entre o SND e seu antecedente é direta ou, quando muito, ocorre uma generalização, no uso anafórico associativo a utilização de um SND parece indicar que existe conhecimento mútuo entre transmissor e receptor sobre as entidades envolvidas. Porém a relação entre estas não é mais uma entidade em comum, mas sim de uma entidade (antecedente) que dá suporte a existência de um SND [Bos, Buitelaar e Mineur 1995]. Veja os exemplos a seguir:

(2.6) a. Gustavo comprou *um carro usado. O motor* estava quase fundindo.

b. O André comprou *um carro novo.* Mas a Célia não gostou *da cor.*

c. O Ayrton comprou *um livro sobre carros. O autor* era conhecido.

Note que nos três exemplos (2.6) a entidade introduzida pelo SND não pode estar ligada diretamente ao seu antecedente, porém é a existência do antecedente que dá contexto à existência da entidade introduzida pela utilização de um SND.

2.2.1.3 Uso contextual imediato

Neste tipo de utilização a entidade referenciada ou está presente (e visível) durante a sua utilização no contexto do discurso [Freitas 1993] ou pode ser inferida do mesmo. Alguns exemplos são:

(2.7) a. Por favor, passe-me *o sal.*

b. Dê-me *o lápis.*

Note que em ambos os casos do texto (2.7) os interlocutores estão num contexto onde os objetos citados, através da utilização de SNDs, estão visualmente presentes. Não existe um antecedente no discurso. Já nas frases a seguir:

(2.8) a. (placa) Cuidado com *o cão*.

b. Não alimente *os animais*.

Ambas as entidades introduzidas por um SND podem ser inferidas a partir do contexto onde elas aparecem: se o interlocutor está num portão e lê a frase (2.8a) ele inferirá a existência de um cão sem que o veja. O mesmo acontece quando alguém lê a frase (2.8b) na entrada de um zoológico. Mesmo sem ver os animais, é de se supor que eles existem ali!

2.2.1.4 Uso contextual abrangente

Contrastando com o uso contextual imediato onde o transmissor espera que o receptor esteja presente no contexto de emissão do discurso, neste caso o emissor espera que o receptor compartilhe conhecimento sobre a entidade introduzida pelo SND e a situação que dá contexto à sua existência[2]. Por exemplo numa situação onde o tema é a novela "Escrava Isaura":

(2.9) *A escrava* fugiu para se casar.

Note que o SND "a escrava" está inserido na situação descrita pela novela e basta que ambos os interlocutores tenham informação sobre este contexto para que o SND possa ser facilmente utilizado.

2.2.1.5 Usos não relacionais com modificadores explicativos

Hawkins classifica como não relacionais aqueles SNDs que não são anafóricos, não recaem sobre informação da situação do discurso e não são associados a nenhum antecedente do discurso anterior. Por fim, ele agrupa este SNDs em classes de acordo com propriedades léxicas e sintáticas:

[2]Note que isto difere do uso contextual onde o contexto de existência de uma entidade é dado por outra entidade e não por uma situação.

Complementos do Sintagma Nominal Esta forma de SND não relacional é caracterizada pela presença de um complemento para o substantivo:

(2.10) a. Carlos está maravilhado com *a descoberta de que existe vida em Marte.*

b. Flávio falou sobre *a vida que todos nós podemos ter.*

c. Luana só pensa *na vida que se segue ao casamento.*

Em todas as frases do exemplo (2.10) a utilização do SND não está condicionada à existência de uma relação (direta ou não) com outra entidade (contextual ou do discurso). Aqui o uso do SND é não anafórico devido à quantidade de informação associada sintaticamente a este, podendo o mesmo ser interpretado como um Sintagma Nominal Indefinido[3] e, portanto, introduzir uma nova entidade no discurso [Kamp e Reyle 1993, Heim 1982].

Modificadores Nominais De acordo com Hawkins, a presença de modificadores nominais é o que distingue as seguintes frases:

(2.11) a. Eu não gosto *da cor rosa.*

b. *O número três* é meu numero de sorte.

Frases relativas As frases relativas podem ser consideradas como autocontidas, de forma que o receptor vai introduzir um referente sem que haja a necessidade de uma menção anterior.

2.2.2 A Teoria de Russell

Na análise de Russell [Russell 1919] os SNDs, por ele chamados de descrições definidas, não pertencem à classe de termos de referenciamento tais como os nomes próprios, mas sim à classe de frases denotativas tais como os quantificadores. De forma que a proposição de uma frase com a seguinte forma:

O F é um G

é representado por uma sentença quantificadora consistindo de:

1. uma condição existencial (existe pelo menos um F),

[3]Sintagmas iniciados por um artigo indefinido: um, uma, uns, umas, algum, alguma, etc.

2. uma condição de unicidade (existe no máximo um F), e

3. uma proposição (tudo que é F é G).

O resultado é expresso formalmente em:

$(\exists x)(Fx\&(\forall y)(Fy \rightarrow y = x)\&Gx)$

Alguns trabalhos mais atuais na área de semântica de linguagem natural ainda utilizam a análise de Russell para o tratamento dos SNDs, um exemplo é a semântica de Montague (cf. [Dowty, Wall e Peters 1981]). A análise de Russell é adequada para o tratamento de conceitos funcionais (SNDs tais como o "cão do Jorge" ou "a praça da cidade"), mas a condição de unicidade é muito forte para descrições em linguagem natural, tal como em:

(2.12) (na frente de uma estante na seção de biologia de uma biblioteca) Esqueci o nome do autor, porém peguei *o livro de botânica*.

A análise de Russell vem sendo revisada no que se refere ao tratamento da condição de unicidade relativa à situação [Cooper 1993].

2.2.3 A proposta de Prince

Prince estudou em detalhes a ligação entre as suposições que o transmissor e o receptor fazem entre si e a maneira como isto é expresso sob a forma de sintagmas nominais [Prince 1992, Prince 1981]. No trabalho de Prince é feita a distinção entre dois tipos de familiaridades, o que não é feito no trabalho de Hawkins.

Em sua crítica Prince avalia como muito simplista a distinção tradicional das entidade do discurso em dois tipos: preexistente ('given') e nova ('new'). Ela propõe uma autonomia muito mais rica para a preexistência de uma entidade ou, como ela chama, familiaridade assumida ('assumed familiarity'). Ela faz a distinção entre a familiaridade do discurso e familiaridade ao receptor.

2.2.3.1 Entidade nova/velha para o receptor

Um fator que afeta a escolha de um **SN** é se a entidade do discurso é velha ou nova em relação ao conhecimento do receptor. Tipicamente, um transmissor vai usar um

nome próprio ou um SND quando ele assume que seu interlocutor já conhece a entidade mencionada pelo transmissor, como em:

(2.13) Estou esperando a hora certa para falar com a Lucia Catabriga.

Por outro lado, se o transmissor acredita que seu interlocutor não conhece quem é a Lucia (neste contexto), então um indefinido vai ser usado:

(2.14) Estou esperando a hora certa para falar um professor do DI.

As entidades do discurso podem ainda ser novas ou velhas com relação ao modelo do discurso.

2.2.3.2 Entidade nova ou velha no discurso

De acordo com Prince, um SN pode referenciar uma entidade que já foi anteriormente referenciada no discurso corrente (textualmente referenciada), ou pode referenciar uma entidade que não foi anteriormente mencionada. Assim uma entidade nova no discurso é distinta de uma entidade nova para o receptor.

2.2.3.3 Familiaridade assumida

Esta é dividida em algumas categorias, a saber:

Entidade nova Um SN pode introduzir uma entidade nova tanto para o discurso quanto para o receptor. Entidade novas são, freqüentemente, introduzidas por SN indefinidos tais como "um professor do DI" no exemplo (2.14).

Entidade nova ancorada Uma entidade é ancorada se ela está ligada a outra entidade do discurso, se esta ligação está contida no SN que expressa a entidade e se esta ligação não é, por si, nova. Prince considera apenas os SN indefinidos nesta classe, mas SN definidos tais como "O pessoal com quem trabalha ..." podem, a nosso ver, ser classificados como tal.

Entidade referenciada Os SNs podem referenciar entidades contextuais ou textuais. Apenas as entidades textuais referenciadas são velhas no discurso. Entidades referenciadas contextualmente correspondem ao uso contextual imediato descrito por Hawkins.

Entidades não usadas Os SNs podem referenciar entidades velhas para o receptor mas novas para o discurso. Os SNs não usados descrevem entidades que são conhecidas por ambos os interlocutores mas que não foram mencionadas/usadas previamente no discurso. Estes casos são semelhantes aos descritos por Hawkins no "uso contextual abrangente".

Entidades inferíveis Algumas entidades do discurso não são velhas no discurso ou velhas ao receptor, mas, por outro lado, elas não são inteiramente novas para ambos. Hawkins chama a tais casos de "uso anafórico associativo": um livro ... o autor. Prince denomina tais entidades de *inferíveis*, porém não introduz uma classe para aquelas entidades que podem ser inferidas da situação (o "uso contextual abrangente" de Hawkins).

Entidades com parte inferível Prince propõe uma categoria para as entidades semelhantes à classe anterior, mas cuja conexão com o conhecimento prévio do receptor é especificada por somente parte do SN, como em:

(2.15) *O portão da Bastilha* foi pintado de cinza.

2.3 Propostas para a Resolução de Anáforas

Precedendo a Teoria do Foco e a Teoria da Centragem, encontra-se o trabalho de Grosz [Grosz 1977] sobre o *foco global* (tópico geral do discurso) e o *foco local* (entidades mais salientes em cada frase). O foco global afeta a produção e interpretação de sintagmas nominais definidos enquanto o foco local afeta a produção e a interpretação de pronomes e elipses[4]. Grosz descreve um mecanismo de identificação e representação do foco, denominado *focalização*, para um sistema de diálogos voltado para a execução de tarefas. O sistema é construído sobre uma base de dados que contém informações sobre as tarefas a serem executadas. A base de dados é uma rede semântica cujos nós representam: objetos, eventos, relações e conjuntos. Os arcos representam relações binárias (estáticas no

[4]Caso em que parte do material sintático é omitido da frase, por exemplo:

(2.16) a. O Antônio chegou.
b. e já Φ saiu.

Neste caso há a elisão do sujeito na segunda frase (representado pelo símbolo Φ). A frase (2.17a) pode ser substituída por:

(2.17) b'. e *ele* já saiu.

Note que a elisão de material sintático foi substituída por um pronome. A interpretação da elipses é semelhante ao tratamento dado à anáforas pronominais.

tempo) entre nós. Nesta base de dados, o foco representa (ou antes marca) as entidades que estão mais salientes dentro de cada *espaço focal*[5], limitando o espaço de procura dos antecedentes de uma anáfora. Este trabalho teve considerável influência sobre outros trabalhos, nomeadamente: sobre a Teoria do Foco [Sidner 1981, Sidner 1979], sobre a Teoria da Centragem [Grosz, Joshi e Weinstein 1995, Grosz, Joshi e Weinstein 1983, inter alia] e sobre a proposta da Teoria da Estrutura do Discurso [Grosz e Sidner 1986].

A Teoria do Foco [Sidner 1979] e o trabalho de Grosz são pioneiros no tratamento computacional das anáforas. A Teoria do Foco propõe uma separação entre o processo de resolução de anáforas (identificação de um antecedente) e o processo de validação do antecedente escolhido. Esta separação reduz consideravelmente a carga computacional necessária para a interpretação de uma anáfora. O processo de validação do par anáfora/antecedente somente vai testar o antecedente proposto anteriormente e não vai fazer o teste sobre todo o universo de possíveis antecedentes.

Para conseguir esta separação, Sidner utilizou a noção de foco local, ou seja, a informação de quais entidades são mais salientes em determinada parte do discurso. Uma das propriedades conhecidas do foco é a sua associação às anáforas [Grosz, Joshi e Weinstein 1995, Grosz e Sidner 1986, Polanyi 1988, inter alia]. De uma maneira *recursiva:* a determinação do foco de uma frase depende das entidades anafóricas da mesma e a resolução das entidades anafóricas de uma frase depende do foco corrente. O foco marca o antecedente preferencial para a resolução de futuras anáforas e portanto não há, à partida, necessidade de inferências na determinação de um antecedente. É apenas necessário fazer a validação do mesmo (no caso, o foco).

Na mesma linha de utilização do foco, Grosz et al [Grosz, Joshi e Weinstein 1983], no âmbito da Teoria da Centragem, propõem um modelo para a descrição da coerência do discurso baseando-se no centro de atenção (foco) e na escolha das expressões de referenciamento (i.e. sintagmas nominais definidos e pronomes) por parte do emissor.

A Teoria da Centragem não foi, no entanto, construída visando a resolução de anáforas [Grosz, Joshi e Weinstein 1995], mas, devido à utilização do foco e à ligação intrínseca deste com as anáforas, era de se supor que a Teoria da Centragem viesse a ser utilizada na resolução das mesmas.

E assim aconteceu. Um dos trabalhos pioneiros na utilização da Teoria da Centragem como ferramenta para a resolução de anáforas pronomi-

[5] Os espaços focais contêm as entidades em foco num determinado fragmento do diálogo e são ordenados de acordo com uma hierarquia que reflete a estrutura do discurso em análise.

nais foi o de Brennan et al [Brennan, Friedman e Pollard 1987]. Outros trabalhos se seguiram [Kehler 1993, Kehler 1993, Walker, Lida e Cote 1994, Walker 1989, Kameyama, Passanneau e Poesio 1993, inter alia] abordando novos fenômenos mediante a utilização direta do algoritmo básico (seção 2.3.2) ou fazendo-lhe extensões.

Paralelamente às propostas anteriores, situam-se os trabalhos na lingüística formal e em lógica que tratam dos aspectos de representação da língua, normalmente baseados em Montague (cf. [Dowty, Wall e Peters 1981]), e que impõem restrições semânticas à interpretação das frases. Os trabalhos mais conhecidos nesta área são: A Teoria da Representação do Discurso - DRT [Kamp e Reyle 1993] que guarda uma estreita ligação com a "File Change Semantics" [Heim 1982] e a Semântica Dinâmica [Groenendijk e Stokhof 1991].

2.3.1 Teoria do Foco

Sidner [Sidner 1981, Sidner 1979] propõe um algoritmo de focalização cujas principais funções são: (1) reduzir o universo de possíveis antecedentes introduzidos no universo do receptor durante a interpretação de novas frases proferidas em contexto ou escritas num todo coerente e (2) propor um caminho mais eficiente para percorrer este universo já reduzido em busca de um antecedente. Para tal ela propõe que as entidades mais salientes de uma frase devem ser os antecedentes preferenciais para a resolução de uma anáfora numa frase seguinte. Propõe ainda que, caso as entidades mais salientes não possam servir de antecedente, a procura de antecedentes deve ser então feita de forma ordenada no conjunto das entidades introduzidas ou referenciadas na frase anterior.

Para resolver uma anáfora, para determinar o foco e para ordenar as entidades de uma frase, Sidner utiliza as seguintes informações: a informação temática (agente e tema) [Gruber 1976], a informação gramatical (sujeito, objeto direto, objeto indireto etc) e a informação sobre quais são as entidades mais salientes da frase anterior – o foco local [Grosz 1977].

São dois os tipos de informação temática utilizadas: o *agente* que é definido como sendo o sujeito de um verbo transitivo, caso este seja animado[6], e o *tema* que também será o sujeito a menos que este já seja o agente da frase e exista um objeto direto, caso em que o objeto direto será o tema.

Sidner define dois centros de atenção ou focos: o *foco do ator* (FA) e o *foco do discurso* (FD), que são preferencialmente determinados pelo agente e pelo tema de cada frase. A

[6] O critério de animacidade representa aqui a preferência do agente por entidades que possam atuar sobre outras. Não há aqui a consideração de uma hierarquia de animacidade [Dahl e Fraurud 1996].

utilização de anáforas é outro fator importante para a determinação dos focos.

A utilização de dois focos em cada frase foi questionada por Grosz et al [Grosz, Joshi e Weinstein 1995], argumentando que cada frase tem somente uma entidade mais saliente. Este ponto de vista encontra respaldo se o objetivo da utilização do foco é a medição da coerência entre duas frases, como é o caso da Teoria da Centragem [Grosz, Joshi e Weinstein 1995, Grosz, Joshi e Weinstein 1983]. Porém, como a Teoria do Foco visa a resolver anáforas pronominais e como pode acontecer que, numa mesma frase, exista mais de uma anáfora, a utilização de dois focos torna-se necessária. Cada anáfora é associada a um foco. Por outro lado, Carter [Carter 1987] argumenta que a existência de dois focos não é suficiente para resolver os casos em que existam mais de duas anáforas que não estejam na posição de agente na frase e sugere que a utilização das listas de focos potenciais deve ter prioridade nestes casos.

Como será visto no capítulo (4), esta tese considera dois focos diretamente relacionados, que não dependem da informação temática mas sim do grau de acessibilidade da entidade referenciada [Ariel 1996, Chafe 1996]. Anáforas pronominais e elipses referenciam entidades mais acessíveis e dão origem ao aqui designado por *foco explícito* ($foco$). Anáforas nominais definidas representam entidades menos acessíveis e dão origem ao *foco implícito* ($ifoco$).

2.3.2 Teoria da Centragem

Ao contrário da Teoria do Foco (apresentada na seção 2.3.1), a Teoria da Centragem [Grosz, Joshi e Weinstein 1983] não foi criada para resolver anáforas, mas sim para medir como a coerência do discurso é influenciada pela compatibilidade entre os centros de atenção[7] (foco) e a escolha das expressões de referenciamento [8] [Grosz, Joshi e Weinstein 1995].

A Teoria da Centragem baseia-se: (1) no trabalho de Grosz [Grosz 1977] sobre a existência de dois níveis de foco: global e local, (2) no trabalho de Sidner [Sidner 1979] sobre a Teoria do Foco (apresentado na seção 2.3.1) com a definição do foco corrente e da lista de focos potenciais e (3) no trabalho de Grosz, Joshi e Weinstein [Grosz, Joshi e Weinstein 1983] sobre as inferências necessárias para integrar a interpretação de uma frase no discurso previamente interpretado.

[7] Grosz et al utilizam o termo *centros de atenção* e não foco para evitar confusão com outras definições de foco, nomeadamente o foco da prosódia.

[8] Por expressões de referenciamento se deve entender os sintagmas nominais anafóricos, as expressões anafóricas.

É importante destacar que a Teoria da Centragem é uma proposta diretamente ligada ao trabalho conjunto de Grosz e Sidner [Grosz e Sidner 1986] sobre a Teoria da Estrutura do Discurso, dentro da qual a Teoria da Centragem é uma proposta para explicar a coerência dentro de um mesmo segmento (seqüência de frases obedecendo a um determinado critério de agrupamento), isto é, a *coerência local*.

Na Teoria da centragem cada frase de um segmento (constituído pela seqüência de frases $F_1, F_2, ...F_i, ..., F_n$) é representada por uma lista ordenada de todas as entidades da frase e pela entidade em foco em cada frase, definidas, respectivamente, por: lista ordenada de potenciais centros de atenção ($Cf(F_i)$)[9] da frase i e centro de atenção ($Cb(F_i)$)[10] da frase i.

O $Cb(F_i)$ da frase i faz co-referência a um elemento de $Cf(F_{i-1})$ da frase anterior *i-1*. $Cf(F_i)$ contém os elementos da frase i que potencialmente podem ter ligações com a próxima frase, $Cf(F_{i+1})$. A lista $Cf(F_i)$, que inclui também o $Cb(F_i)$, é ordenada pela posição de aparecimento na frase, de modo que a escolha dos próximos centros de atenção vai depender desta ordem. Finalmente o $Cp(F_i)$ é o elemento melhor classificado em $Cf(F_i)$ e que portanto será o preferido para ser escolhido como o centro de atenção da frase corrente. É de realçar que os centros de atenção são entidades semânticas no modelo do discurso e são usadas para interpretar sintagmas nominais singulares, designando objetos no mundo (pessoas, animais, objetos, etc).

A utilização dos centros de atenção do discurso na interpretação permite a obtenção de ligações entre as entidades das frases, estabelecendo critérios de coesão e também de coerência.

O processo de ordenamento, embora de vital importância para a Teoria da Centragem, não está claro em [Grosz, Joshi e Weinstein 1983]. Brennan et al [Brennan, Friedman e Pollard 1987] usam as relações gramaticais subcategorizadas pelo verbo principal: sujeito, objeto direto, objeto indireto, outras subcategorizações e por fim os adjuntos. Esta ordem geralmente coincide com a ordem de escrita da frase no Inglês e caso não coincida pode levar à escolha incorreta de um antecedente.

A teoria desenvolvida ao longo desta tese não adota esta ordenação, por duas razões: primeiro porque ela só pode ser mais eficiente nas línguas com ordem quase fixa (inglês, francês e português), não tendo resultados satisfatórios em línguas que não tenham ordem fixa (alemão e checo, por exemplo) [Hahn e Strube 1996, Strube e Hahn 1996]. E segundo

[9] forwards centers, Cf.
[10] backwards centers, Cb.

porque esta ordenação impõe restrições muito fortes no caso da resolução das anáforas nominais definidas, que buscam o seu antecedente nos elementos menos acessíveis da lista de potenciais centros de atenção. Por estas razões adotou-se uma ordenação funcional modificada a partir da proposta de Strube e Hahn [Strube e Hahn 1996]. O primeiro critério de ordenação é a função da entidade: entidades anafóricas são melhor classificadas que as não anafóricas. Esta ordenação e a sua influência na determinação do foco estão detalhadas no capítulo 4.

2.3.3 Abordagens semânticas

As abordagens semânticas têm como primeiro objetivo encontrar um ambiente de trabalho formal para a representação da linguagem articulada: criar mecanismos de interpretação das frases visando a uma representação semântica num formalismo lógico.

Destaca-se o trabalho de Montague [Dowty, Wall e Peters 1981] que se tornou o mais difundido e aceito, tendo influenciado diversos outros formalismos, nomeadamente a DRT [Kamp e Reyle 1993], a "File Change Semantics" [Heim 1982] e a "Dynamic Predicate Logic" [Groenendijk e Stokhof 1991].

2.3.3.1 Montague

Montague [Dowty, Wall e Peters 1981] introduziu a semântica baseada em modelos no estudo da língua natural. A idéia central é descrever a semântica da língua utilizando técnicas desenvolvidas na lógica matemática. Aqui a interpretação das frases é feita mediante a utilização de mundos possíveis, definindo as condições para as quais uma determinada interpretação seja verdadeira.

Outro fator importante em Montague é a idéia de composicionalidade: a representação do significado de uma frase é construída a partir do significado de suas partes (sintagmas) e da maneira como estas estão organizadas. O significado de cada constituinte é determinado por um conjunto de regras semânticas replicando um conjunto de regras sintáticas.

Montague considera que algumas expressões básicas da língua têm uma interpretação fixa num dado modelo (particularmente os determinantes: um, todo, todos etc). Por exemplo, o artigo indefinido *um* introduz o quantificador existencial $\exists$, enquanto que o determinante *todo* introduz o quantificador universal $\forall$.

Para Montague a noção de domínio sintático – que delimita o universo de procura do

antecedente para uma anáfora – é substituído pela idéia de escopo de quantificação. Os pronomes são vistos como variáveis livres que devem ser instanciadas dentro do escopo de quantificação ao qual pertencem. Porém os pronomes não são traduzidos diretamente em entidades individuais, mas sim em expressões que denotam o conjunto de propriedades do indivíduo.

Um dos problemas no tratamento de anáforas dentro da proposta de Montague é levantado pelas conhecidas "frases de burro" – do Inglês *donkey sentences*: frases compostas relativas ou condicionais que contêm um sintagma nominal indefinido na frase subordinada que é co-referido por um pronome inserido na frase principal. Alguns exemplos:

(2.18) a. Se um fazendeiro tem um burro, (ele) bate-lhe.

b. Todo homem que tem um burro bate-lhe.

Segundo Montague, existem duas interpretações possíveis para as frases em 2.18, porém nenhuma delas dá a leitura correta:

$$\forall x[[fazendeiro(x) \wedge \exists y[burro(y) \wedge ter(x,y)]] \rightarrow bater(x,y)] \tag{2.19}$$

$$\forall x \exists y[[homem(x) \wedge [burro(y) \wedge ter(x,y)]] \rightarrow bater(x,y)] \tag{2.20}$$

Em (2.19), representação lógica da frase (2.18a), o quantificador existencial assume um escopo restrito sobre a implicação, deixando a segunda ocorrência de y (que representa o pronome) fora do escopo de $\exists$. Nestes casos, o pronome tem um valor cujo referente é independente do SN *"o burro"*.

Por outro lado em (2.20), representação lógica da frase (2.18b), o escopo alargado do quantificador existencial sobre a implicação permite que a segunda ocorrência de y possa ser instanciada com a entidade *"o burro"*. O problema surge na verificação das condições de verdade: a interpretação da frase será verdadeira quando o antecedente da implicação for falso (i.e., a existência de um y que não é um burro). São necessárias condições suplementares para garantir que a interpretação esteja correta: que cada fazendeiro tem pelo menos um burro e que a proposição ainda é verdadeira se o fazendeiro tiver mais de um burro, mas apenas bater num deles.

2.3.3.2 DRT e as DRSs

A Teoria da Representação do Discurso - DRT [Kamp e Reyle 1993] foi proposta no âmbito da lingüística formal e da Filosofia da Linguagem e tem as suas raízes no trabalho de Montague [Dowty, Wall e Peters 1981], constituindo como tal um formalismo semântico baseado na teoria dos modelos[11], sem que contudo seja apenas uma extensão ao trabalho de Montague, especialmente porque visa à interpretação do discurso e não das frases isoladas.

O trabalho inicial de Kamp teve duas motivações iniciais: dar um tratamento adequado às anáforas pronominais singulares (em particular à representação das "frases de burro", veja o exemplo (2.18)) e fazer um estudo sobre o tempo e o aspecto, em especial sobre a referência temporal.

A DRT resolve o problema das variáveis não instanciadas das "frases de burro" através da criação de uma estrutura sem variáveis livres, as Estruturas de Representação do Discurso - **DRS**. Estas estruturas abstratas são a representação semântica das frases (e mesmo do discurso), e são obtidas como resultado do *Algoritmo de Construção*: um mecanismo recursivo que tem como entrada a estrutura sintática da frase e como saída a DRS correspondente.

As DRSs têm condições de verdade bem definidas, porém estas condições de verdade não são especificadas para as frases do discurso, mas sim para as DRSs construídas a partir delas. Em suma, as condições de verdade de uma frase são definidas via DRS global do discurso.

Nas DRSs, os indivíduos introduzidos por um sintagma nominal (na árvore de derivação sintática) são representados por entidades semânticas chamadas *Referentes do Discurso*. As DRSs contêm ainda um conjunto de *condições da DRS*, que representa a informação descritiva veiculada por uma frase: a identidade dos indivíduos representados pelos referentes e a proposição que representa a predicação da frase (freqüentemente veiculada pela interpretação do verbo principal). As condições de uma DRS podem ser simples ou complexas, isto é, são recursivas e podem conter outras DRSs.

Em resumo, as DRSs têm dois componentes: um conjunto de referentes do discurso, chamado de *Universo da DRS* e um conjunto de condições. Por exemplo:

[11] A teoria de modelos é o estudo da representação de conceitos em termos de teoria de conjuntos. É assumido que existem alguns objetos pré-existentes, e investiga-se o que pode ser concluído de tal coleção de objetos, algumas operações e/ou relações entre estes objetos, e alguns axiomas.

(2.21) Marcos foi ao médico.

é representado como:

x,y,e
Marcos(x) médico(y) e:ir(x,y)

(2.22)

onde x e y constituem o universo da DRS e *Marcos(x), médico(y)* e *ir(x,y)* as condições associadas à DRS[12].

As condições de verdade da DRS são verificadas se existir um indivíduo para cada referente do discurso (no universo da DRS) de forma que as condições contidas na DRS para tais referentes sejam satisfeitas por estes indivíduos.

A DRT trata as referências anafóricas como uma relação entre os referentes introduzidos pelos pronomes e os referentes introduzidos na DRS que representa todo o discurso anteriormente interpretado. Os referentes do discurso no universo da DRS principal e no universo das subDRSs representam os indivíduos aos quais os pronomes podem fazer referência. A DRT tem regras de construção para os pronomes que contemplam a resolução anafórica destes. Para estabelecer a ligação entre o referente introduzido pelo pronome e o referente do seu antecedente, este último tem que satisfazer restrições sintáticas, semânticas e pragmáticas. O resultado é a inserção de uma condição DRS que liga os dois referentes. Por exemplo:

(2.23)

a. João gosta da Maria.

b. (e) Ela gosta do Pedro.

Onde as regras de construção das DRSs aplicadas sobre a frase (2.23a) resultam na DRS (2.24):

[12]Esta tese não considera a interpretação do tempo e do aspecto verbal da frase que daria origem a dois referentes adicionais ([Rodrigues e Lopes 1995, Rodrigues e Lopes 1994, Rodrigues 1995, Rodrigues e Lopes 1992]): *e* e *t*, o primeiro associado à eventualidade descrita pelo verbo principal e o segundo associado ao intervalo de tempo em que a eventualidade decorreu, e a pelo menos mais duas condições: *occurs(e, t)* ou *holds(e, t)* descrevendo, no primeiro caso, que o evento *e* ocorreu no intervalo de tempo *t* e, no segundo caso (*holds(e, t)*), que o estado *e* foi observado no intervalo de tempo *t* e a condição *evt(e,ir(x,y))* descrevendo que o referente é uma instância da eventualidade *ir(x,y)* [Rodrigues e Lopes 1995, Rodrigues 1995].

x,y,e_1
joão(x) *maria(y)* e_1*:gostar(x,y)*

(2.24)

A qual constituirá o contexto para a interpretação da DRS (2.23):

z,w,e_2
z=? *pedro(w)* e_2*:gostar(z,w)*

(2.25)

Como resultado desta interpretação em contexto, o referente z, introduzido pela interpretação isolada do pronome na segunda frase, será resolvido com o referente x. O resultado será então uma nova DRS:

x,y,z,w,e_1,e_2
João(x) Maria(y) e_1 :gostar(x,y) z=x Pedro(w) e_2 :gostar(x,y)

(2.26)

Com relação ao tratamento dado às "frases de burro" (exemplo (2.18)) elas introduzem subDRSs. Não podendo os referentes do universo das subDRS ter acesso de forma arbitrária (existem regras de acessibilidade) ao universo da DRS que as contém, logo o universo de escolha do antecedente de uma anáfora é limitado por esta restrição de acessibilidade. A interpretação das frases em (2.18) resulta numa mesma representação, a DRS (2.25):

x,y
fazendeiro(x) *burro(y)* *ter(x,y)*

$\Rightarrow$

x,y,u,v
fazendeiro(x) *burro(y)* *ter(x,y)* *u=x* *v=y* *bater(u,v)*

(2.27)

2.3.4 A Proposta de Dagan e Itai

As abordagens simbólicas baseiam-se no pressuposto de que, quando um transmissor cria um texto, ele o faz seguindo regras léxicas, sintáticas, semânticas e pragmáticas definidas. Estas regras não são, necessariamente, conscientes para o transmissor. Porém a aplicação destas regras, conjuntamente com adaptações culturais, conhecimento de senso comum, exceções etc, produz textos com alguns padrões, os quais podem ser, parcialmente, detectados por abordagens estatísticas. Esta é a premissa básica dos métodos estatísticos.

Dagan e Itai [Dagan e Itai 1990] propõem uma abordagem para a escolha de pronomes ambíguos. Eles realizaram um experimento com a resolução de referências para o pronome "it" (do inglês) em frases aleatórias de um texto, com resultados específicos para o tratamento da língua inglesa. O modelo usa os pares de ocorrência observados em um corpus como padrão de seleção para futuras resoluções anafóricas. Os candidatos a antecedente são substituídos pelas anáforas, sendo que somente podem ser selecionados aqueles que possuem uma ocorrência regularmente observada nos textos.

Um exemplo desta abordagem é ilustrada em:

(2.28) They knew fully well that the companies held tax money$_k$ aside for collection later on the basis that the government$_i$ said it$_i$ was going to collect it$_k$.

Existem duas ocorrências do pronome "it" no exemplo 2.28. A primeira é o sujeito do verbo *collect* e a segunda seu objeto. A estatística é então usada para selecionar entre os três candidatos na frase: *money, collection* e *government*. A tabela 1 relaciona os padrões produzidos pela substituição de cada candidato pela sua anáfora e o número de vezes que estes padrões ocorrem no corpus analisado (Harvard):

Relação Sintática	Palavra 1	Palavra 2	Nº ocorrência
subject-verb	collection	collect	0
subject-verb	money	collect	5
subject-verb	government	collect	198
verb-object	collect	collection	0
verb-object	collect	government	0
verb-object	collect	money	149

Tabela 1: Co-ocorrência de palavras no corpus Harvard.

De acordo com a tabela 1, *government* é o antecedente preferido para o primeiro *it* (198 ocorrências), enquanto *money* é o antecedente preferido para o segundo *it* (149 ocorrências).

Este exemplo demonstra que as restrições semânticas definidas podem eliminar os resultados absurdos, selecionando os mais corretos. Esta idéia pode ser usada para a seleção de diversos outros tipos de ocorrências na análise de corpus, porém isto não elimina a possibilidade de que, em determinados casos, mais de um candidato possa ser selecionado como antecedente, havendo então a necessidade de outro método para a escolha do mais adequado. Esta escolha é, em geral, baseada no significado das palavras, na semântica e em conhecimento de senso comum.

2.4 Avaliação das Propostas para Resolução de Anáforas

Nesta seção são avaliadas as propostas para resolução de anáforas.

Veja o seguinte exemplo:

(2.29) a. Esta casa foi assaltada.

b. A porta ficou aberta.

c'. A fechadura foi arrombada.

c". A mobília foi roubada.

Na primeira frase, a interpretação do sintagma nominal definido[13] *esta casa* introduz uma entidade que é indiretamente referenciada na frase seguinte pela entidade introduzida pela interpretação do sintagma nominal definido *"a porta"*. Note-se aqui que a entidade *a porta* não poderia ser interpretada relativamente ao contexto do discurso se a entidade *casa* não tivesse sido previamente introduzida, ou seja, é a entidade *casa* que serve de antecedente para a entidade *porta*.

A abordagem de Sidner (Teoria do Foco) considera *a casa* como sendo o foco da primeira frase. Esta entidade servirá de referência para a resolução de futuras entidades anafóricas. Assim a interpretação da entidade *porta* introduzida na segunda frase terá como antecedente preferencial a entidade *casa* e o contexto *implícito* da mesma. O contexto implícito é representado por uma taxinomia de entidades e atributos. Apesar do foco servir de base para o posicionamento nesta estrutura [Cohen e Erteschik-Shir 2002], não é claro qual é o seu limite de atuação, ou seja, considerando uma representação em

[13] Apesar da entidade se apresentar, sintaticamente, como um SN definido, por ser a primeira frase do discurso ela comporta-se com um indefinido e portanto não é anafórica, apesar de ser específica.

grafo, qual o número de nós que podem ser considerados como sendo entidades envolvidas no contexto.

Já na Teoria de Centragem, para a interpretação da entidade *"porta"*, não sendo possível estabelecer uma relação de especificação direta entre a porta e a casa, opta-se por estabelecer uma relação de especificação indireta entre ambas, denominada *realização*.

Outra diferença entre anáforas nominais definidas e as anáforas pronominais é que o processo de resolução das últimas necessita apenas identificar o antecedente, enquanto as primeiras necessitam também da identificação da relação entre a anáfora e o seu antecedente [Bos 2003]. Nenhuma das propostas para resolução de anáforas apresentadas tratam desta identificação. O mais que fazem é introduzir relações genéricas, como o caso da "realização direta ou não" da Teoria da Centragem, ou apenas prever esta necessidade (sem construir algo), como é o caso da Teoria do Foco.

Esta tese propõe uma metodologia que permite tanto identificar o antecedente quanto propor uma identificação de algumas relações estereotipadas e funcionais (vide equação 2.4). A identificação destas relações enriquece a interpretação do discurso por parte do receptor (vide o princípio-I de Levinson cf. [Huang 1994]). No caso específico as entidades *casa* e *porta* estão ligadas por uma relação estrutural: *parte_ de(porta, casa).*

Retomando o exemplo 2.29, na sua última frase a entidade introduzida pelo SND *a mobília* também faz referência indireta à entidade *casa* introduzida na primeira frase. Esta noção de que certas entidades são referenciadas explicitamente (por pronomes, por exemplo) e outras implicitamente (por SND, por exemplo) dá origem à idéia da existência de dois níveis de interpretação de entidades anafóricas: referência direta ou explícita e referência indireta ou implícita. Nesta proposta estes dois níveis são contemplados através da introdução de duas listas de entidades: os sintagmas nominais indefinidos, os pronomes e as elipses dão origem à lista de entidades representativas do nível explícito. Já os SNDs dão origem à lista de entidades representativas do nível implícito. A resolução de uma anáfora (dependendo do seu tipo) é feita mediante a procura de um antecedente: primeiro na sua lista correspondente e depois na lista complementar. O capítulo 4 trata em maior profundidade as entidades implícitas e explícitas, apresentando a forma como a lista de entidades é ordenada e também a noção de foco implícito e foco explícito, os quais são essenciais para a criação da estrutura nominal apresentada no mesmo capítulo.

2.4.1 Teoria do Foco

A Teoria do Foco [Sidner 1981] foi proposta para ser uma solução computacional para a resolução de anáforas pronominais, onde fosse requerido um mínimo de conhecimento sobre as entidades envolvidas na escolha dos antecedentes. Porém a complexidade das regras de interpretação e atualização das estruturas internas, sendo que muitas das regras são ad hoc, torna a Teoria do Foco de difícil (senão impossível) adaptação para outros tipos de fenômenos de co-referência, em especial para as Anáforas Nominais Definidas.

Outro problema implícito na Teoria do Foco é a utilização da informação temática na determinação do foco. Apesar dos pesquisadores da área admitirem papéis temáticos tais como agente, tema, paciente, instrumento, localização etc, não existe uma definição consensual sobre a definição de cada um deles, o que, no caso da Teoria do Foco, levou Sidner a adaptar a noção de animacidade para suprir a definição de agente (entidade animada).

A principal contribuição da Teoria do Foco é a utilização de dois focos, mas que foram reformulados visando a sua utilização na teoria proposta nesta tese: olhando para o tratamento das anáforas nominais definidas é necessário que os critérios para determinação dos focos tenham uma componente semântica bem influente, visto que o conhecimento necessário para a determinação de uma anáfora nominal definida é maior do que o conhecimento necessário para a resolução de uma anáfora pronominal.

2.4.2 Teoria da Centragem

Os centros de atenção de uma frase são, preferencialmente, "efetivados" nas frases seguintes através do uso de pronomes ou sintagmas nominais definidos. Um *Cb* é preferencialmente efetivado como um pronome e é o sujeito da frase. Grosz et al [Grosz, Joshi e Weinstein 1995] argumentam que os sintagmas nominais podem ter argumentos implícitos que podem vir a ser o *Cb* de uma frase. Por exemplo, se uma casa é o centro de atenção corrente, suas portas e janelas são argumentos implícitos que não necessariamente foram citados.

É comum que o receptor não tenha o conhecimento completo de todas as entidades envolvidas e o seu relacionamento. Esta falta de informação de senso comum é suprida pela forma sintática com que o emissor introduz as entidades, pelo grau de saliência das mesmas e por algumas pequenas regras pragmáticas. Isto permite ao receptor identificar minimamente uma relação entre a anáfora e o seu antecedente para além de dizer-se

apenas que "a casa foi *efetivada* pela porta".

A Teoria da Centragem tem algumas restrições quanto à forma como os centros de atenção podem ser utilizados. Para cada frase do segmento:

1. existe somente um Cb,
2. todos os elementos de $Cf(F_n)$ devem ser efetivados na frase F_n,
3. o $Cb(F_n)$ é o elemento melhor classificado de $Cf(F_{n-1})$ que é efetivado na frase F_n.

As regras para a realização do Cb nas frases subseqüentes são:

1. se um elemento de $Cf(F_{n-1})$ é efetivado na frase F_n por um pronome, então o $Cb(F_n)$ também deve ser efetivado por um pronome.
2. existem três transições possíveis para os centros de atenção entre duas frases consecutivas F_n e F_{n+1}:
   (a) *continuação:* $Cb(F_{n+1})$ e $Cb(F_n)$ são a mesma entidade e o elemento melhor classificado de $Cf(Fn+1)$,
   (b) *retenção:* $Cb(F_{n+1})$ e $Cb(F_n)$ são a mesma entidade mas o centro de atenção $Cb(F_{n+1})$ não é o melhor elemento em $Cf(Fn+1)$,
   (c) *mudança:* $Cb(F_{n+1})$ é diferente de $Cb(F_n)$.

O elemento melhor classificado em $Cf(F_n)$ é o centro de atenção preferido – $Cp(F_n)$. O resultado desta ordenação pode ser visto na tabela 2 onde já estão incorporadas as modificações propostas por Brennan et al [Brennan, Friedman e Pollard 1987], nomeadamente na divisão do item *mudança*.

	$Cb(F_n) = Cb(F_{n-1})$	$Cb(F_n) \neq Cb(F_{n-1})$
$Cb(F_n) = Cp(F_n)$	**continuação**	**mudança**
$Cb(F_n) \neq Cp(F_n)$	**retenção**	**mudança radical**

Tabela 2: Movimentação dos focos segundo a Teoria da Centragem.

Um segmento coerente do discurso tem frases sobre uma mesma entidade (continuação) antes de introduzir uma entidade relacionada (retenção), a qual poderá vir a ser o novo centro de atenção (mudança ou mudança radical). Os centros de atenção vão mudando de acordo com o comportamento das anáforas de cada frase.

Comparando a Teoria da Centragem com a Teoria do Foco, pode-se dizer que o Cb corresponde ao Foco do discurso (DF), enquanto que a lista Cf corresponde à lista de focos potenciais do discurso. O Cp captura a noção de foco esperado. A noção de foco do ator, introduzida para lidar com frases com muitos pronomes, não existe aqui.

2.4.3 DRT

A DRT [Kamp e Reyle 1993] é uma abordagem essencialmente semântica, onde há a limitação do universo de escolha do referente de acordo com a interpretação do escopo dos quantificadores em frases simples, o que torna o universo de possíveis antecedentes bem limitado e simplifica, e muito, o processo de validação de um antecedente recorrendo apenas às informações sintáticas e informações semânticas.

A DRT clama a representação do discurso como um todo, representação esta que é obtida incrementalmente a partir da interpretação de cada frase no contexto anterior. Porém esta representação do discurso impões poucas restrições à escolha de antecedentes para anáforas onde as restrições semânticas sejam escassas, por exemplo, em frases sem palavras como: todo, alguns, etc. O número de indivíduos acessíveis, expresso, por exemplo, no número de referentes do universo da DRS, tende a crescer, aumentando assim o número de possíveis antecedentes. Aqui seria necessário, por exemplo, um processo de segmentação da DRS que não é feito na proposta inicial da DRT [Kamp e Reyle 1993]. Asher [Asher 1993] no seu trabalho sobre entidades abstratas propõe uma DRT segmentada que, apesar de fornecer um processo de segmentação para as DRSs, continua a sofrer da deficiência destas abordagens: não permitir a separação do processo de escolha de um antecedente do processo de validação do mesmo.

3 Relações de Ligação

"*Deus pensa, o homem sonha, a obra nasce.*"
Fernando Pessoa

Neste capítulo é apresentada a metodologia de obtenção da relação $\mathcal{R}$ da fórmula $\mathcal{R}(\mathcal{T}, \mathcal{A})$ introduzida no capítulo 1. Para tal, são definidas a linguagem de representação semântica (uma versão estendida da DRT [Kamp e Reyle 1993]), as relações $\mathcal{R}$ que são tratadas nesta tese e a metodologia que permite obtê-las computacionalmente.

3.1 Introdução

Considere um discurso D coerente, formado por n frases, tal que: $D = f_1, f_2, ..., f_{i-1}, f_i, ..., f_n$ é a concatenação das frases f_i, com $1 \leq i \leq n$. Considere agora que se deseja obter uma interpretação I_D para D, i.e., uma representação semântica que seja válida em circunstâncias equivalentes à validade de D, de acordo com o conhecimento daquele que recebe o discurso – **o receptor**[1]. O processo de obtenção de I_D é incremental, feito frase a frase. Dada uma frase f_i qualquer, primeiro obtém-se sua interpretação $I_i^{parcial}$ – denominada **interpretação parcial ou fora de contexto** – que é então *agregada* à interpretação $I_{(i-1)}$ das $(i-1)$ frases anteriores – o denominado **contexto**.

Esta *agregação* é mais do que a mera soma das interpretações parciais de i frases do discurso. Decorre da distribuição do conhecimento pelo **emissor**[2] nas diversas frases de um discurso, onde cada frase carrega em si uma parcela da informação transmitida. Mais ainda, este conhecimento é estruturado e codificado nas frases em duas partes: (1) na informação autocontida na frase (interpretação parcial) e (2) na informação de como ligar esta interpretação parcial com o conhecimento expresso pelo discurso como um todo. O resultado é que o processo de obtenção de I_D é realizado em duas etapas:

1. Cada frase f_i do discurso é interpretada individual e incrementalmente, buscando-se uma representação semântica válida $I_i^{parcial}$ para a informação expressa em f_i (de maneira semelhante ao proposto na Core Language por Alshawi [Alshawi 1990]). Esta etapa é chamada de **processo de interpretação fora de contexto**.

2. $I_i^{parcial}$ deve então ser *agregada* ao contexto formado pela interpretação das frases anteriores $I_{(i-1)}$. Esta etapa, chamada de **processo de interpretação em contexto**, é dividida em duas fases:

   (a) A interpretação fora de contexto é adicionada ao contexto obtendo-se a interpretação I_i^{fase1}, mas sem apresentar coesão:

   $$I_i^{fase1} = I_{(i-1)} \cup I_i^{parcial} \tag{3.1}$$

   (b) É estabelecida a coesão entre $I_i^{parcial}$ e $I_{(i-1)}$ através da identificação de um conjunto de ligações semânticas $\omega_{(i-1)}^{i}$. Este conjunto nada mais é do que no-

[1] O receptor é qualquer pessoa ou programa de computador que leia/ouça/receba o discurso e queira "entender", mesmo que parcialmente, o que ali está expresso.

[2] O emissor é, em complemento ao receptor, qualquer pessoa ou programa de computador que escreva/fale/emita um discurso com a intenção de transmitir algum conhecimento, o qual é, pelo menos "parcialmente", desconhecido para o receptor.

vas informações que não estão presentes nem em $I_i^{parcial}$ nem em $I_{(i-1)}$, mas que devem ser acrescentadas a I_i como resultado da interpretação de fenômenos do discurso tais como: anáforas, elipses, paralelismo, referência temporal, referência espacial entre outros[3]. O resultado é a equação (3.2), onde r é uma relação:

$$I_i = I_i^{fase1} \cup \omega_{(i-1)}^i \,|\, \forall r\, r \in \omega_{(i-1)}^i \wedge r \notin I_{(i-1)} \wedge r \notin I_i^{parcial} \tag{3.2}$$

A interpretação obtida, seja em contexto ou fora de contexto, nada mais é do que uma **Estrutura de Representação do Discurso** (DRS do inglês, *Discourse Representation Structure*). Uma DRS é a forma de representação semântica para representação de frases e discursos proposta na **Teoria da Representação do Discurso** [Kamp e Reyle 1993]. A DRT, proposta por Kamp e Reyle e baseada em Montague [Dowty, Wall e Peters 1981], tem sido amplamente utilizada em PLN para a representação semântica de frases e textos.

A DRT é caracterizada por: (1) apresentar um formalismo simples para a representação de frases, (2) adotar a idéia de processamento incremental do discurso, (3) basear-se na teoria dos modelos, apresentando uma distinção entre a representação semântica obtida e a forma como esta pode ser *logicamente* provada e (4) é facilmente extensível ao tratamento/representação de fenômenos do discurso não tratados na DRT. Em virtude destas características, adotou-se a DRT como formalismo para representação semântica do discurso/frases, fazendo-lhe extensões para comportar o tratamento das Anáforas Nominais Definidas. Uma versão simplificada da DRT é apresentada na seção 3.2 e as extensões feitas serão apresentadas no decorrer do capítulo.

Com relação aos fenômenos tratados, como apresentado no capítulo 2, este trabalho trata da interpretação das Anáforas Nominais Definidas, a qual pode ser resumida no processo pelo qual os elementos $\mathcal{T}$ (antecedente) e $\mathcal{R}$ (relação) da fórmula $\mathcal{R}(\mathcal{A}, \mathcal{T})$ são determinados. Formalmente, as instâncias de $\mathcal{R}(\mathcal{A}, \mathcal{T})$ são elementos do conjunto $\omega_{(i-1)}^i$ (equação 3.2), onde cada elemento $\omega_{(i-1)}^i$ é uma *resposta* à interpretação das condições especiais $snd(Ref)$[4] que diferenciam os referentes Ref introduzidos por SNDs do restante dos referentes existentes em $I_i^{parcial}$. Tais condições disparam, durante a interpretação em contexto, o processo de resolução anafórica e são, por esta razão, denominadas de **condições gatilho**.

Em relação à identificação de $\mathcal{T}$ e $\mathcal{R}$, este capítulo *trata somente* da determinação da

[3] São estes fenômenos que vão estabelecer a estrutura do conhecimento distribuído no discurso.

[4] Introduzidas pela interpretação dos Sintagmas Nominais Definidos. Tal condição *marca* as entidades, determinando que elas muito provavelmente são anafóricas e necessitam ser resolvidas.

relação $\mathcal{R}$. Para tal, considera que $\mathcal{T}$ já seja conhecido (o capítulo 4 trata da determinação de $\mathcal{T}$). A determinação da relação $\mathcal{R}$ é um processo onde se deve considerar que:

1. Dadas duas entidades quaisquer, supostamente expressão anafórica e antecedente, o processo de identificação das relações entre ambas deve considerar um número expressivo de possibilidades quando não se tem nenhuma indicação contextual [Markert, Strube e Hahn 1996, Hovy 1990]. Por exemplo:

   - carro e elefante.
   - casa e carro.
   - sala e sapato.

   Sem contexto, as relações possíveis entre cada par podem ser as mais variadas: podem ser cinco elefantes dentro de um carro (piada), ou elefante na frente/trás/cima/lado/embaixo do carro etc. Carro em casa, na frente da, dentro da etc. Isto torna complexa, ou pelo menos muito trabalhosa, a tarefa de identificar as relações de ligação, ainda mais se forem considerados os objetos abstratos [Asher 1993].

2. Considerando-se que todas as possíveis relações pudessem ser totalmente mapeadas, então tem-se um problema ainda maior na sua utilização: num dado contexto podem ser identificadas mais de uma relação entre duas entidades. Conseqüentemente, o processo de eliminação/classificação das possíveis soluções será oneroso considerando que as possíveis interpretações de um discurso são o produto cruzado das interpretações parciais de cada frase. As interpretações parciais de cada frase serão, por sua vez, o produto cruzado das relações identificadas e dos possíveis antecedentes. Com isto a interpretação do discurso tende a ser um processo computacional pesado. Melhorias na restrição do produto cruzado, como a proposta nesta tese, servem para otimizar o processo computacional.

Esta complexidade é decorrente de uma única consideração, implícita até agora: para se determinar as relações foi necessário considerar o domínio de conhecimento que engloba as entidades referenciadas. O que esta tese propõe é um rompimento com este paradigma, passando a utilizar a informação lingüística (i.e. informação léxica e sintática) como base para a obtenção das relações entre entidades. A estrutura do conhecimento transmitido também é expressa e estruturada nas diversas frases do discurso na forma lingüística através de: anáforas, elipses, foco do discurso entre outras.

Por sua vez, o processo para determinação das relações é uma interpretação pragmática das entidades do discurso [Filho e Freitas 2003]. Interpretação pragmática no sentido de que é o conhecimento de senso comum da língua e suas estruturas léxicas e sintáticas que permitem estabelecer relações entre duas entidades [Abbott 1993, Polanyi 1988].

3.2 A Teoria da Representação do Discurso

A DRT tem como seu elemento básico as Estruturas de Representação do Discurso (DRS). Cada frase f_i ao ser interpretada fora de contexto dá origem a uma DRS **básica** $K_i^{parcial}$ (interpretação fora de contexto) que deverá ser *agregada* à DRS **global** $K_{(i-1)}$ (contexto de interpretação), resultante da interpretação das $(i-1)$ frases anteriores, isto é, $K_{(i-1)}$ é o contexto de interpretação para f_i.

3.2.1 A obtenção das DRS

Uma DRS é constituída por dois componentes:

1. O **universo da DRS,** um conjunto de *referentes do discurso* (ou simplesmente, referentes), cada qual representando uma única entidade semântica.
2. O conjunto de **condições DRS,** um conjunto de proposições aplicadas sobre os referentes do discurso.

Esquematicamente uma DRS é apresentada sob a forma de um caixa dividida em duas partes. Na parte superior está o conjunto de referentes do discurso, representando as entidades introduzidas no próprio discurso (universo da DRS) e, na parte inferior, estão as condições semânticas aplicadas aos referentes do universo da DRS, conforme o diagrama (3.3):

Universo DRS
condição 1 *condição 2* . . .

(3.3)

Formalmente uma DRS é um par ordenado $< U_k, Conds_k >$, onde U_k é o universo da DRS e $Conds_k$ as condições da DRS.

O significado lingüístico de uma DRS é dado pelo conjunto de condições e referentes. A avaliação semântica das DRSs é feita em modelos. A representação DRS é recursiva, i.e., uma DRS pode ter como condição outra DRS. As DRSs que estão como condições de uma DRS são denominadas ***subDRSs*** e a DRS que engloba todas as outras é chamada ***DRS principal***.

A obtenção de uma DRS é um processo que envolve a utilização de regras de reescrita: algoritmos que transformam um determinado tipo de subárvore da análise sintática em referentes do discurso e condições DRS, integrando-os à representação semântica da frase que está sendo interpretada.

De maneira a ilustrar o processo, considere o exemplo:

(3.4) Marcelo ama Patrícia.

E sua respectiva Árvore de Derivação Sintática (**ADS**):

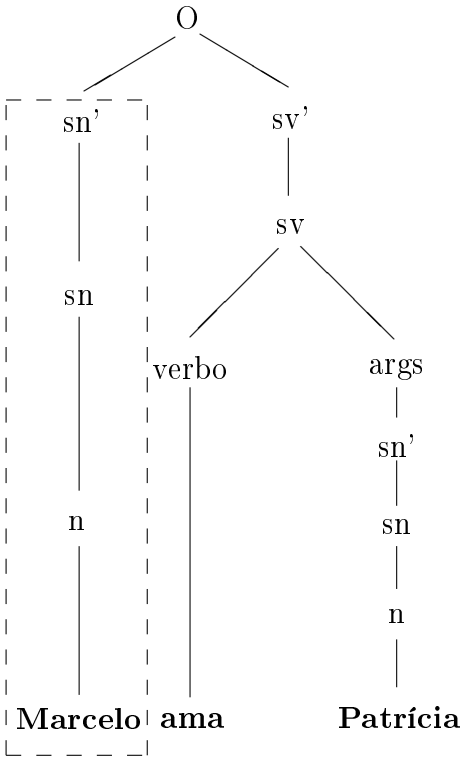


Figura 1: Árvore de derivação sintática

A subárvore do SN *Marcelo* (destacado na figura 1) é reduzida através das seguintes operações:

- Sintagmas nominais de nomes próprios criam novos referentes e devem ser introduzidos no universo da DRS principal.
- A redução da subárvore, cria uma condição formada pelo nome próprio seguido do referente entre parênteses.
- O referente criado substituirá a subárvore.

O resultado destas operações é a DRS (2a).

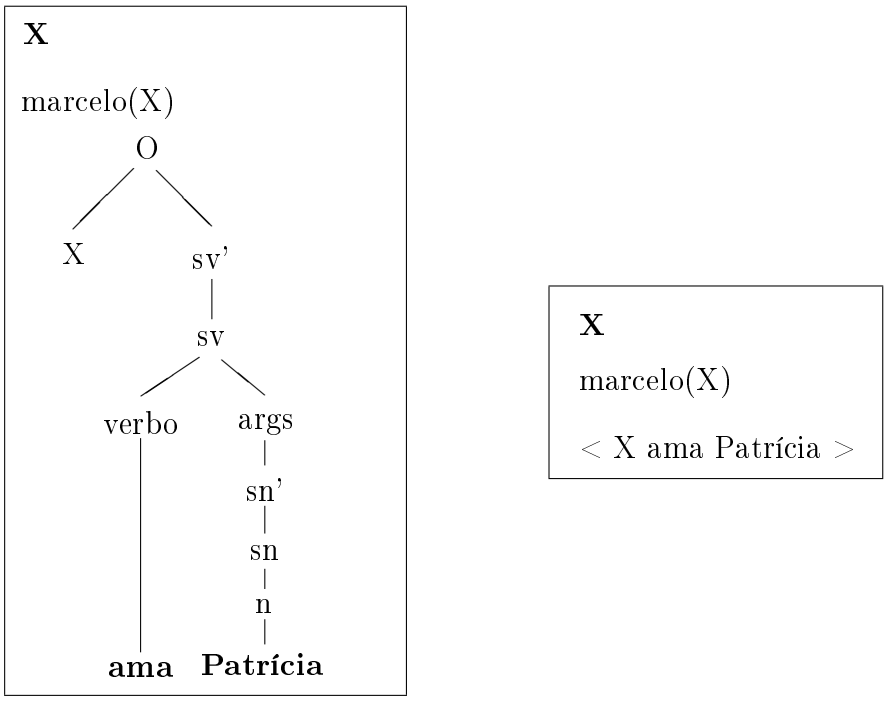


Figura 2: Formas para representação da redução de uma frase.

Existem duas formas gráficas para a representação das consecutivas reduções de uma árvore sintática. Na figura (2a), a DRS contém, além das condições, também a ADS já parcialmente reduzida. Na figura (2b), a representação da árvore é *linear*. Em virtude das duas traduzirem a mesma informação, padronizou-se o uso da ADS linear, por motivos de facilidade de representação.

Continuando a redução da árvore na figura (2a), as mesmas operações utilizadas na redução do sintagma nominal *Marcelo*, são utilizadas na redução do sintagma nominal

Patrícia (DRS (3.5)). Nela somente falta a redução da árvore < x ama y>, que será substituída pela condição amar(x,y)[5], resultando na DRS (3.6).

x, y
marcelo(x) *patrícia(y)* *<x ama y>*

(3.5)

x, y
marcelo(x) *patrícia(y)* *amar(x, y)*

(3.6)

Assim todos os indivíduos e verbos da primeira frase do texto 3.4 foram reduzidos, dando lugar ao conjunto referentes e condições. Ao conjunto de operações que governam a introdução de condições, criação de referentes e redução de árvores sintáticas, dá-se o nome *regras de reescrita.* O conjunto de regras de reescrita aplicado sobre frases individuais simples (sem negação ou conjunções) é formalizado da seguinte maneira:

- Uma **DRS Simples** K – que não contém uma subDRS como condição – confinada a um vocabulário $\mathcal{V}$ e a um conjunto $\mathcal{R}$ de referentes, é um par $< U_K, Cond_K >$ constituído por um subconjunto U_K de $\mathcal{R}$ e por um conjunto de condições DRS confinadas a $\mathcal{V}$ e a $\mathcal{R}$.

- Uma condição DRS (simples ou atômica) confinada a $\mathcal{V}$ e a $\mathcal{R}$ é uma expressão que tem uma das seguintes formas:

  1. $x = y$, pertencendo x e y a $\mathcal{R}$.
  2. $\pi(x)$, sendo x um referente de $\mathcal{R}$ e π um nome próprio de $\mathcal{V}$.
  3. $\eta(x)$, sendo x um referente de $\mathcal{R}$ e η um predicado unário (correspondente a um substantivo comum) de $\mathcal{V}$.

[5] Os verbos vão ser sempre representados no infinitivo, a não ser que, por uma questão de facilidade de leitura, seja necessário colocá-los no tempo do discurso. Este tratamento não traz prejuízo para o trabalho em si, pois nem os tempos verbais nem as entidades por eles introduzidas são tratados nesta tese.

4. $\zeta(x)$, sendo x um referente de $\mathcal{R}$ e ζ um predicado unário (correspondente a um verbo intransitivo) de $\mathcal{V}$.
5. $\xi(x, y)$, sendo x e y referentes de $\mathcal{R}$ e ξ um predicado binário (correspondente a um verbo transitivo) de $\mathcal{V}$.
6. $snd(x)$, sendo x referente de $\mathcal{R}$ e snd um predicado unário, correspondente à indicação de que um referente foi introduzido por um SND.
7. $\rho(x, y)$, sendo x e y referentes de $\mathcal{R}$ e ρ um predicado binário correspondente a uma possível relação existente entre x e y.

Salientando que os itens 6 e 7 foram introduzidos nesta tese. Alguns exemplos que dão origem a estas condições anteriores são:

1. *Marcelo ama Patrícia. Ela o fascina.* Considerando x e y como sendo os referentes introduzidos, respectivamente, por *Marcelo* e *Patrícia*. O referente z introduzido pelo pronome pessoal **Ela** está co-referindo o referente introduzido pelo nome próprio *Patrícia*, resultando em $z = y$ e o referente w introduzido pelo pronome oblíquo **o** está co-referindo o referente introduzido pelo nome próprio *Marcelo*, resultando em $w = x$.
2. A condição $marcelo(x)$, retirado do exemplo anterior.
3. *Marcelo tem um carro*, o sintagma nominal indefinido *um carro* introduz a condição $carro(w)$.
4. *Patrícia partiu.* O nome próprio *Patrícia* introduz o referente x e o verbo intransitivo insere a condição $partir(x)$.
5. Usando o exemplo do item 3, o verbo transitivo *ter* é transformado na condição $ter(x, w)$.

3.2.2 Algoritmo para a obtenção das DRSs

Considere:

- o discurso $D = f_1, f_2, ..., f_{i-1}, f_i, ..., f_n$,
- $K_i^{parcial}$como sendo a DRS resultante da interpretação fora de contexto da frase f_i,

- K_i como sendo a interpretação em contexto das i frases do discurso e
- $K_0 = \phi$, i.e. sem nenhum referente nem condições, como sendo o contexto de partida para a interpretação da frase f_1.

Logo define-se o algoritmo para obtenção de uma DRS K_i em contexto, como:

Repetir para $i = 1, 2, \ldots, n$ onde n é o número de frases do discurso.

1. Juntar a árvore de derivação sintática da frase $f_i[O_i]$ às condições da DRS $K_i^{semi-parcial}$, inicialmente vazia.
2. Acrescentar às regras de reescritas originais o tratamento para subárvores que indiquem sintagmas nominais definidos, caso no qual estas introduzirão um novo referente do discurso x, a condição $snd(x)$ indicando que x é um SND e todas as condições aplicáveis sobre este referente que foram derivadas das informações sobre a entidade expressa pelo SND (e.g. o carro: $carro(x), singular(x), masculino(x), objeto\_direto(x), etc$).
3. Dado o conjunto de condições redutíveis em $K_i^{semi-parcial}$, aplicar as *regras de reescrita* até que não existam mais condições redutíveis em $K_i^{semi-parcial}$, caso em que a DRS só tem informação semântica e, portanto, $K_i^{parcial} = K_i^{semi-parcial}$.

   A proposta original da DRT [Kamp e Reyle 1993] define apenas os passos 1 e 3, pois considera que as regras de reescrita são suficientes para interpretar os fenômenos do discurso tais como anáforas pronominais (co-referência, onde surgem condições do tipo: $x = y$). Porém, as regras não são suficientes para o tratamento de SNDs, as quais necessitam também identificar a relação, isto é, uma nova condição DRS ou o predicado da fórmula $R(A, T)$. Além da introdução do passo 2, os passos a seguir são extensões ao algoritmo proposto originalmente pela DRT e os quais vão na linha do processo de interpretação proposto na introdução deste capítulo.
4. Unir a DRS $K_i^{parcial}$ com a DRS contexto K_{i-1} obtendo-se K_i^{f1}:

$$K_i^{f1} = K_{i-1} \cup K_i^{parcial} \tag{3.7}$$

   lembrar que $K_i^{f1} =< Universo_{K_i^{f1}}, Conds_{K_i^{f1}} >$ e portanto (3.7) é desmembrada em duas:

$$U_{K_i^{f1}} = U_{K_{i-1}} \cup U_{K_i^{parcial}} \tag{3.8}$$

e

$$Conds_{K_i^{f1}} = Conds_{K_{i-1}} \cup Conds_{K_i^{parcial}} \tag{3.9}$$

5. Cada condição $snd(x)$ encontrada em K_i^{f1} pode indicar que a entidade denotada pelo referente x é anafórica, caso em que é disparado o processo que encontra um provável antecedente y (a descrição de como isto é feito está no capítulo 4).

6. Encontrado ou não um antecedente, sempre haverá um conjunto de possíveis relações $\omega^i_{(i-1)}$ entre as entidades marcadas como *snd* na frase f_i e alguma entidade do contexto K_{i-1}. $\omega^i_{(i-1)}$ nada mais é do que um conjunto de condições DRSs as quais devem ser adicionadas ao conjunto de condições de K_i^{f1}:

$$K_i =< U_{K_i^{f1}}, Conds_{K_i^{f1}} \cup \omega^i_{(i-1)} > \tag{3.10}$$

O restante deste capítulo vai centrar-se na forma como as relações podem ser obtidas (passo 6).

3.3 Determinação das relações de ligação

As ANDs são SNDs cuja interpretação é anafórica (no inglês - bridging [Clark 1977, Heim 1982]). Como já visto, nestes casos deve-se determinar tanto o antecedente $\mathcal{T}$, quanto a relação $\mathcal{R}$ existente entre a entidade introduzida pela interpretação da expressão anafórica e este antecedente. Considere o exemplo (3.11):

(3.11)

a. Levei meu carro para fazer uma revisão.

b. O motor estava com um barulho estranho.

Neste exemplo, o uso do SND *"o motor"* na frase (3.11b) indica a existência de uma entidade previamente introduzida (e já conhecida pelo receptor) a qual dá contexto à existência do motor no discurso. Seu antecedente é o carro[6] da frase (3.11a). Mas qual será a ligação entre o motor e o carro? É plausível assumir que o motor *é parte* do carro. Note que esta informação *parece vir* do conhecimento sobre o domínio automobilístico que os interlocutores têm. Isto leva à conclusão de que a relação está preestabelecida [Strand 1996]

[6] A forma como o antecedente pode ser localizado é tema do capítulo 4.

na mente dos interlocutores, sendo apenas necessário validá-la com argumentos extraídos da interpretação do discurso.

Caso isto fosse uma verdade absoluta no processo de interpretação, seria como dizer que este não é mais do que um *casamento de padrões preestabelecidos*. Hobbs et al [Hobbs et al. 1993] identificaram que em cada frase do discurso existe uma componente de aprendizagem para o receptor motivando o transmissor a emitir o discurso. De uma maneira simples, se o receptor já conhece tudo, então nada precisa ser dito! Hobbs utiliza a abdução com pesos para identificar o que está sendo aprendido. Com relação às ANDs ele identifica que a componente de aprendizagem é o encapsulamento da expressão anafórica, do seu antecedente e da relação entre eles no contexto do discurso, deixando em aberto que a relação também pode ser algo novo. Porém não detalha a forma com que uma relação pode ser determinada.

Retomando a discussão sobre o conhecimento do domínio, não se nega que este é necessário. Porém existe um tipo de conhecimento o qual pode ser utilizado: a informação codificada na linguagem em si. Esta informação pode ser utilizada para a interpretação de ANDs e também para a identificação de alguns tipos de relações entre entidades. É bem conhecida a importância da estrutura do discurso para a resolução de anáforas pronominais [Grosz e Sidner 1986], o paralelismo para a resolução de elipses [Kehler 1993] e o movimento dos centros de atenção para a coerência do discurso [Grosz, Joshi e Weinstein 1995]. Na mesma linha dessas propostas, este trabalho usa a informação morfológica, a informação sobre coletivos (existente em qualquer dicionário de português) e a informação sobre a animacidade das entidades [Sidner 1979] para calcular a relação existente entre as duas entidades de uma resolução anafórica.

3.3.1 As regras pragmáticas

Baseado no conhecimento que as pessoas têm sobre a língua que falam é possível estabelecer um conjunto pragmático de regras a serem utilizadas na determinação da relação entre a expressão anafórica e seus antecedentes. As informações sobre gênero, número e grau, coletivos e animacidade [Sidner 1979], podem ser utilizadas na determinação das seguintes relações:

co-referência: indicando que tanto $\mathcal{A}$ quanto $\mathcal{T}$ denotam a mesma entidade: $\mathcal{A} = \mathcal{T}$.

membro de: indicando que a entidade denotada por $\mathcal{A}$ é um membro do conjunto de entidades denotada por $\mathcal{T}$.

parte de: indicando que a entidade denotada por $\mathcal{A}$ é parte (estrutural) da entidade denotada por $\mathcal{T}$.

subcategorizado por: indicando que a entidade denotada por $\mathcal{A}$ é, de alguma forma, uma *parte* conceitual da entidade denotada por $\mathcal{T}$.

3.3.2 A relação de co-referência

Esta é a relação *tradicional* usada na resolução de anáforas pronominais e elipses [Sidner 1979, Hirst 1981, Carter 1987], veja o exemplo:

(3.12)

a. [Márcia]$_i$ comprou [um presente]$_j$.

b. [Ela]$_i$ deu-[o]$_j$ ao Marcelo.

Onde o referente do discurso introduzido pela interpretação do pronome *Ela* co-referencia o referente introduzido por *Márcia.* A representação semântica [Kamp e Reyle 1993] da co-referência é o símbolo '=', aplicado aos referentes do discurso. Numa primeira aproximação, é proposto que a relação de co-referência entre $\mathcal{A}$ e $\mathcal{T}$ pode ser estabelecida pela seguinte regra pragmática:

Regra Pragmática 1 *Proposta inicial para a determinação da relação de co-referência*

Se o referente $\mathcal{A}$ foi introduzido pela interpretação de um pronome ou de uma elipse, então a relação $\mathcal{R}$ é a de co-referência.

Mas a co-referência também pode ser introduzida por um SND, como em:

(3.13)

a. [Fernando]$_i$ foi a uma festa.

b'. [O idiota]$_i$ nem me escutou.

b". [Ele]$_i$ nem me escutou.

A interpretação da frase (*3.13*a) dá origem à seguinte DRS:

$$K_1 = \boxed{\begin{array}{|l} \hline f,g \\ \hline fernando(f) \\ festa(g) \\ ir(f,g) \\ \hline \end{array}} \tag{3.14}$$

Já a interpretação fora de contexto da frase (*3.13*b') dá origem à DRS:

$$K_2^{parcial} = \boxed{\begin{array}{|l} \hline i,e \\ \hline idiota(i) \\ snd(i) \\ eu(e) \\ \neg\ escutar(i,e) \\ \hline \end{array}} \tag{3.15}$$

Note que na interpretação em contexto de $K_2^{parcial}$, o referente i que denota a entidade **idiota** vai co-referenciar o referente do discurso introduzido pela interpretação do nome próprio *Fernando*, de forma que uma nova condição semântica será introduzida na DRS K_2 resultante da interpretação de $K_2^{parcial}$ no contexto K_1:

$$K_2^{parcial} + K_1 = \boxed{\begin{array}{|l} \hline f,g,i,e \\ \hline fernando(f) \\ festa(g) \\ ir(f,g) \\ idiota(i) \\ \underline{i=f} \\ \ldots \\ \hline \end{array}} \tag{3.16}$$

Quando um transmissor usa um SND (frase *3.13*b') em vez de usar um pronome, por exemplo *ele* na frase (*3.13*b"), ele está tentando enriquecer o conhecimento do receptor com mais informações sobre uma mesma entidade (no caso, o Fernando). Em termos semânticos, foram introduzidas duas novas condições no universo da DRS final K_2: uma condição $idiota(i)$ definindo o atributo *idiota* para um novo referente i e uma condição $i = f$ relacionando o referente introduzido por um SND (provavelmente anafórico) e seu antecedente, no caso f.

Isto leva à seguinte constatação: uma regra para determinação da relação de co-referência não pode ser baseada apenas na informação de tipo da expressão anafórica. Por outro lado, a informação léxica é uma fonte de informações que pode ser usada nesta determinação:

Número: Se $\mathcal{A}$ e $\mathcal{T}$ *não concordam em número, então não é possível estabelecer uma relação de co-referência entre ambas, tal como em:*

(3.17)

a. Fernando foi à festa.

b. (*) Os idiotas não me ouviram.

Note que o SND *"os idiotas"* da frase (3.17b) está no plural e não pode ter como antecedente (no discurso) a entidade introduzida pelo nome próprio *Fernando* (frase *3.17*a). Salientando que este texto poderia ter uma interpretação possível, caso fosse considerado o ambiente que o discurso foi emitido e os outros membros do conjunto de "idiotas" estivessem presentes, sendo então possível referenciá-los por um apontamento de dedo ou uso de um pronome demonstrativo [Freitas 1993]. Porém este tratamento está fora do escopo desta tese.

Gênero: Se $\mathcal{A}$ e $\mathcal{T}$ *não concordam em gênero, então não é possível estabelecer uma relação de co-referência entre ambos. Compare as frases do exemplo a seguir*:

(3.18)

a'. Fernando foi à festa.

a". Claúdia foi à festa.

b. A idiota não me ouviu.

Nas frases (3.18a') e (3.18b), *Fernando* não está co-referenciando *a idiota*, porque não concordam em gênero. Caso as frases fossem (3.18a") e (3.18b), então a relação de co-referência poderia ser estabelecida, pois ambos concordam em gênero.

Grau: A diferença de grau entre $\mathcal{A}$ e $\mathcal{T}$ não interferem na determinação da relação de co-referência, como em:

(3.19)

a. *Fernandão* foi a uma festa.

b. **O idiota** não me ouviu.

Apesar de *Fernandão* estar no aumentativo, isto não interfere com a sua ligação com o SND **o idiota**, podendo ser estabelecida a relação de co-referência. Veja ainda a seguinte variação:

(3.20)

a. *Fernando* foi a uma festa.

b. **O idiotão** não me ouviu.

O SND **o idiotão** indica que o transmissor queria acentuar o quão idiota foi *Fernando*. Mais uma vez, a relação de co-referência pode ser estabelecida. Concluindo a série de exemplos:

(3.21)

a. *Fernandão* foi a uma festa.

b. **O idiotão** não me ouviu.

Onde a relação estabelecida é de co-referência.

A proposta para identificação da relação de co-referência fica então da seguinte forma:

Regra Pragmática 2 *Determinação da relação de co-referência*

a) Se $\mathcal{A}$ tiver sido introduzido no discurso por meio de um pronome ou de uma elipse, então $\mathcal{R}$ é uma relação de co-referência.

b) Se $\mathcal{A}$ tiver sido introduzido no discurso por meio de um SND e $\mathcal{A}$ e $\mathcal{T}$ concordam em número e gênero, então $\mathcal{R}$ pode ser uma relação de co-referência.

3.3.3 A relação membro de

A relação "membro de" pode ser estabelecida entre indivíduos e conjuntos de indivíduos, como em:

(3.22)

a. Antônio abriu *algumas portas*.

b. e entrou pela **porta mais larga**.

A interpretação do substantivo indefinido plural *algumas portas* na frase (3.22a) introduz um referente do discurso p e as condições: $porta(p), plural(p)$ indicando que p é do tipo *porta* e é uma entidade coletiva (por estar no plural) composta de diversos indivíduos deste mesmo tipo:

$$K_1 = \begin{array}{|l|}\hline a,p \\ \hline antônio(a) \\ porta(p) \\ abrir(a,p) \\ plural(p) \\ \\ \hline \end{array} \tag{3.23}$$

Na frase (3.22b) a interpretação do SND *a porta mais larga* introduz o referente l e as condições $porta(l), mais\_larga(l), snd(l)$:

$$K_2^{parcial} = \begin{array}{|l|}\hline x,l \\ \hline elipse(x) \\ porta(l) \\ mais\_larga(l) \\ snd(l) \\ entrar(x,l) \\ \\ \hline \end{array} \tag{3.24}$$

A interpretação de $K_2^{parcial}$ no contexto K_1 implicará na justificativa do porquê da existência da condição $snd(l)$ considerando o referente p como sendo seu antecedente. Esta explicação levará à introdução da condição $membro\_de(l,p)$, considerando as seguintes informações:

1. A informação lingüística de que tanto *portas* quanto *porta mais larga* têm a mesma raiz (substantivo).
2. *Portas* está no plural, indicando que é uma entidade coletiva.
3. *Porta mais larga* está no singular, indicando que é um indivíduo simples.
4. Ambos concordam em gênero: feminino.

O resultado é a DRS K_2:

$$K_2 = \begin{array}{|l|}\hline a,p,x,l \\ \hline antônio(a) \\ porta(p) \\ plural(p) \\ x{=}a \\ porta(l) \\ mais\_larga(l) \\ snd(l) \\ entrar(x,l) \\ \underline{membro\_de(l,p)} \\ \\ \hline \end{array} \tag{3.25}$$

Com estas informações é possível estabelecer uma regra pragmática (inicial) para a determinação da relação *membro_de,* dando origem à seguinte regra pragmática:

Regra Pragmática 3 *Proposta inicial para determinação da relação membro_de*

Considere $\mathcal{A}$ e $\mathcal{T}$ como sendo, respectivamente, os referentes do discurso denotados pelo SND e por um dos seus possíveis antecedentes. Considere $T_{\mathcal{A}}$ como o tipo expresso pela raiz léxica do SND e $T_{\mathcal{T}}$ como o tipo expressado pela raiz do antecedente. Se $T_{\mathcal{T}} = T_{\mathcal{A}}$, $singular(\mathcal{A})$, $plural(\mathcal{T})$ e $genero(\mathcal{A}) = genero(\mathcal{T})$ então $\mathcal{A} \in \mathcal{T}$ ou $membro\_de(\mathcal{A}, \mathcal{T})$.

Considere agora o seguinte exemplo (3.26):

(3.26)

a. Ontem **uma matilha** veio até o galinheiro.

b. *Os cães* mataram cinco galinhas.

Observe os substantivos *matilha* e *cães* do texto (3.26):

1. apesar de *matilha* estar no singular, ela é uma entidade coletiva[7] e
2. *cães,* por estar no plural, também é uma entidade coletiva.

Isto é representado nas DRSs K_1 e $K_2^{parcial}$:

$$K_1 = \begin{array}{|l|}\hline m,l \\ \hline matilha(m) \\ plural(m) \\ galinheiro(l) \\ vir(m,g) \\ \\ \hline \end{array} \qquad K_2^{parcial} = \begin{array}{|l|}\hline c,g \\ \hline cão(c) \\ plural(c) \\ snd(c) \\ galinha(g) \\ plural(g) \\ matar(c,g) \\ \\ \hline \end{array} \tag{3.27}$$

Como resultado a regra pragmática 3 não pode ser aplicada ($singular(\mathcal{A})$ é falso) e a relação *membro de* não pode ser estabelecida.

Por outro lado, a relação de co-referência entre a matilha m e os cães c é preferida apesar do gênero ser diferente: $genero(c) \neq genero(m)$. Como conseqüência, a regra pragmática (2) deverá ser alterada de forma a introduzir uma nova opção (c):

[7] A qual pode ser localizada em qualquer dicionário de coletivos da língua portuguesa.

Regra Pragmática 4 *c. Se $\mathcal{A}$ tiver sido introduzido no discurso por meio de um SND e $\mathcal{A}$ e $\mathcal{T}$ concordam em número e $\mathcal{A}$ ou $\mathcal{T}$ são coletivos qualificados, então $\mathcal{R}$ pode ser uma relação de co-referência.*

Neste exemplo fica claro que a informação de número isolada não é suficiente para determinar se uma entidade é membro de outra. É necessário excluir os coletivos qualificados.

Considere agora o seguinte texto:

(3.28) a. Ontem passou *uma matilha* por aqui.

b'. *Um cão* revirou a minha lata de lixo.

b". (*) **O cão** revirou a minha lata de lixo.

b"'. **O cão líder** revirou a minha lata de lixo.

Note que o indefinido *um cão* introduzido na frase (3.28b') comporta-se como um SND, pois não foi um cão qualquer que revirou a lata de lixo, mas sim um cão pertencente à matilha da frase (3.28a). Caso fosse utilizado o definido *o cão,* como na frase (3.28b"), não seria possível estabelecer uma ligação entre este cão e a matilha (conjunto de cães) da primeira frase: o uso de um SND indica que *o cão* é uma entidade conhecida e individualizada pelo receptor, o que no contexto da frase (3.28a) não é verdade. Isto torna difícil a ligação deste cão com entidades do contexto. Note que o mesmo aconteceria caso *a matilha* fosse substituída pelo seu sinônimo: *cães*. Já na frase (3.28b"') o SND **o cão líder**, que é uma especificação de um tipo de cão, é interpretado com sendo um *membro da matilha* (ou do conjunto de cães) da frase (3.28a).

Com isto a regra pragmática 3 é revisada, alterando-se três pontos:

1. O tipo de uma entidade passa a ser determinado pela cabeça lingüística, caso o indivíduo esteja no plural, e pelo sinônimo da entidade, caso o indivíduo seja uma entidade coletiva.
2. Ou o número do antecedente está no plural ou o antecedente está marcado como uma entidade coletiva.
3. O gênero somente é aplicável caso o antecedente esteja no plural.

Como resultado destas alterações, a proposta para determinação da relação *membro de*, fica da seguinte forma:

Regra Pragmática 5 *Determinação da relação* membro_de

Considere $\mathcal{A}$ e $\mathcal{T}$ como sendo, respectivamente, um SND e um dos seus possíveis antecedentes. Considere $\mathbf{T}$ *como sendo o tipo da entidade $\mathcal{E}$, o qual é determinado da seguinte forma: se $\mathcal{E}$ está no plural então* $\mathbf{T}$ *é um conjunto único formado pela cabeça lingüística de $\mathcal{E}$ no singular; se $\mathcal{E}$ é uma entidade coletiva então* $\mathbf{T}$ *é o conjunto de sinônimos de $\mathcal{E}$. Considere agora $\mathbf{T_A}$ como sendo o tipo de $\mathcal{A}$ e $\mathbf{T_T}$ como sendo o tipo de $\mathcal{T}$. Se $T_A \cap T_T \neq \{\} \wedge singular(\mathcal{A}) \wedge plural(\mathcal{T})$ então pode-se* ***assumir*** *$\mathcal{A} \in \mathcal{T}$, ou simplesmente, $membro\_de(\mathcal{A}, \mathcal{T})$.*

De acordo com a taxinomia de Strand, a regra pragmática 5 determina uma relação de especificação: inicialmente é introduzida uma entidade mais genérica a qual é depois referenciada por uma mais específica.

Segundo Strand também é possível haver uma relação de generalização: introduzir-se uma entidade mais específica e depois generalizá-la através de um SND, como no exemplo seguinte:

(3.29) a. Comprei um *cão*.

b. **A matilha** foi toda vendida.

onde **a matilha** da segunda frase referencia e generaliza o *cão* introduzido na primeira frase. Casos de generalização para a relação *membro de* não são muito comuns, pois na maioria das vezes obriga o emissor a colocar mais informação no SND de forma a tornar a ligação mais clara. Por outro lado, quanto maior o número de entidades léxicas que compõem um SND, menor a sua dependência de uma referência:

(3.30) a. Comprei um *cão*.

b. **Este** era **o último da matilha**.

Neste exemplo, a entidade **o último da matilha** tem um contexto que depende da resolução do pronome *este*. Ou seja, é uma reafirmação da mesma anáfora, comportando-se como uma anáfora intrafrase. As anáforas intrafrases estão fora do escopo desta tese.

3.3.4 A relação *parte_de*

Esta relação é definida quando uma entidade é parte estrutural de outra. Por exemplo:

(3.31) a. Márcia chegou com seu carro.

b. Abriu a porta e desceu.

O SND *"a porta"* na segunda frase pode ser justificado se esta for *"parte do* carro" introduzido na primeira frase. A seguir são especificados os tipos de informação necessários para se chegar a esta conclusão:

1. Uma ontologia de conceitos, parcial ou não, que seja conhecida pelo receptor e baseada na relação "parte de": portas são partes de carros, motores são partes de carros, janelas são partes de casas etc. A ontologia é usada da seguinte maneira: existindo no discurso uma entidade carro e uma entidade porta, pesquisa-se se existe uma relação entre estas duas. Existindo, esta será a relação utilizada.

2. As informações textuais do discurso: o antecedente *carro* está no singular, não é um indivíduo coletivo e a expressão *porta* é um definido, então **pode-se assumir** que a porta é parte do carro. Note que não é necessário conhecer o gênero da expressão anafórica, veja o exemplo: "Márcia chegou com seu carro. Abriu *as portas* e desceu". Em termos de relacionamento, tanto *a porta* quanto *as portas* são parte do carro.

Como já foi argumentado anteriormente, este trabalho usa, preferencialmente, as informações presentes no texto e como estas podem gerar regras para uma interpretação pragmática do discurso. Assim, apesar de admitir-se que é perfeitamente possível o receptor ter uma relação *preestabelecida* entre carros e portas, optou-se por considerar que esta informação não é estritamente necessária. Isto pode ser exemplificado no exemplo a seguir (3.32):

(3.32) a. Wilson trouxe uma cesta de lanche.

b. A cerveja estava quente.

Como de costume, o SND *a cerveja* deve ser resolvido com algum antecedente. Escolhendo-se *a cesta de lanche* da primeira frase é plausível assumir, de acordo com o item 2 e em contexto, que a *cerveja* é parte da *cesta de lanche*: o receptor não necessita ter uma ontologia onde esteja explícito que **cervejas sejam parte das cestas de lanche**. O conhecimento que o receptor provavelmente tem é: **não existe nenhuma informação neste contexto invalidando a proposição** onde a cerveja é parte da cesta de lanche.

Isto pode, em termos do descrito no item 2, ser implementado como: após ter-se, de forma pragmática, identificado que existe uma relação *parte de* entre *a porta* e *um carro,* verifica-se se não existe uma informação, no conhecimento de senso comum do receptor, que torne a ligação inválida.

Uma proposta para determinação da relação *parte de* é:

Regra Pragmática 6 *Determinação da relação* parte_de

a. Se o antecedente $\mathcal{T}$ está no singular - singular($\mathcal{T}$), $\mathcal{A}$ é a entidade introduzida por um SND, $\mathcal{A}$ não é uma entidade coletiva (determinada pelo plural ou por estar presente num dicionário de coletivos), então pode-se assumir a relação parte_de($\mathcal{A}, \mathcal{T}$).

b. A relação parte_de($\mathcal{A}, \mathcal{T}$) somente será válida se não existir nada em contrário no contexto de interpretação : parte_de($\mathcal{A}, \mathcal{T}$) $\wedge \neg$anormal(parte_de($\mathcal{A}, \mathcal{T}$)).

A anormalidade da relação é quando existe informação que impossibilite que um objeto seja parte de outro. Por exemplo: caso o tamanho de $\mathcal{A}$ seja maior que o tamanho de $\mathcal{T}$.

3.3.5 A relação *subcategorizado_por*

A definição e obtenção da relação *subcategorizado_por* é semelhante à da relação *parte de* (seção 3.3.4) havendo uma única diferença crucial: o antecedente tem que ser uma entidade que possa exercer uma ação autônoma e *intencional* sobre outras – a denominada **animacidade** [Dahl e Fraurud 1996, Sidner 1979]. Considere o seguinte exemplo:

(3.33) a. Um ônibus chegou à rodoviária.

b. O motorista era calvo.

Onde considerando o SND *o motorista* introduzido na segunda frase o qual pode ter uma das suas ligações estabelecidas com *o ônibus* introduzido na primeira frase. De acordo com a definição de como é obtida a relação *parte de* (regra pragmática 6) é possível obter-se que "o motorista é parte do ônibus". Porém um motorista não é uma parte estrutural do ônibus, apesar de num determinado instante estar dentro deste. Por outro lado é a existência de um ônibus que permite ao emissor usar um SND representando aquele motorista específico. Visando solucionar esta situação propomos a relação *subcategorizado por* que deriva do fato das duas entidades poderem ser parte uma da outra, mas não o são porque uma delas é *animada.*

Assim a proposta para determinação da relação *subcategorizado por* é:

Regra Pragmática 7 *Determinação da relação* subcategorizado por

Se o antecedente $\mathcal{T}$ está no singular - $singular(\mathcal{T})$, $\mathcal{A}$ é um definido animado - $snd(\mathcal{A}) \wedge animado(\mathcal{A})$ e $\mathcal{T}$ não é um indivíduo coletivo - $\neg plural(\mathcal{T})$, pode-se assumir a relação $subcategorizado\_por(\mathcal{A}, \mathcal{T})$ entre $\mathcal{A}$ e $\mathcal{T}$.

3.3.6 A pseudo relação *acomodação*

A *acomodação* surge quando todas as outras possibilidades de interpretação de um SND terminaram e nenhuma das relações anteriores pôde ser estabelecida. Mesmo neste caso, algo tem que ser feito para que a interpretação do discurso continue [Levine, Guzmán e Klin 2000], pois um emissor não deseja transmitir discursos desconexos. Como conseqüência a entidade introduzida pelo SND, a qual não se configurou como sendo anafórica, deve então ser acomodada na representação semântica [Freitas e Lopes 1996, Spenader 2003, Cohen 2000], comportando-se de maneira semelhante a um indefinido [Heim 1982].

A proposta para a pseudo relação de acomodação fica:

Regra Pragmática 8 *Determinação da relação* acomodação

Se $\mathcal{A}$ é um SND e não é possível estabelecer nenhuma das relações anteriores (parte de, membro de, subcategorizado por) entre $\mathcal{A}$ e o antecedente $\mathcal{T}$, então assume-se a relação $acomodacao(\mathcal{A})$.

3.4 Implementação das regras pragmáticas

A interpretação de um SND é, na sua essência, um raciocínio sobre a referência *mental* da entidade introduzida pelo SND com alguma entidade conhecida pelo receptor [Ariel 1996, Strawson 1950, Donnellan 1966, Chafe 1996]. Esta tese, diferente da desses últimos autores, e alinhada com os trabalhos de referência anafórica [Grosz, Joshi e Weinstein 1995, Kamp e Reyle 1993, Carter 1987, Grosz, Joshi e Weinstein 1983, Sidner 1981], considera o termo *referência* como sendo *interno ao discurso*, i.e., somente as entidades introduzidas pela interpretação do discurso são consideradas como possíveis antecedentes. Qualquer referência externa é con-

siderada uma entidade do discurso [Heim 1982], sendo então acomodada [Spenader 2003, Cohen 2000] na sua interpretação.

Considere o seguinte exemplo:

(3.34)

a. Horácio tirou a cesta de piquenique do carro.

b. A cerveja estava quente.

Embora não esteja explícito, é necessário inferir que existe uma relação *parte de* entre a cerveja introduzida na frase (3.34a) e a cesta de piquenique na frase (3.34b). Este fenômeno, conhecido em inglês como *bridging [Clark 1977, Huang 1994],* indica que é necessário achar uma explicação para o uso de um SND pelo transmissor num contexto onde não existe uma relação direta entre duas entidades.

Os trabalhos anteriores sobre resolução de anáforas, apesar de citarem o fenômeno, não o tratam. As *Teoria do Foco* [Sidner 1979], *Teoria da Centragem* [Grosz, Joshi e Weinstein 1995] e a proposta de Carter [Carter 1987], consideram que a única relação possível entre o antecedente e a expressão anafórica é a relação de co-referência. Esta relação não é suficiente para tratar fenômenos como os do texto (3.34) onde a *cesta de piquenique* é o antedecente do SND *a cerveja,* mas a entidade *cerveja* não é uma *cesta de piquenique,* mas sim **parte desta.**

Quando essas propostas tentam avaliar este fenômenos, elas não consideram a influência do contexto no processo de resolução. No exemplo (3.34), não se é esperado que cestas de piquenique tenham cerveja, e mais ainda, um sistema automatizado não deveria gastar tempo tentando predizer (antecipamente), todas as partes de um dado objeto. Neste trabalho adota-se a idéia de que as entidades introduzidas no discurso por SNDs podem estar relacionadas com outras entidades do mesmo discurso e a língua fornece pistas de como pode ser este relacionamento. O resultado é que cerveja é contextualmente parte da cesta de piquenique.

Isto levou ao desenvolvimento da metodologia aqui apresentada na qual o processo de resolução é um processo abdutivo: dada a observação de que uma entidade foi introduzida por um SND, uma explicação seria que esta entidade está ancorada numa entidade anteriormente introduzida. Formalmente, esta explicação é dada pela inserção de uma nova condição DRS no formato da já conhecida fórmula $\mathcal{R}(\mathcal{A}, \mathcal{T})$.

3.4.1 Interpretação por abdução

Nesta tese adaptou-se a metodologia proposta de Hobbs et al. [Hobbs et al. 1993]: a interpretação pragmática de uma frase deve ser feita provando a fórmula lógica desta numa base de conhecimentos com os fatos do texto, recorrendo a raciocínio abdutivo. O esquema abdutivo proposto por Eshghi e Kowalski [Eshghi e Kowalski 1989] é utilizado na especificação das regras. Este esquema permite a escrita das regras em programação em lógica com negação por falha e é facilmente implementável.

A abdução é a formalização de um tipo de raciocínio de senso comum: raciocinar para explicar [Brewka, Dix e Konolige 1997]. Num exemplo clássico, quando se observa que a grama está molhada pela manhã, pode-se inferir que choveu à noite ou que o aspersor ficou ligado [Brewka, Dix e Konolige 1997, Kakas, Kowalski e Toni 1992]. A abdução é caracterizada pela regra de inferência 3.35. Não é seguro considerar "a" verdadeiro com esse tipo de inferência. Em outras palavras a abdução é não monotônica, pois se é notado que as ruas não estão molhadas, no exemplo da grama, não se deve assumir que incontestavelmente choveu [Brewka, Dix e Konolige 1997].

$$\begin{array}{r} b \leftarrow a \\ \underline{\quad\quad b} \\ a \end{array} \tag{3.35}$$

O esquema abdutivo utilizado é constituído por uma tupla (P, A, I), onde:

- P - é um programa em lógica estendido para utilizar negação por falha.
- A - é um conjunto de literais que podem ser abduzidos.

  O conjunto de literais que podem ser abduzidos é prefixado, o que pode reduzir o número de justificações possíveis.
- I - é um conjunto de restrições de integridade.

  As restrições de integridade são da forma:

  $\leftarrow L_0, ..., L_n$, com $L_i = l_i$ ou $L_i = not\, l_i$, um literal positivo ou a negação por falha de um literal positivo.

Só são justificações as soluções abdutíveis que verificam as restrições de integridade e as melhores soluções são as básicas e minimais.

Uma justificação é **básica** se nenhum dos fatos na justificação pode ser explicado pela teoria e é **minimal** se não existe nenhuma justificação que seja subconjunto desta. Assim, para justificar o fato p na teoria 3.36, a explicação $\Delta = \{r\}$ é básica e minimal; a explicação $\Delta = \{q\}$ é minimal, mas não é básica e a explicação $\Delta = \{q, r\}$ não é básica nem minimal.

$$\begin{aligned} p &\leftarrow q \\ p &\leftarrow q, r \\ q &\leftarrow r \end{aligned} \tag{3.36}$$

Em termos de resolução anafórica, implica dizer que a observação da utilização de um SND por parte do transmissor tem como explicação a identificação, por parte do receptor, de uma relação entre o SND e um possível antecedente. Esta explicação será minimal, pois existe um conjunto definido de possíveis relações sem que nenhuma seja subconjunto da outra e é básica, pois tais relações não podem ser inferidas diretamente pela informação presente no discurso.

Por fim, as regras pragmáticas constituem um conjunto de restrições de integridade que permite eliminar as justificativas menos plausíveis na interpretação de um SND.

O programa em lógica O programa em lógica P é o conjunto resultante da união das condições representadas no contexto $K_{(i-1)}$ com as condições resultantes da interpretação das entidades introduzidas em $K_i^{parcial}$:

$$P = U_{K_{(i-1)}} \cup U_{K_i^{parcial}}$$

O conjunto de abdutíveis Os elementos que podem ser abduzidos são as relações de ligação, as quais permitem explicar a razão da utilização de um SND.

membro_de(Ref,Ref2): a entidade denotada pelo referente Ref é membro do conjunto de entidades denotadas por $Ref2$. Respeitando a restrição de que tanto Ref quanto $Ref2$ são do mesmo tipo.

coref(Ref,Ref2): a entidade denotada pelo referente Ref co-referencia a entidade denotada pelo referente $Ref2$ desde que respeitadas as regras pragmáticas que limitam o estabelecimento da relação.

parte_de(Ref,Ref2): a entidade denotada pelo referente Ref é parte estrutural da entidade denotada pelo referente $Ref2$ desde que respeitadas as condições para que tal relação seja estabelecida.

subcategorizado_por(Ref,Ref2): a entidade denotada pelo referente Ref é parte conceitual da entidade denotada pelo referente $Ref2$.

As restrições de integridade São o conjunto de restrições usadas na verificação da consistência da base de dados resultante de uma inferência abdutiva. Em termos desta proposta, a base resultante é K_i, ou seja, a DRS resultante da interpretação da frase f_i no contexto $K_{(i-1)}$.

Inicialmente tem-se que garantir que não existam inconsistências simples, isto é feito através das seguintes regras:

$$\begin{aligned}
&\Leftarrow membro_de(Ref, Ref2),\ not\, gen_membro_de(Ref, Ref2).\\
&\Leftarrow coref(Ref, Ref2),\ not\, gen_coref(Ref, Ref2).\\
&\Leftarrow parte_de(Ref, Ref2),\ not\, gen_parte_de(Ref, Ref2).\\
&\Leftarrow subcategorizado_por(Ref, Ref2),\ not\, gen_subcategorizado_por(Ref, Ref2).
\end{aligned} \tag{3.37}$$

Ou seja, é inconsistente assumir que exista uma relação entre dois referentes Ref e $Ref2$ e ao mesmo tempo não haja condições para que esta relação exista "$gen\_...$". As condições genéricas para que uma relação exista são as fornecidas pelas regras pragmáticas apresentadas na seção 3.3, as quais são traduzidas para:

$$\begin{aligned}
gen_membro_de(Ref, Ref2) \Leftarrow\ & snd(Ref),\\
& tipo(Ref, T_A),\\
& tipo(Ref2, T_T),\\
& T_A \cap T_T \neq \{\},\\
& singular(Ref),\\
& plural(Ref2).
\end{aligned} \tag{3.38}$$

$$
\begin{aligned}
gen\_parte\_de(Ref, Ref2) \Leftarrow\ & snd(Ref), \\
& singular(Ref2), \\
& not\, plural(Ref), \\
& not\, anormal\_parte\_de(Ref, Ref2).
\end{aligned}
\tag{3.39}
$$

$$
anormal\_parte\_de(Ref, Ref2) \Leftarrow animado(Ref2) \tag{3.40}
$$

$$
anormal\_parte\_de(Ref, Ref2) \Leftarrow tamanho(Ref) > tamanho(Ref2). \tag{3.41}
$$

$$
gen\_coref(Ref, Ref2) \Leftarrow pronome(Ref). \tag{3.42}
$$

$$
gen\_coref(Ref, Ref2) \Leftarrow elipse(Ref).
$$

$$
\begin{aligned}
gen\_coref(Ref, Ref2) \Leftarrow\ & snd(Ref), \\
& numero(Ref, Nref), \\
& numero(Ref2, Nref2), \\
& Nref = Nref2, \\
& genero(Ref, Gref), \\
& genero(Ref2, Gref2), \\
& Gref = Gref2.
\end{aligned}
$$

$$
\begin{aligned}
gen\_coref(Ref, Ref2) \Leftarrow\ & snd(Ref), \\
& numero(Ref, Nref), \\
& numero(Ref2, Nref2), \\
& Nref = Nref2, \\
& plural(Ref).
\end{aligned}
$$

$$
\begin{aligned}
gen\_coref(Ref, Ref2) \Leftarrow\ & snd(Ref), \\
& numero(Ref, Nref), \\
& numero(Ref2, Nref2), \\
& Nref = Nref2, \\
& plural(Ref2).
\end{aligned}
$$

$$
\begin{aligned}
gen\_subcategorizado\_por(Ref, Ref2) \Leftarrow\ & snd(Ref), \\
& animado(Ref2), \\
& singular(Ref2), && (3.43)\\
& not\,plural(Ref). && (3.44)
\end{aligned}
$$

Outra restrição é que algumas relações não podem ser reflexivas:

$$
\begin{aligned}
&\Leftarrow membro\_de(Ref, Ref2), membro\_de(Ref2, Ref). \\
&\Leftarrow parte\_de(Ref, Ref2), parte\_de(Ref2, Ref). && (3.45)\\
&\Leftarrow subcategorizado\_por(Ref, Ref2), subcategorizado\_por(Ref2, Ref).
\end{aligned}
$$

Isto é: uma entidade não pode ser membro de um conjunto e o conjunto ser membro da entidade ou um objeto ser parte de outro objeto e vice-versa, ou ainda uma entidade subcategorizar a existência de outra entidade e vice-versa.

O mecanismo

Para ativar a máquina de abdução [Damásio, Nejdl e Pereira 1994, Damásio, Pereira e Schroeder 1996] é necessário inserir a contraprova da observação, assim para cada condição $snd(Ref)$ que esteja presente na interpretação fora de contexto, é introduzida uma cláusula $\tilde{}snd(Ref)$.

Considerando agora que cada $snd(Ref)$ é na verdade a observação da seguinte equação:

$$
\begin{aligned}
snd(\mathcal{A}) \Longleftarrow\ & existe(\mathcal{T}), \\
& \mathcal{R}(\mathcal{A}, \mathcal{T}).
\end{aligned}
\qquad (3.46)
$$

onde $\mathcal{R}$ é um dos abdutíveis (relação de ligação). Como $\mathcal{A}$ e $\mathcal{T}$ são conhecidos, logo se existir uma relação $\mathcal{R}$ (predicado aplicado a $\mathcal{A}$ e $\mathcal{T}$) esta seria o que falta para dizer que *snd* é uma conseqüência lógica da base de conhecimento atual.

Mais de um $\mathcal{R}$ podem ser identificados. É neste instante que as restrições de integridade atuam. Somente os valores de $\mathcal{R}$ que mantenham coerentes a união da base de conhecimento anterior com as restrições de integridade são aceitos e vão gerar modelos

válidos.

3.4.2 A Inferência das relações de ligação

Considere o discurso $D = f_1, f_2, ..., f_{i-1}, f_i, ..., f_n$. A interpretação da frase f_i e de seus SNDs é feita em duas etapas:

1. Primeiro, a frase f_i é transformada em uma DRS:
   - cada sintagma nominal indefinido de f_i introduz um referente x e um conjunto de condições descritivas aplicadas a este referente $\Theta(x)$,
   - cada SND de f_i introduz um referente y, a condição $snd(y)$ que marca o referente y para futura interpretação e o conjunto de condições descritivas aplicada ao referente introduzido $\Omega(y)$,
   - finalmente, cada verbo introduz uma condição ϵ aplicada sobre os referentes de seus argumentos. Como já destacado, não está sendo considerado, no caso dos verbos, o restante das condições introduzidas pela interpretação do tempo verbal, em especial a introdução de dois outros referentes: a eventualidade e o tempo associado [Rodrigues e Lopes 1994].

Em relação aos conjuntos $\Theta(x)$ e $\Omega(y)$ estes têm em comum os seguintes itens:

1. informação léxica: (1) se a entidade é singular ou plural, (2) masculino ou feminino, (3) aumentativo, normal ou diminutivo e (4) radical do substantivo composto.
2. informação sintática: se a entidade é o sujeito, objeto direto ou objeto indireto da frase.

O conjunto $\Theta(x)$ aplicado a um referente x no contexto de interpretação de f_i, pode ter mais duas condições não disjuntas:

1. $foco(x)$: indicando que o referente x é uma entidade saliente no contexto.
2. $ifoco(x)$: indicando que o referente x está saliente, de forma implícita, no contexto.

O resultado da interpretação fora de contexto é uma DRS parcial $K_i^{parcial}$. Tome o exemplo:

(3.47)

a. Samuel comprou um cão.

b. O animal late toda noite.

Considerando que a frase (3.47a) tenha sido interpretada:

$$K_{3.47a} = \left| \begin{array}{l} s, c \\ \hline samuel(s) \\ singular(s) \\ sujeito(s) \\ cão(c) \\ singular(c) \\ objeto(c) \\ comprar(s,c) \end{array} \right| \tag{3.48}$$

e a frase (3.47b) tenha a seguinte interpretação parcial:

$$K_{3.47b}^{parcial} = \left| \begin{array}{l} a \\ \hline animal(a) \\ singular(a) \\ sujeito(a) \\ snd(a) \\ latir\_toda\_a\_noite(a) \end{array} \right| \tag{3.49}$$

Durante o segundo passo, na interpretação em contexto, cada $K_i^{parcial}$ deve ser interpretada no contexto dado por K_{i-1} e todas as condições $snd(Ref)$ necessitam ser abdutivamente provadas. Dado um conjunto de k condições snd para referentes introduzidos em f_i: $\varsigma_i = \{snd(ref_i^0), snd(ref_i^1), ..., snd(ref_i^j), ..., snd(ref_i^k)\}$, deve-se provar que:

$$K_{i-1} \cup K_i^{parcial} \cup \Delta_i \models \varsigma_i \tag{3.50}$$

Onde Δ_i é o conjunto de condições que, quando acrescentadas à $K_{i-1} \cup K_i^{parcial}$, permite justificar por que algumas entidades foram introduzidas em f_i através de SNDs – ς_i. A princípio, qualquer inferência que permita relacionar os referentes de $snd(ref_i^j)$ com quaisquer referentes de K_{i-1} é uma prova. Isto é plausível em termos humanos: *tenta-se sempre uma ligação entre as coisas do dia-a-dia*! Porém, no âmbito deste trabalho, são

tratadas apenas as inferências *diretas* onde a justificativa é a existência de uma relação estrutural entre ref_i^j e seu antecedente em K_{i-1}. As relações utilizadas são: parte de, membro de, subcategorizado por e co-referência.

Se alguma das condições de ς_i não puder ser provada em $K_{i-1} \cup K_i^{parcial} \cup \Delta_i$ através da utilização da relações estruturais, então considera-se que apesar da entidade ter sido introduzida por um SND ela comporta-se como um indefinido [Kamp e Reyle 1993, Heim 1982]. Neste caso ela deve ser simplesmente *acomodada* no discurso [Spenader 2003, Cohen 2000, Sandt 1992]. Isto é assinalado através da pseudo relação $acomoda(Ref)$, que também faz parte de Δ_i.

Com a introdução da condição $acomoda(ref)$ fica claro que pode existir mais de um modelo que verifique (3.50). Isto é expresso em:

$$\forall M \forall x \forall y [modelo(M) \wedge snd(x) \wedge (snd(x) \in M) \wedge antec(x,y) \wedge (antec(x,y) \in M) \rightarrow$$
$$(\exists! R\,(\,R \in M) \wedge R(x,y)) \oplus (acomoda(x) \wedge acomoda(x) \in M)]$$

onde $R \in \{coreferencia, membro\_de, parte\_de, subcategorado\_por\}$. Isto é, para cada modelo M existe somente uma única relação R entre a entidade introduzida pelo SND: $snd(x)$ e seu antecedente: $antec(x,y)$. Caso não exista tal relação, então *necessariamente* a entidade x deve ser acomodada em M.

Note que existe um modelo M_0 onde todas as condições em ς_i são acomodações: caso em que apesar de *todas* entidades terem sido introduzidas por SNDs, elas foram interpretadas como indefinidos. Isto contradiz a motivação que leva um transmissor a utilizar um SND: a suposição de que o receptor já conhece a entidade que pode ser referenciada através do uso de um termo *supostamente* anafórico. Como conseqüência, este modelo deve ser preterido em favor dos outros.

Seguindo esta linha de raciocínio, foi estabelecido um critério de escolha dos modelos válidos. O critério adotado foi: **quanto mais informativo for a representação semântica final, melhor**. Isto pode ser obtido comparando-se o número de relações em Δ_i para um dado modelo M. Aqueles com maior número de relações são os melhores, tal que:

$$K_{i-1} \cup K_i^{parcial} \cup \Delta_i \models_{M_0} \varsigma_i \tag{3.51}$$

onde $\forall y \,|\, snd(y) \in \varsigma_i, acomoda(y) \in \Delta_i$. Em M_0 todos os SNDs de f_i foram acomodados.

Nos demais modelos M existe uma única relação entre um referente y de uma condição $snd(y)$ e seu antecedente.

Depois da interpretação de $K_i^{parcial}$ no contexto K_{i-1}, a DRS resultante K_i será: $K_i =< U_{i-1} \cup U_i^{parcial}, Cond_{i-1} \cup Cond_i^{parcial} \cup \Delta_i >$

3.5 Avaliação das regras pragmáticas

Na avaliação do mecanismo aqui proposto, usou-se o corpus marcado do CETEN-Folha [Paulo 2002]: um corpus baseado no jornal "A Folha de São Paulo", contendo 1.597.807 frases e 25.475.272 palavras. Na avaliação foram considerados apenas os indivíduos, os quais foram identificados da seguinte forma: localizou-se todas as frases (*tags* <s>...</s>), e dentro de cada frase localizou-se os artigos definidos e indefinidos (*tags* <artd> e <arti>) e considerou-se que a entidade são as próxima três *tags*.

A identificação dos antecedentes foi feita usando a implementação, descrita no capítulo 5, para o algoritmo de identificação de antecedente (seção 4.3.6). Os textos marcados são traduzidos de frases marcadas para a representação semântica que é a entrada do algoritmo.

Foram feitos dois experimentos. No primeiro foi utilizado o processo automatizado, de acordo com a metodologia abdutiva utilizada aplicada sobre o corpus. Num segundo experimento, foi feito um teste menor com apenas algumas frases e foi utilizado como comparação o experimento feito com um testador humano.

O resultado do primeiro teste é apresentado na tabela 3:

Relação de ligação	Processo automatizado
co-referência	936.070
membro de	125.262
parte de	630.159
subcategorizado por	477.341
acomodação	1.548.124

Tabela 3: Resultado do teste automatizado.

As conclusões tiradas dos resultados foram:

1. Foi identificado que existe uma predominância da acomodação (45.15%), indicando que apesar de uma entidade ter sido introduzida por um SND, ela não é anafó-

rica. Isto não foi surpresa, pois Vieira e Poesio já haviam observado tal comportamento [Vieira e Poesio 2000].

2. O restante dos casos: 2.168.832, não correspondem ao percentual restante (54.85%). Isto ocorre porque existe ambigüidade na determinação de algumas relações, em especial entre *subcategorizado_por* e *parte_de*, onde não se pode determinar a animacidade ou não de uma determinada entidade.

3. A co-referência também ocorreu num número elevado de casos, levando à conclusão de que o transmissor apenas usou um SND para não repetir o mesmo termo. Vale salientar que no teste não foram considerados os pronomes.

4. A relação entre conjuntos (*membro_de)* é pouco freqüente em textos (pelo menos jornalísticos).

No segundo experimento foi utilizado um experimentador humano e comparado com os resultados obtidos pelo processo automatizado. Foram analisados 54 extratos de textos, num total de 236 frases, 520 SNDs e 46 indefinidos. O resultado está na tabela 4:

Relação de ligação	Exp. humano	Processo automatizado
co-referência	123	93
membro de	21	15
parte de	28	87
subcategorizado por	85	71
acomodação	190	271
nenhuma das anteriores	73	-

Tabela 4: Resultado com o experimentador humano.

As conclusões tiradas da tabela 4 são:

1. As relações *parte_de* e *subcategorizado_por* são melhor definidas pelo experimentador e por esta razão houve uma discrepância entre os resultados obtidos no processo automatizado. Estes valores podem ser melhorados caso um dicionário completo (digitalizado) de coletivos fosse utilizado.

2. Houve alguns casos (73 correspondendo a 14.04% dos SNDs) cuja relação não pôde ser identificada pelo experimentador humano como sendo as que são tratadas nesta tese. Foram casos em que a anáfora era uma nominalização de eventos, por exemplo: "O avião do presidente aterrissou às 17:15h... A chegada foi". Note que neste caso

é o evento que é referenciado pelo SND **a chegada.** Então nenhuma relação foi considerada. Por outro lado, casos como este foram classificados como acomodação pelo programa.

3. Houve menos acomodações do experimentador humano do que no processo automatizado. A razão é clara: o experimentador humano utilizou-se de muito do seu conhecimento para estabelecer uma relação, enquanto o processo automatizado usou pouco conhecimento. Um exemplo é o SND "a chegada" citado anteriormente.

4. O experimentador humano detectou mais relações de co-referência. A razão é a mesma do número de acomodações: como o experimentador *conhece* um maior número de relações e sinônimos entre palavras, ele acabou por detectar co-referências onde o programa detectou acomodações.

A conclusão que se chegou com estes resultados é que apesar das discrepâncias com o resultado do experimentador humano, o sistema automatizado consegue estabelecer relações com um grau relativamente bom de precisão. Ainda mais caso fosse incorporado um dicionário de coletivos e informações sobre a animacidade das entidades, o que permitiria uma melhor separação das relações *parte_de* e *subcategorizado_por*.

4 *Estrutura Nominal do Discurso*

"*Só sei uma coisa: que nada sei.*"
Sócrates

Neste capítulo é apresentada a metodologia de obtenção do antecedente T da fórmula $R(T, A)$. Para tal, é criada a *Estrutura Nominal do Discurso* que permite ao sistema de interpretação de anáforas restringir o número de antecedentes T possíveis para uma expressão anafórica A. Para apresentar a forma como tal estrutura deve ser construída é feito um estudo sobre as características necessárias ao processo de estruturação do discurso. A seguir é definido o conceito de *centro de atenção ou foco*, em especial são criados o *foco explícito* e *foco implícito,* elementos centrais para a criação da Estrutura Nominal do Discurso (END) defendida nesta tese. Por fim é apresentada a metodologia para criação e manutenção da END.

4.1 Introdução

Relembrando a fórmula $\mathcal{R}(\mathcal{A}, \mathcal{T})$, considere que $\mathcal{R}$ (a relação) é um elemento fixo, sinalizando que $\mathcal{A}$ está ligado a $\mathcal{T}$ (o antecedente). Nesta hipótese simplificadora, a resolução de anáforas seria um processo em que apenas $\mathcal{T}$ deve ser determinado. Os problemas que teriam que ser resolvidos nesta determinação são explicitados a seguir:

1. Freqüentemente existe mais de um candidato para $\mathcal{T}$. O processo de interpretação tem então duas alternativas: (a) escolher um candidato e prosseguir com a interpretação ou (b) considerar uma interpretação para cada um dos n candidatos e prosseguir com n interpretações. A primeira hipótese tem a vantagem da velocidade de resolução, porém pode acontecer que informações introduzidas posteriormente (pressuposto de não monotonicidade) invalidem a solução anterior, obrigando a um *reprocessamento da informação já interpretada*. Na segunda hipótese, todo o processamento das alternativas possíveis já foi feito. Assim, já não é necessário o reprocessamento, mas sim a busca por uma interpretação alternativa (já pronta). A desvantagem desta segunda hipótese é que a interpretação do discurso é o produto cruzado das interpretações possíveis para cada frase, o que torna o processamento oneroso. O ideal seria uma metodologia que utilizasse o melhor de cada uma destas hipóteses.

2. $\mathcal{T}$ pode estar em qualquer frase do discurso. A conseqüência para a interpretação é que, à medida que o discurso vai sendo interpretado, é maior o número de possíveis antecedentes $\mathcal{T}$, por um dado $\mathcal{A}$ e é maior o esforço computacional do processo de resolução como um todo. A solução encontrada na literatura consiste em limitar o espaço de busca a um determinado número m de frases anteriores. Como encontrar o valor ideal para m? A escolha de um valor pequeno pode impossibilitar a escolha de um antecedente $\mathcal{T}$ que esteja num frase anterior à frase m. A escolha de um valor grande torna o processamento oneroso.

3. Finalmente, a limitação do número de frases conjuntamente com a consideração de que as entidades nelas introduzidas constituem apenas um conjunto de simples escolhas reduz a contribuição semântica destas mesmas entidades para a interpretação do discurso como um todo [Freitas e Lopes 1996]. Cada frase (e suas entidades) traz uma contribuição semântica tanto para a sua própria interpretação (fora de contexto) quanto para a estruturação do conhecimento disperso em cada frase do discurso. Considerar a contribuição da frase para a estruturação do discurso permite

ao processo de interpretação inserir restrições *naturais*[1] ao processo de escolha de $\mathcal{T}$, podendo então aumentar o número de frases consideradas. A Teoria da Centragem [Grosz, Joshi e Weinstein 1995] utiliza esta abordagem considerando a contribuição de cada entidade para o acompanhamento da movimentação do centro de atenção (foco) à medida que o discurso avança e como resultado deste acompanhamento são geradas restrições para a escolha de um antecedente. O que a Teoria da Centragem não considera é que a informação sobre a movimentação do foco não só gera restrições imediatas no processo de resolução de anáforas (e.g. nas próximas duas frases) como também pode gerar restrições estruturais sobre a interpretação de qualquer entidade do discurso [Lopes e Freitas 1994, Freitas e Lopes 1994].

Objetivando criar uma metodologia que leve em consideração todos estes itens, este capítulo propõe a *Estrutura Nominal do Discurso* (**END**). Esta estrutura permite: (1) restringir o espaço de busca por antecedentes sem contudo limitar o número de frases e (2) criar um semiprocessamento de interpretações, i.e. intermediário entre uma interpretação completa e um reprocessamento, de forma a agilizar uma reinterpretação. Esta estrutura permite assim explicitar a movimentação dos focos durante todo o discurso.

A END é uma árvore onde cada folha representa o conteúdo semântico (DRS) de uma determinada frase do discurso e cada nó interno representa o conteúdo semântico resultante do acompanhamento das entidades mais salientes (focos) de seus filhos. Uma propriedade importante desta árvore é que somente os nós mais à direita estão abertos para interpretação [Polanyi, Berg e Ahn 2003, Polanyi 1988]. Uma forma esquemática desta árvore pode ser vista na figura 3:

[1] Impostas pelo emissor e codificadas no discurso.

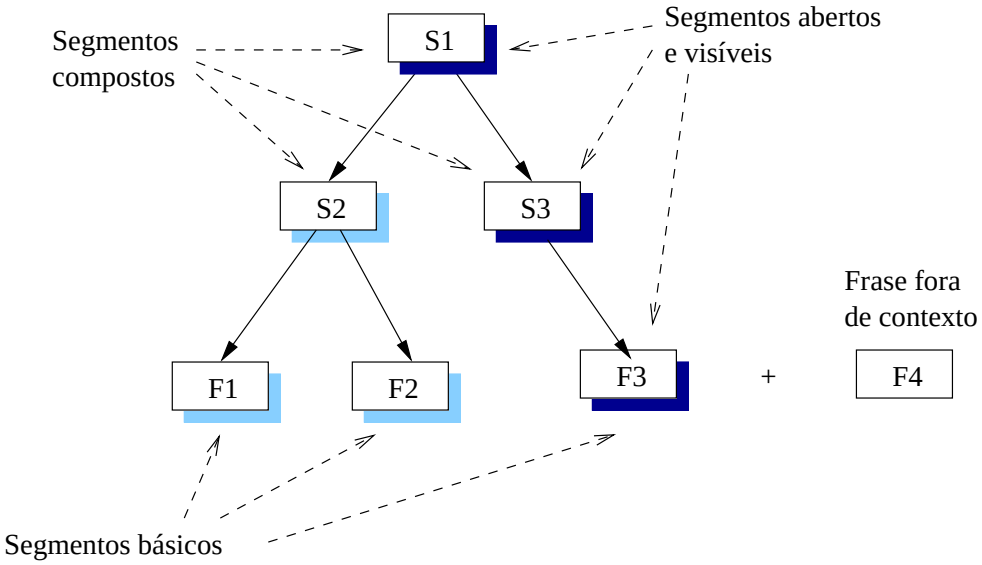


Figura 3: Esquematização da Estrutura Nominal do Discurso

onde: F1, F2, F3 e F4 são frases, S1 e S2 são nós internos e F4 é a frase a ser interpretada em relação aos nós visíveis S1 e F3.

Por fim, foi necessário definir dois tipos de entidades salientes: um **foco explícito** e um **foco implícito**.

O foco explícito é resultante da utilização de anáforas pronominais, elipses e ANDs diretas (relação de co-referência), tal como em: **"O João** partiu a perna. **Ele** ficou sete dias de cama". Onde o pronome pessoal *ele* referencia o nome próprio *João*.

O foco implícito é resultante da utilização de conhecimento subjacente ao discurso, tal como acontece no exemplo 4.1. Neste exemplo, ao contrário do exemplo anterior, há necessidade de encontrar também a relação $\mathcal{R}$. Há que saber que há motoristas de ônibus, como de táxis, de caminhões, que há portas de ônibus, de táxis, de caminhões, de casas etc. Mas que não há portas (no plural) de motorista.

(4.1) a. Um ônibus acabou de chegar.

b. O motorista abriu as portas.

c. Os passageiros desceram pela porta de trás.

onde o sintagma nominal definido "os passageiros" na frase (4.1c) tem como antecedente *implícito*[2] o ônibus introduzido na frase (4.1a) (a resolução deve descartar as entidades

[2] A entidade que continua a ser referenciada de forma indireta.

introduzidas na frase (4.1b)), com a ligação entre o ônibus e os passageiros não sendo uma relação de co-referência direta, mas sim uma relação em que os passageiros *são parte do* ônibus[3].

Este capítulo apresenta a construção Estrutura Nominal do Discurso, estando assim estruturado: na seção 4.2 é feito um estudo sobre as propostas para a estruturação do discurso existentes na literatura, na seção 4.3 são apresentados os fundamentos de uma nova abordagem para o foco do discurso. O foco do discurso é o elemento central da teoria para estruturação do discurso proposta na seção 4.4.

4.2 Estrutura do Discurso

Existe concordância nas áreas da Lingüística Computacional [Sibun 1993, Grosz e Sidner 1986, Sidner 1979], Filosofia da Linguagem [Polanyi e Berg 1996, Mann e Thompson 1987] e Inteligência Artificial [Hobbs 1993, Hobbs 1985] de que um agente cooperativo[4], ao produzir um discurso, fá-lo de maneira planejada e organizada, reduzindo o esforço de interpretação por parte de seu interlocutor – o receptor. Esta forma organizada de transmissão é expressa sob a forma de uma estrutura que na maioria das vezes está implícita no discurso, a denominada *Estrutura do Discurso*.

A estrutura é fundamental para a compreensão do discurso pois organiza a informação transmitida, auxiliando sua interpretação por parte do receptor. O processo de estruturação do discurso está diretamente relacionado com a comunicação entre o transmissor e o receptor [Freitas e Lopes 1994, Abbott 1993], sendo que sua eficiência pode ser medida pela rapidez com que o receptor recupera as interpretações possíveis para um dado trecho do discurso [Blutner 2000]. Em termos computacionais isto equivale a um menor tempo de processamento e em termos lógicos a um menor número de inferências sobre o menor número possível de modelos (os modelos aqui citados são construídos sobre as DRSs).

O que, porém, não é consenso é a forma de representar a estrutura do discurso. Esta discordância tem suas origens na falta de uma resposta largamente aceita para a seguinte questão: "tem o discurso uma *estrutura genérica* ?", isto é, uma estrutura única que organize todo o conjunto de informações transmitidas (e.g. frases, gestos, intenções, proposições, conceitos, entidades, tempo, espaço etc).

A posição desta tese sobre o assunto é que, pelo menos em termos formais e

[3]Pelo menos quando os passageiros estão dentro do ônibus.

[4]Um agente que não tem a intenção de enganar transmitindo, deliberadamente, informações incorretas.

computacionais, a estrutura genérica ainda não existe. O que já é possível obter, computacionalmente, é um conjunto de estruturas específicas para o tratamento de certos fenômenos do discurso: anáforas [Huang 2000, Freitas e Lopes 1994, Lopes e Freitas 1994], elipses e informação temporal [Rodrigues e Lopes 1995, Rodrigues e Lopes 1993, Rodrigues e Lopes 1992]. Também para o agrupamento de entidades [Polanyi, Berg e Ahn 2003, Polanyi e Berg 1996] e representação de intenções, crenças e planos [Grosz e Sidner 1998, Grosz e Sidner 1990, Grosz e Sidner 1986]. Provavelmente, a estrutura genérica do discurso vai ser um *metanível de raciocínio* que promova a interação destas estruturas específicas. Neste contexto, esta tese apresenta na seção 4.4 a proposta de uma *estrutura específica* para o acompanhamento do centro de atenção de entidades de um discurso. Esta estrutura é utilizada para a resolução de anáforas (e de elipses), mais especificamente na determinação do antecedente $\mathcal{T}$ da equação $\mathcal{R}(\mathcal{A}, \mathcal{T})$.

4.2.1 Características de uma Estrutura do Discurso

As principais características que uma metodologia para estruturação do discurso deve levar em conta são:

Unidades básicas da estrutura: Considerando a estrutura com sendo um conjunto organizado de elementos básicos, estes podem ser: frases, sintagmas, interjeições ou parágrafos.

Forma de representação das unidades básicas: As unidades básicas, dependendo da teoria de estruturação adotada, podem ser representadas como: (1) forma lingüística pura (*surface structure*) [Passonneau e Litman 1997, Mann e Thompson 1987], (2) árvore de derivação sintática [Reinhart 1981, Reinhart 1976] ou (3) conteúdo proposicional ou semântico [Kamp e Reyle 1993, Groenendijk e Stokhof 1991].

Forma de representação da estrutura: Considerando o conjunto de unidades básicas, o próximo passo para caracterizar uma estrutura é a definição da forma estrutural com que estas unidades são ligadas umas às outras. As opções existentes são: seqüências [Sidner 1981], grafos [Hobbs 1985], pilhas [Grosz e Sidner 1986], árvores [Polanyi, Berg e Ahn 2003, Rodrigues 1995, Polanyi e Berg 1996, Rodrigues e Lopes 1992] ou uma combinação delas [Grosz e Sidner 1986].

Relações entre unidades: A existência de relações entre as unidades básicas independentemente da forma de representação da estrutura. As relações

variam de uma simples coordenação/subordinação [Polanyi, Berg e Ahn 2003, Polanyi 1988] de unidades até relações mais complexas tais como as relações retóricas [Mann e Thompson 1987].

Herança entre unidades básicas: Como resultado da forma de representação adotada, das relações utilizadas na estrutura e dos atributos internos das unidades básicas, é possível estabelecer critérios para a herança de atributos entre as unidades. Esta herança terá um impacto na interpretação do discurso [Lopes e Freitas 1994, Rodrigues 1995, Rodrigues e Lopes 1994].

4.2.1.1 Unidades Básicas da Estrutura

Ao considerar a existência de uma estrutura do discurso, surge naturalmente a pergunta: *qual é a menor unidade sobre a qual a estrutura deve ser construída?* A escolha da unidade básica e de sua forma de representação constitui um passo importante para a construção da estrutura, pois tem repercussões diretas sobre esta, nomeadamente:

1. na determinação das relações entre unidades básicas: caso sejam demasiadamente grandes, fica difícil encontrar ligações entre as diversas unidades básicas, impedindo que um maior número de relações possam ser estabelecidas e contribuindo para que a representação fique pobre de informação,

2. na definição da herança de atributos entre unidades: certas informações das unidades básicas devem ser herdadas pelas unidades básicas subseqüentes, permitindo assim uma interpretação mais coesa do discurso, por exemplo: *"O João foi ao supermercado. O açougue estava fechado".* Neste caso existe uma herança da entidade "supermercado" da primeira frase na segunda frase, permitindo que a entidade açougue tenha um contexto. Sem a herança de *supermercado,* o *açougue* poderia ser qualquer outro.

3. na caracterização das operações de (re)construção da estrutura, as quais permitem a obtenção de interpretações equivalentes sem a necessidade do processo de reinterpretação do discurso.

Os tipos de unidades básicas encontradas na literatura são:

Frases: são as unidades básicas mais simples e intuitivas [Polanyi, Berg e Ahn 2003, Komagata 2003, Mann e Thompson 1987, Grosz e Sidner 1986]. Podem ser grama-

ticalmente bem formadas ou não, e mesmo incompletas (i.e. não terminadas, interrompidas). Cada frase possui informações léxicas, sintáticas, semânticas e mesmo pragmáticas (minimamente contextuais), que lhe permite estabelecer um conteúdo informativo quase autosuficiente. Só não é totalmente autosuficiente, em virtude da interpretação de fenômenos do discurso que envolvem, em muitos casos, mais de uma frase. Por exemplo: anáforas (interfrases), elipses e tempo verbal.

Sintagmas: Outra proposta seria a utilização dos sintagmas constituintes da frase como elementos básicos [Reinhart 1983, Reinhart 1981, Langacker 1966]. Apesar desta proposta ter como atrativo a minimização da quantidade de informação no elemento básico (uma frase possui diversos fenômenos que devem ser tratados, e.g., anáforas intrafrases), ela apresenta uma grande desvantagem: o aumento da carga de interpretação para as unidades constituintes "imediatamente superiores", que no caso são as frases, sem um reflexo sensível no processo de interpretação do discurso como um todo.

Interjeições e palavras-marcas: apesar de não possuírem uma estrutura interna e se limitarem, quando muito, a carregarem informação léxica e sintática, aparentemente sem nenhum conteúdo semântico, as interjeições e as palavras-marcas desempenham um papel importante no processo de segmentação do discurso, atuando como operadores [Polanyi, Berg e Ahn 2003, Jr. e Duffy 2001, Grosz e Sidner 1986] que delimitam a abertura e o fecho dos segmentos considerados.

Alguns exemplos de interjeições são: Ops, Uhh, Uff, entre outros. Alguns exemplos de palavras-marcas são: porém, entretanto, embora etc.

Apesar da influência destes operadores frente à construção de uma *estrutura genérica do discurso*, o presente trabalho não considera sua utilização na construção da estrutura.

Parágrafos: no caso de textos escritos, cada parágrafo representa um determinado assunto ou subassunto que, em geral, são autocontidos em termos de informações, participando de forma direta na "composição" do assunto global do discurso, gerando uma estrutura de assuntos e subassuntos que muito se assemelha à estrutura genérica do discurso. O parágrafo deve ser considerado como o limite do processo de interpretação, onde levando-se adiante um determinado número de interpretações válidas, seria o parágrafo o limite no qual uma das interpretações válidas deve prevalecer [Rodrigues 1995].

Esta tese considera a frase como sendo a unidade básica formadora da estrutura, assumindo que estas são sempre bem construídas e completas.

Nas teorias que adotam a frase como elemento básico, estas são representadas das seguintes formas:

1. integral: quando não se recorre a nenhuma forma de representação, por exemplo, semântica. As frases são consideradas unas. Este tipo de "representação" é adotado por modelos tais como a RST [Mann e Thompson 1987].

2. árvore sintática: o segmento básico é constituído pela árvore de representação sintática da frase.

3. conteúdo proposicional: a interpretação léxica e sintática dá origem a uma representação semântica das entidades expressas pelos sintagmas nominais, das eventualidades e tempos expressos pelos verbos e entidades temporais, e das relações explícitas e implícitas entre estas entidades (e.g. DRT [Kamp e Reyle 1993]).

Nesta tese o processo de interpretação fora de contexto de uma frase isolada dá origem a uma representação semântica, semelhante à DRS, aqui denominada **segmento básico**. A interpretação do segmento básico frente à estrutura prévia dá origem a um bloco único de informações denominado **segmento**. O segmento nada mais é do que uma representação composta e/ou resumida da informação representada nos segmentos básicos que o compõem (seção 4.4.1).

4.2.1.2 Forma de representação das unidades básicas

As unidades básicas, dependendo da teoria de estruturação adotada, podem ser representadas como: (1) forma lingüística pura (*surface structure*) [Passonneau e Litman 1997, Mann e Thompson 1987], (2) árvore de derivação sintática [Reinhart 1981, Reinhart 1976] ou (3) conteúdo proposicional ou semântico [Kamp e Reyle 1993, Groenendijk e Stokhof 1991].

A representação na forma lingüística pura é feita através da utilização das frases como elemento básico. Cada frase é um elemento *de representação* que deve ser ligado diretamente a outras frases. A utilização deste tipo de representação é mais propícia para processos de avaliação feitos por agentes humanos, daí a sua utilização em teorias tais como a RST [Mann e Thompson 1987]. Computacionalmente não é viável a utilização da

forma lingüística pura, pois os mecanismos utilizados por um ser humano na avaliação de uma frase estão bem além dos processos de interpretação realizados por uma máquina. Logo é necessário antes *desmembrar* os componentes de uma frase para que cada peça (i.e. palavras, sintagmas, análise léxica, fenômenos lingüísticos etc) possam ser antes avaliados individualmente e depois compostos numa interpretação mais abrangente. Esta representação com base na forma lingüística completa não será utilizada nesta tese.

A representação na forma de árvore de derivação sintática proporciona uma análise em maior profundidade não só dos componentes internos a uma frase quanto do relacionamento destes (e.g. anáforas intrasentenciais [Reinhart 1981, Reinhart 1976], ambigüidades estruturais [Carter 1987] etc). Porém quando se considera a interpretação do discurso como um todo verifica-se que o processo de análise de duas estruturas sintáticas em frases distintas é complexo, a não ser para fenômenos localizados tais como as elipses com paralelismo [Kehler 1993, Hahn, Markert e Strube 1996].

Recorrendo a uma representação semântica, os elementos internos são denominados *atributos.* Por exemplo, caso a forma de representação seja uma DRS [Kamp e Reyle 1993], cada referente do discurso existente no universo da DRS é considerado um atributo (interno) da unidade básica. Esta representação é mais adequada para a interpretação do discurso: a representação semântica proporciona uma independência da sintaxe utilizada, possibilitando a análise do discurso de acordo com as informações contidas em cada frase deste. Por fim, as informações léxicas e sintáticas podem também ser representadas de forma semântica, impedindo que estas sejam perdidas na representação final.

Esta tese, como já visto no capítulo 3, utiliza uma representação semântica para cada frase e para o discurso como um todo.

4.2.1.3 Forma de representação da estrutura

A interpretação do discurso envolve mais do que a simples soma das interpretações parciais de cada frase. É necessário interligar, completar, corrigir e por fim unir as interpretações parciais de modo a obter uma interpretação para o discurso como um todo.

Sendo a estrutura do discurso uma das fases desta representação, a escolha da forma da estrutura[5] vai ter necessariamente conseqüências diretas no processo de interpretação:

1. na maneira com que as novas frases são interpretadas relativamente ao discurso

[5] Que é, pelo menos hipoteticamente, um reflexo da forma com que o emissor estruturou a idéia.

anterior e

2. na maneira com que as informações de uma nova frase são acrescentadas ao discurso previamente interpretado.

A seguir são apresentadas algumas formas de representação da estrutura do discurso:

seqüências: cada *segmento* tem um conjunto de informações que não depende de um segmento anterior. Este tipo de estrutura somente pode ser obtido quando se olha para o discurso "em alto nível", isto é, em grandes blocos de assuntos. Por exemplo: no primeiro capítulo desta tese é apresentada uma introdução, no segundo capítulo os trabalhos relacionados, no terceiro capítulo a representação semântica e assim por diante (fig. 4).

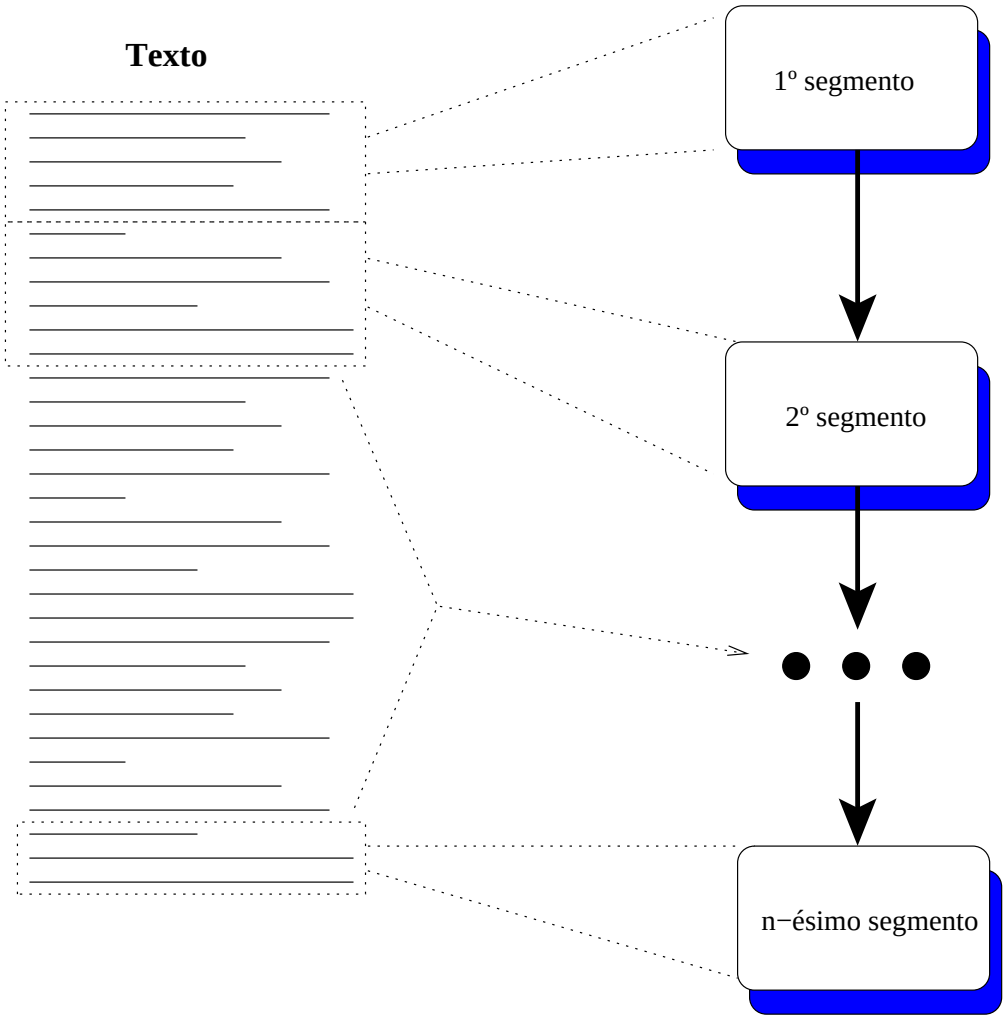


Figura 4: Representação seqüencial da estrutura.

grafos: cada segmento pode estar relacionado com qualquer bloco anterior em qualquer ordem [Hobbs 1985, Hobbs 1979], permitindo que segmentos menores, por exemplo parágrafos e frases, possam ter informações compartilhadas ou correlacionadas com segmentos previamente interpretados. Apesar desta ser a representação mais próxima da realidade, computacionalmente ela peca pelo excesso de testes e por não oferecer restrições à escolha de qual dos segmentos anteriores deve-se relacionar o último segmento interpretado. Como resultado, a cada segmento básico interpretado tem-se uma quantidade considerável de interpretações possíveis, tornando o processo computacional bem oneroso a cada nova frase (fig 5).

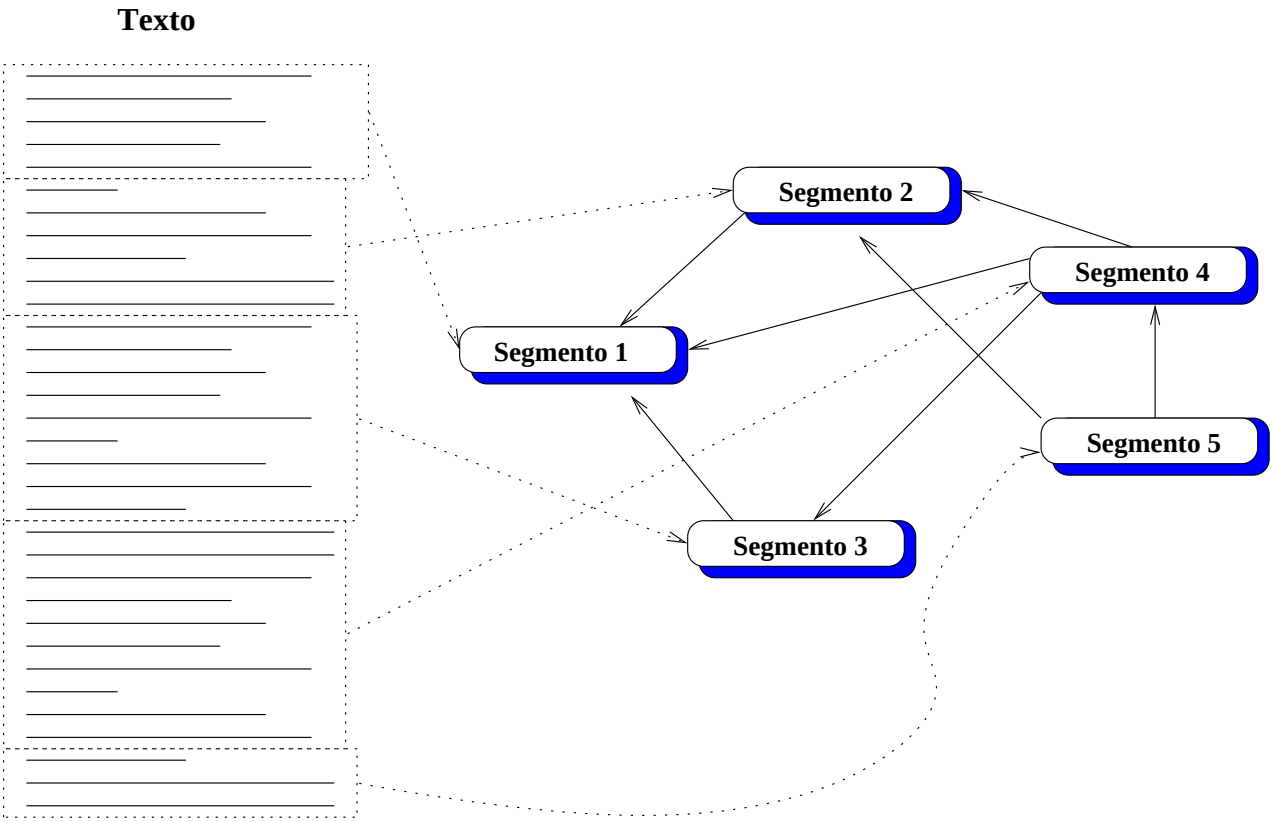


Figura 5: Representação da estrutura em grafo.

pilhas: uma simplificação da estrutura em grafo. Aqui os segmentos maiores podem englobar outros segmentos menores e mais específicos sobre o assunto em pauta, deixando explícita a idéia de que os segmentos maiores estão não só relacionados, como também dependem da informação dos blocos menores. Um segmento somente pode ser desempilhado quando todos os seus "subsegmentos" já foram desempilhados (fig. 6). As desvantagens desta metodologia são:

1. não existe um histórico da estrutura sobre os segmentos empilhados e desempilhados que permita uma revisão da interpretação feita e

2. impõe restrições fortes à interpretação pois a única relação possível é de subordinação de blocos, não sendo possível, por exemplo, concluir-se que dois blocos contenham informações complementares sobre um terceiro que os coordena.

   Esta estrutura em pilha foi apresentada como sendo o "estado atencional" da "estrutura lingüística" [Grosz e Sidner 1986] e tem uma dependência intrínseca da estrutura intencional (propósitos de cada bloco). As restrições impostas pela pilha são muito rígidas, inviabilizando qualquer interpretação relativa a um bloco que já tenha sido desempilhado. A "cache" de Walker [Walker 1996] é uma extensão à proposta original que tenta contornar algumas das limitações de uma pilha.

   Finalmente, é também encontrado o mesmo tipo de estruturação expresso nas pilhas de entidades (AFS e DFS) propostas por Sidner na Teoria do Foco [Sidner 1981, Sidner 1979].

árvores: corresponde a uma estrutura intermediária entre a estrutura em pilha e a estrutura em grafo. Esta estrutura permite tirar partido *de dois mundos*: as restrições impostas pela estrutura em pilha e o histórico gerado pela estrutura em grafo (figura 7).

Um exemplo desta utilização é a estrutura apresentada por Polanyi [Polanyi, Berg e Ahn 2003, Polanyi 1988], que permite relações de subordinação e coordenação entre os segmentos constituintes. Outro fato interessante são as restrições à interpretação que a própria estrutura pode apresentar, permitindo considerar, por exemplo, que somente os nós mais à direita da árvore estão *abertos* para a interpretação de novas frases (figura 8).

Desta forma o processo de interpretação de um novo segmento relativamente aos segmentos previamente introduzidos é simplesmente um processo de encontrar um dos segmentos mais à direita na árvore, o qual possa servir de referência para interpretação da nova frase.

Esta tese utiliza uma estrutura em árvore motivada pelo trabalho de Polanyi [Polanyi, Berg e Ahn 2003, Polanyi 1988].

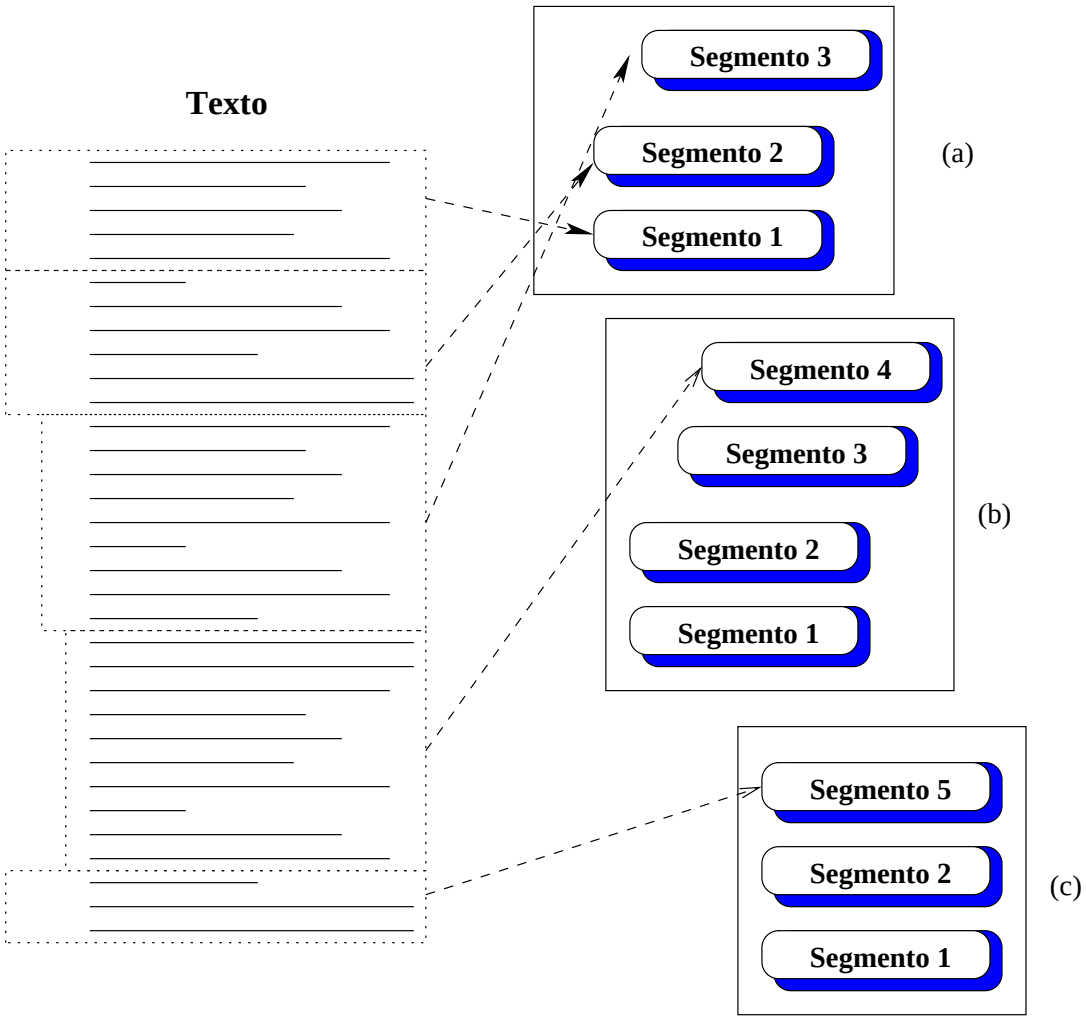


Figura 6: Representação da estrutura em pilha.

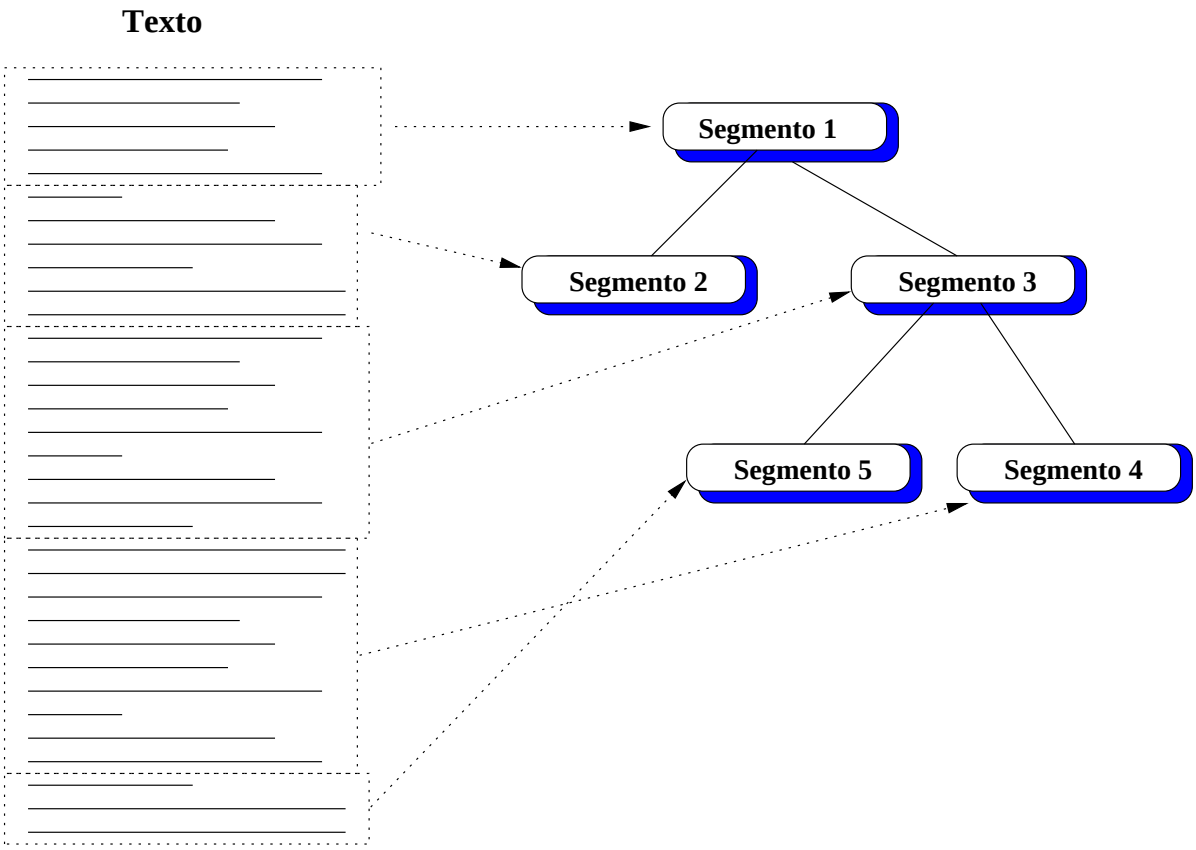


Figura 7: Representação da estrutura em árvore.

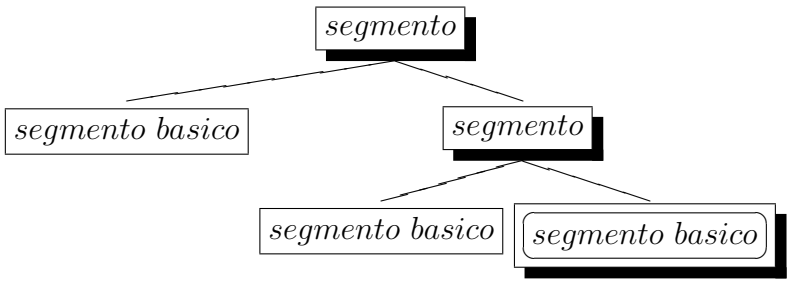


Figura 8: Árvore com os nós mais à direita abertos

4.2.1.4 Herança entre unidades básicas

Como resultado da forma de representação adotada, das relações utilizadas na estrutura e considerando os atributos internos das unidades básicas, é possível estabelecer critérios para a herança de atributos entre as unidades. Esta herança, como será visto em detalhes na Estrutura Nominal do Discurso (seção 4.4), terá um forte impacto na interpretação do discurso [Lopes e Freitas 1994, Rodrigues 1995, Rodrigues e Lopes 1994].

A herança sofre influência de operações sobre a estrutura tais como: agrupamento de unidades básicas, inserção de novas unidades, remoção de unidades, reorganização da estrutura e equivalência entre estruturas [Rodrigues 1995, Rodrigues e Lopes 1994, Rodrigues e Lopes 1995].

Na proposta de estruturação deste trabalho é considerada uma árvore cujos nós mais à direita estão abertos para a interpretação das novas frases.

4.2.2 Propostas de Estrutura do Discurso

A seguir são apresentadas as principais propostas para estruturação do discurso.

4.2.2.1 A proposta de Grosz e Sidner

A proposta de Grosz e Sidner [Grosz e Sidner 1986] assume a existência de uma estrutura genérica para o discurso. Esta estrutura é tripartida: sob a forma de pilha e árvore. São utilizadas duas pilhas: uma para a estrutura intencional e outra para a estrutura atencional do discurso. A árvore é utilizada como estrutura lingüística do discurso. O elemento básico representativo é a frase.

Grosz, a partir de seus estudos sobre diálogos orientados por tarefas (o conhecido diálogo instrutor-aprendiz [Grosz 1977]), concluiu que, neste tipo de discurso, as frases são estruturadas em árvore de forma a refletir a estrutura das tarefas e subtarefas envolvidas (estrutura intencional). Posteriormente junto com o trabalho realizado por Sidner na Teoria do Foco [Sidner 1981, Sidner 1979] e na Teoria de Centragem [Grosz, Joshi e Weinstein 1983], elas propõem uma estrutura tripartida para discurso:

1. estrutura da seqüência de frases, ou estrutura lingüística,

2. a estrutura dos propósitos de cada bloco de frases ou estrutura intencional,

3. e o estado atencional, destacando as entidades mais salientes em cada bloco de frases.

4.2.2.2 O Modelo Lingüístico do Discourse

Polanyi (Modelo Lingüístico do Discourse - LDM) [Polanyi, Berg e Ahn 2003, Polanyi 1988], baseando-se em parte no trabalho de Grosz ([Grosz 1977]), assume uma estrutura em árvore para o discurso. Esta árvore é gerada a partir da aplicação recursiva de um *conjunto de regras de formação* (coordenação e subordinação) sobre os diversos tipos de *unidades constituintes do discurso*[6] (DCUs): seqüências, expansões, unidades binárias e interrupções, ou recursivamente sobre uma composição anterior de DCUs. Como resultado do modelo adotado, tem-se uma árvore cujos nós são DCUs ou composições destas. Note-se ainda que Polanyi generaliza a idéia de Grosz [Grosz 1977] de que somente os nós mais à direita na árvore estão visíveis para futuras interpretações (por exemplo, de anáforas pronominais).

4.2.2.3 A Teoria da Estrutura Retórica

O trabalho de Mann e Thompson [Mann e Thompson 1987, Passonneau e Litman 1997], com a *Teoria da Estrutura Retórica* (RST), é uma metodologia para análise do discurso, propondo o agrupamento de frases satélites em torno da frase central do agrupamento – núcleo. O núcleo relaciona-se com seus satélites e com outros núcleos através de uma série de relações predefinidas.

As principais contribuições da RST para o estudo sobre os processos de estruturação do discurso são: (1) a criação de relações que agrupem dois ou mais elementos básicos e (2) a noção de que certos elementos da estrutura (núcleos) são num dado instante centrais ao processo de interpretação.

Apesar de reconhecer que a proposta da RST foi elaborada visando a análise do discurso (e uma possível estruturação deste) por humanos, o autor deste trabalho é partidário da idéia de que esta análise não é adequada, no presente momento, à interpretação computacional do discurso, porque exige uma análise em múltiplos níveis (lingüístico, semântico, intencional etc) [Moore e Pollack 1992, Hovy 1990], para o qual, em grande parte, ainda não existe um arcabouço computacional adequado. Porém o grande complicador é que a RST não depende somente da análise nos diversos níveis, mas principalmente da *interação*

[6] Frases, partes de frases ou interjeições.

entre estes níveis.

A RST, por outro lado, tem se mostrado muito útil na parte da geração de textos [Gardent 2002], onde a representação semântica de um discurso é usada na geração de frases.

4.2.2.4 Relações de Coerência

Hobbs [Hobbs 1985, Hobbs 1979] propõe uma estrutura em grafo onde os segmentos básicos representam o conteúdo proposicional das frases e os segmentos mais interiores representam as "relações de coerência" entre estes segmentos básicos.

Para gerar a estrutura, cada frase é comparada com todos os segmentos anteriormente interpretados, gerando um grafo com diversas possibilidades de interpretação entre o segmento básico da frase correntemente interpretada e o discurso previamente interpretado.

Como resultado, o grafo estabelece uma série de ligações da frase corrente com o discurso anterior, permitindo ao receptor fazer uma ligação coerente de todo o discurso. Esse modelo apresenta um conteúdo melhor em relação às propostas anteriores. Nele os agentes têm um coeficiente de informações mútuas elevado referente à comunicação sobre determinado assunto. Porém, em termos computacionais, encontrar estas relações é oneroso e, em termos práticos, desnecessário dependendo do tipo de fenômeno que se queira tratar.

4.2.3 Considerações finais

Consideradas as características de uma estrutura do discurso e a sua funcionalidade, as estruturas existentes podem ser classificadas em três casos:

- de um ponto de vista mais alargado, tal como nas propostas de Grosz e Sidner [Grosz e Sidner 1986] e da RST [Mann e Thompson 1987], onde se propõe uma estrutura geral para o discurso, fortemente baseada na identificação da intencionalidade (propósito) de cada frase ou conjunto de frases, tornando difícil a criação de processos computacionais que permitam implementá-las.

- de um ponto de vista intermediário estão as propostas de Polanyi [Polanyi 1988] e Hobbs [Hobbs 1985, Hobbs 1979]. Apesar do seu caráter também genérico, tais propostas são mais aplicáveis à obtenção de estruturas de resumo do discurso e a uma implementação computacional.

- e, finalmente, de um ponto de vista específico, existem as teorias que não vêem a estrutura do discurso como um objetivo final, mas sim como uma ferramenta auxiliar adequada aos fenômenos que se pretende analisar. Como exemplo deste tipo de estrutura está a proposta de Rodrigues e Lopes [Rodrigues e Lopes 1992, Rodrigues 1995] onde é criada uma estrutura para a interpretação temporal do discurso.

A estrutura nesta tese é uma árvore, específica para a resolução de anáforas. O conteúdo proposicional de cada frase, representado sob a forma de uma DRS modificada (seção 4.4.1), é a unidade básica. E por ter como objetivo a resolução de anáforas, a estrutura é fortemente baseada no acompanhamento do modo como as entidades introduzidas em cada frase, nomeadamente as que estão mais em evidência ou em foco, evoluem durante o discurso (mantendo-se salientes ou não).

Independente da forma de representação, esta tese considera que a interpretação de um discurso $D = f_1, f_2, ..., f_{i-1}, f_i, ..., f_n$, quando considerada a estrutura, segue os seguintes passos:

1. A estrutura E_0, contexto para a interpretação da frase f_1, está inicialmente vazia: $E_0 = \emptyset$.
2. repetir para $i = 1, ..., n$
   (a) A interpretação fora de contexto de f_i gera $I_i^{parcial}$.
   (b) $I_i^{parcial}$ é interpretada no contexto dado por E_{i-1}, gerando I_i.
   (c) Implícito ao processo de obtenção de I_i está a *localização do ponto de interpretação* pi_i *de* $I_i^{parcial}$ *em relação à estrutura* E_{i-1}. pi_i é a posição da estrutura E_{i-1} que serve de referência para a interpretação da frase f_i.
   (d) I_i é inserido no ponto pi_i, gerando E_i. Note que esta inserção é mais do que colocar I_i num determinado ponto. Em geral, esta inserção envolve uma reorganização de E_{i-1}de pi_i em diante.

Por fim o processo de inserção e geração da estrutura é fortemente baseado na característica de que as entidades mais salientes do discurso podem ser destacadas [Hajičová, Skoumalová e Sgall 1995, Grosz, Joshi e Weinstein 1995, Sidner 1981], que o acompanhamento destas é o acompanhamento de como o discurso evolui [Cohen e Erteschik-Shir 2002, Lopes e Freitas 1994, Grosz, Joshi e Weinstein 1995] e que a estrutura gerada por este acompanhamento permite a resolução de anáforas

[Strohner et al. 2000, Freitas e Lopes 1994]. Na próxima seção é apresentada a teoria do foco do discurso desenvolvida nesta tese.

4.3 O Foco do discurso

Foco do discurso é o termo utilizado para designar a entidade mais saliente do discurso [Hajičová, Skoumalová e Sgall 1995, Sidner 1981, Grosz 1977]. Tipicamente, o foco é a entidade sobre a qual o transmissor centra sua atenção em determinado ponto do discurso, sendo que a utilização continuada de uma determinada entidade através do uso de anáforas é um forte indício de que esta entidade está em foco [Grosz, Joshi e Weinstein 1995, Sidner 1981]. Veja o texto a seguir:

(4.2) a. Eram cinco irmãos de Coimbra.

b. O mais velho migrou pra França.

c. O mais novo formou-se advogado e vive em Lisboa.

d. Os outros foram morar no Porto.

A primeira frase (4.2a) introduz um conjunto pessoas, em número de cinco, que são irmãos entre si. Na frase (4.2b) é destacado deste conjunto um indivíduo em especial, *o mais velho*, o qual, supostamente, vive na França porque migrou pra lá. Note que ao utilizar-se uma anáfora nominal definida (*o ...*) o transmissor está querendo dizer que o assunto em discussão no texto continua sendo os irmãos introduzidos na frase (4.2a). Veja que se ele houvesse utilizado: "um irmão mais velho" (artigo indefinido) as duas primeiras frases ficariam desconexas dando a impressão de mudança de assunto[7]. O texto continua nas frases (4.2c) e (4.2d) versando sobre o conjunto de irmãos introduzidos na frase (4.2a).

De um modo geral considera-se que cada frase fala sobre um determinado assunto. Mais especificamente em cada frase existe um ou mais focos [Brennan, Friedman e Pollard 1987, Grosz, Joshi e Weinstein 1995].

Para resolver uma anáfora deve-se levar em conta:

1. foco do discurso,

[7] É possível que o receptor mediante uma série de raciocínios possa chegar a conclusão de que "um irmão mais velho" faça parte do conjunto de "irmãos que vivem em Coimbra", porém para chegar a esta conclusão ele terá que seguir um caminho mais longo do que se fosse utilizado "o irmão mais velho".

2. forma da anáfora (pronome, nome próprio, SND),

3. inferência pragmática.

O foco é freqüentemente influenciado por quão recente uma entidade foi introduzida no discurso (do inglês *recency*) e pelo contexto em que a entidade se enquadra (um garçom é mais saliente no contexto de um restaurante do que um cliente!)

A forma lingüística da anáfora, determinando o grau de informação semântica presente numa expressão anafórica, é que vai contribuir para facilitar a resolução ou não. De um modo geral, o gráfico de informação, apresentado nesta tese, representa bem este critério:

- as SNDs são mais rápidas na resolução do que os pronomes.
- o foco (explícito) influencia mais a resolução de anáforas pronominais (e elipses) do que outros tipos.

Os pronomes podem ter um status privilegiado em termos de acesso às informações conceituais (*deep information*) enquanto que as "formas cheias" (*full forms*) podem apenas produzir uma ativação imediata da estrutura de conhecimento num dado ponto.

4.3.1 Tipos de Foco

A literatura sobre foco do discurso pode ser dividida em duas partes: o foco da prosódia e o foco lingüístico [Brennan, Friedman e Pollard 1987, Grosz, Joshi e Weinstein 1995]. Esta tese não atenta para o foco da prosódia, visto que este trabalho é sobre textos escritos e não sobre fala.

Os focos lingüísticos são caracterizados:

1. Pela forma com que o foco é determinado:

   (a) utilizando o papel gramatical (agente e tema) e pelo uso de anáforas pronominais [Sidner 1981],

   (b) posição das entidades na frase (sujeito, objeto, objeto direto etc) e pelo uso de anáforas pronominais [Grosz, Joshi e Weinstein 1995, Brennan, Friedman e Pollard 1987].

2. Pelo número de focos existentes em cada frase:

(a) um foco para cada frase [Grosz, Joshi e Weinstein 1995, Brennan, Friedman e Pollard 1987, Kehler 1993, Strube e Hahn 1996],

(b) dois focos para cada frase [Sidner 1981].

4.3.2 Foco Implícito e Foco Explícito

Na literatura as propostas sobre o foco do discurso [Brennan, Friedman e Pollard 1987, Grosz, Joshi e Weinstein 1995] colocam-no sob a perspectiva do elemento mais saliente do discurso e, quando muito, existência de dois focos de acordo com o papel gramatical [Sidner 1979]. Esta abordagem é insuficiente para o tratamento das anáforas nominais definidas, onde na maioria das vezes não existe só um componente implícito ao discurso (que é um dos focos do discurso), mas também existe um outro componente explícito (foco) que determina o centro de atenção em cada frase. O primeiro componente é definido como sendo o *foco implícito do discurso* ou $foco^{imp}$, porque ele salienta o assunto sobre o qual versa determinado conjunto de frases, o segundo é definido como *foco explícito da frase* ou $foco^{exp}$ porque determina a entidade sobre a qual é centrada a atenção em cada frase. Veja o exemplo:

(4.3) a. O João trouxe uma cesta de piquenique.

b. A cerveja estava quente.

c´. Os salgadinhos estavam frios.

c´´. Ela estava fora da geladeira.

Na frase (4.3a) é introduzida a entidade *cesta de piquenique*. Na frase (4.3b) esta entidade é referenciada através do uso de uma anáfora nominal definida: *a cerveja*, fazendo com que o assunto do discurso continue a ser implicitamente *a cesta de piquenique*, porém a cerveja passa a ser a entidade mais saliente da frase. O resultado desta diferenciação é claro quando se colocam duas possíveis continuações: na frase (4.3c') é usada novamente uma anáfora nominal definida, o que indicará que existe uma referência ao assunto do discurso e não à frase anterior, logo *os salgadinhos* estão ligados à cesta de piquenique da frase (4.3a) que é o foco implícito do discurso até o momento e não à cerveja da frase (4.3b) que é o foco explícito. Agora se a continuação fosse a frase (4.3c") a anáfora pronominal *ela* faria referência direta ao foco explícito, que no momento é a cerveja da frase (4.3b).

Existe uma íntima ligação entre o uso de anáforas e os focos, mais especificamente, uma ligação entre o foco implícito e o uso de anáforas nominais definidas e o foco explícito e o emprego de anáforas pronominais.

Nenhuma das propostas de foco de atenção apresentadas a seguir faz a diferenciação do foco como relação ao seu papel local ou global no discurso:

A Teoria do Foco de Sidner [Sidner 1981] utiliza dois focos baseados puramente no caráter gramatical (agente e tema) das entidades envolvidas e na subseqüente co-referência destas através de anáforas nas frases seguintes. Os focos utilizados têm apenas um papel local na resolução de anáforas, mesmo considerando a pilha de *ex-focos do ator*. Isto não é adequado à resolução de anáforas nominais definidas, que têm um caráter fundamentalmente global sobre o discurso, exigindo uma busca global pelo seu antecedente. Concluindo, a proposta de Sidner apresenta bons resultados na resolução de anáforas pronominais [Freitas 1993] ao custo de um algoritmo complexo de resolução [Cormack 1992], mas não é adequada à resolução de anáforas definidas.

Grosz et al com a sua proposta da *Centering* [Grosz, Joshi e Weinstein 1995, Grosz, Joshi e Weinstein 1983] e toda a série de propostas nela baseada [Brennan, Friedman e Pollard 1987, Kameyama 1997, Walker, Lida e Cote 1994, Strube e Hahn 1996, Kameyama, Passanneau e Poesio 1993] visam a medir o grau de coerência entre duas frases. Para tal, acompanham a forma como um único foco (*centro de atenção*) evolui no decorrer da interpretação de frases seqüenciais. Este foco, ao ser *medido* entre duas frases consecutivas, é local e tal como a Teoria do Foco não é adequado à resolução das anáforas nominais definidas. Porém a *Centering* apresenta um algoritmo bem simples para a determinação do foco e mais ainda uma maneira de medir o grau de coerência entre duas frases. A coerência é um fator que influencia na delimitação do espaço de interpretação de uma anáfora nominal definida, principalmente porque delimita o foco implícito.

4.3.3 As Listas de Entidades Relevantes

Dado um discurso D constituído das frases $f_1, \ldots, f_{i-1}, f_i, \ldots, f_n$, seja $Refs_{i-1} = [e, \ldots, e^k_{i-1}]$ o conjunto de referentes do discurso introduzidos pela interpretação da frase f_{i-1}, então a **lista de entidades explícitas relevantes** (LR_{i-1}) será a lista ordenada dos referentes (o processo de ordenação é explicado na próxima seção). Os referentes em LR_{i-1} servirão, a priori, de antecedentes para a resolução das anáforas na interpretação da frase seguinte f_i. Eles também são utilizados no cálculo dos focos de cada

frase (seção 4.3.5).

Por definição a lista de entidades explícitas relevantes para a primeira frase será vazia:

$$LR_1 = \phi \tag{4.4}$$

A lista LR_{i-1} é uma ordenação parcial dos referentes introduzidos pela interpretação da frase f_{i-1}. Esta ordenação permite que a busca de um possível antecedente em LR_{i-1} não seja um mero processo de busca exaustiva. As entidades mais salientes, que são as mais prováveis de serem utilizadas como antecedentes, estarão melhor classificadas em LR_{i-1} e necessitarão apenas de uma confirmação semântica: verificar se o modelo que representa as frases anteriormente interpretadas não apresenta contradições [Freitas e Lopes 1998, Pereira, Damásio e Alferes 1993] após a introdução do referente T da expressão anafórica e da relação R deste com seu antecedente.

A ordem na qual as entidades estão dispostas em LR_{i-1} é fundamental para a resolução das anáforas e o cálculo dos focos.

4.3.4 Ordenação da Lista de Relevantes

Como um discurso não muda constantemente de assunto, as entidades sobre as quais são centradas a atenção do emissor/receptor também não vão mudar com muita freqüência. Um indício desta continuação é o uso freqüente de expressões anafóricas para referenciar as entidades que estão em foco [Grosz, Joshi e Weinstein 1995]. Como resultado, as entidades anafóricas deverão ser melhor classificadas do que as entidades não anafóricas no processo de ordenação da lista de relevantes (LR_{i-1}), resultando na seguinte ordem de classificação:

$$\text{entidades anafóricas} > \text{entidades não anafóricas} \tag{4.5}$$

onde "$>$ " significa aqui que estão melhor classificadas e por isso estão à cabeça da lista.

Outro fenômeno que influencia a ordenação de LR_{i-1} é o tipo da expressão anafórica utilizada em f_{i-1} para confirmar a existência da entidade na lista [Garrod, Freudenthal e Boyle 1994]:

elipses: A falta de material sintático resultante do uso de elipses indica que o emis-

sor tem plena confiança que o receptor sabe recuperar o material sintático elidido [Kehler 2000]. Para agilizar a resolução de elipses por parte do receptor o transmissor normalmente associa a resolução da elipse com o foco explícito da frase anterior. Em termos de classificação para a lista de relevantes, significa que as entidades referenciadas por elipses têm preferência para continuarem a ser mais salientes nas próxima frases e, portanto, devem ser melhor classificadas na lista de relevantes.

pronomes: O uso de pronomes para referenciar entidades previamente introduzidas indica que o transmissor acredita que o receptor tem capacidade de recuperar a referência [Gundel, Hegarty e Borthen 2003, Strube e Hahn 1996, Brennan, Friedman e Pollard 1987, Sidner 1981], porém tem uma margem de certeza menor que com as elipses. Por esta razão usa um pronome que, apesar de, em termos semânticos, ter tão pouca informação quanto uma elipse, apresenta algumas informações morfológicas básicas (número, gênero e grau) que, em caso de ambigüidade, podem ser usadas para recuperar o antecedente.

sintagmas nominais definidos: O uso de um sintagma nominal definido por parte do transmissor indica três possibilidades [Blutner 2000, Abbott 1993]: (1) que este não tem certeza que o receptor possa facilmente recuperar o antecedente e, por conseguinte, possa realizar a interpretação da frase em contexto sem a descrição detalhada do antecedente. Por isso ele coloca o máximo de informação na expressão anafórica de forma a facilitar o trabalho do receptor, (2) o receptor pode recuperar o antecedente, porém este foi introduzido e deixou de ser referenciado nas últimas frases interpretadas. Sendo assim é necessária uma maneira de relembrar o receptor do ponto no texto onde o antecedente se encontra, e (3) que a expressão não é anafórica.

Dentro dos tipos de fenômenos citados, nota-se que existe um comprometimento entre a quantidade de informação utilizada na construção da expressão anafórica e a ordem pela qual a expressão referenciada deve ser ordenada na lista de relevantes (LR). Assim esta tese propõe que, quanto mais informação léxica, sintática e semântica estiver presente numa expressão anafórica, pior deverá ser sua classificação na lista de relevantes (4.6).

$$\Big\Downarrow \begin{array}{c} \text{definidos} \\ \text{pronomes} \\ \text{elipses} \end{array} \Big\Downarrow \tag{4.6}$$

O resultado é que a saliência na LR será inversamente proporcional à informação presente na expressão anafórica (4.7).

$$\downarrow \text{informação} \quad \uparrow \text{saliência em LR} \tag{4.7}$$

As elipses[8] estão melhor classificadas do que os pronomes e estes, por sua vez, estão melhor classificados do que os sintagmas nominais definidos.

Como resultado destes dois critérios, a ordenação da LR é feita da seguinte forma:

1. As entidades anafóricas estarão melhor classificadas do que as entidades não anafóricas.

2. Dentro do conjunto das entidades anafóricas, as entidades referidas via elipses estarão melhor classificadas do que as entidades referidas via anáforas pronominais e estas melhor classificadas do que as entidades referidas via anáforas nominais definidas.

3. As entidades anafóricas do mesmo tipo são ordenadas pela ordem gramatical da frase [Grosz, Joshi e Weinstein 1995, Brennan, Friedman e Pollard 1987] (4.8).

$$sujeito > objeto > objeto\,direto \tag{4.8}$$

4. As entidades não anafóricas também são ordenadas pela ordem gramatical da frase.

O resultado desta regras de ordenação é apresentado em 4.9.

$$\begin{array}{ccc} \text{entidades anafóricas} & > & \text{entidades não anafóricas} \\ \text{elipse} > \text{pronomes} > \text{SND} & & \text{sujeito} > \text{objeto} > \text{objeto2} \\ \text{sujeito} > \text{objeto} > \text{objeto2} & & \end{array} \tag{4.9}$$

Se três entidades quaisquer u_i, u_j e u_k estão ordenadas de acordo com (4.9) então existe uma relação de classificação parcial definida através de:

$$u_i \succcurlyeq u_j \succcurlyeq u_k \tag{4.10}$$

[8] O fenômeno da elisão do sujeito é uma situação freqüente no Português, e tem um correspondente nas mesmas condições com o emprego de pronomes pessoais no caso do Inglês e do Francês.

onde u_i está melhor classificada do u_j e esta por sua vez está melhor classificada do que u_k.

4.3.5 Cálculo dos Focos

Considere o conjunto de referentes do discurso $Refs_i = [u, u_2, ..., u_n]$ introduzidos pela interpretação semântica da frase corrente f_i. A ordenação $Refs_i$, segundo os critérios apresentados na seção 4.3.4, produz a Lista de Entidades Relevantes explicitamente introduzidas em f_i, a LR_i^{exp}. A entidade melhor classificada em LR_i^{exp} será o ***foco explícito*** da frase f_i, ou $foco_i^{exp}$.

Este foco assemelha-se ao "*backward looking center*" (Cb) da Teoria de Centragem [Grosz, Joshi e Weinstein 1995, Brennan, Friedman e Pollard 1987, Grosz, Joshi e Weinstein 1983] e ao Foco do Discurso (*Discourse Focus*) da Teoria do Foco [Sidner 1979, Sidner 1981]. Ele utiliza somente as entidades *explicitamente* introduzidas ou co-referenciadas. A diferença dessas propostas para esta tese são os critérios de ordenação: aqui eles levam em consideração a distinção entre elementos anafóricos ou não e o tipo de expressão anafórica.

Agora considere a **Lista de Entidades Implícitas Relevantes** $LR_i^{imp} = [r_1, r_2, ..., r_n]$, composta somente pelas entidades referenciadas (i.e. antecedentes) por anáforas nominais definidas da frase f_i e classificadas pela ordem de aparecimento da expressão anafórica na frase f_i (definição 4.8). A entidade melhor classificada em LR_i^{imp} será o ***foco implícito*** da frase f_i, ou $foco_i^{imp}$. Note que, para a interpretação da primeira frase do discurso f_i, não existe nada previamente interpretado, portanto a LR_1^{imp} é vazia. Logo, o foco implícito da primeira frase será nulo: $foco_1^{imp} = nulo$.

O foco implícito não é desenvolvido na literatura de resolução de anáforas, apesar de ter sido citado por Sanford e Garrod [Sanford e Garrod 1981]. A razão para isto é histórica: em 1979 Sidner cria a teoria do Foco [Sidner 1979] a qual utiliza dois focos para a resolução de anáforas pronominais, porém a metodologia para determinação dos focos era complexa, pois envolvia considerações a respeito de animacidade, posição sintática e referência anafórica. Em 1983, Grosz et al [Grosz, Joshi e Weinstein 1983] propõem a Teoria da Centragem como uma metodologia para aferir o grau de coerência entre as frases do discurso. Para tal eles utilizam a movimentação de um único centro de atenção (foco) como a medida da coerência. Na Teoria da Centragem o foco é facilmente calculado e quase exclusivamente dependente das informações sintáticas. Houve um embate, natural, sobre a utilização de um ou dois focos, o qual foi ganho pela simplicidade na determinação do

foco na Teoria da Centragem. A *vitória*, se assim se pode dizer, foi selada com a proposta conjunta de Grosz e Sidner [Grosz e Sidner 1986] da Estrutura Tripartida do Discurso, a qual utiliza somente um foco.

Com a crescente utilização da Teoria da Centragem (para medição de coerência do discurso) a utilização de dois focos foi gradativamente deixando de ser utilizada. Assim, a proposta de um foco alternativo para o tratamento das Anáforas Nominais Definidas, da maneira proposta nesta tese, acabou sendo protelada, apesar de ser uma maneira elegante de resolver as ANDs e com a mesma simplicidade do que foi proposto na Teoria da Centragem.

4.3.6 Uso dos Focos e da LR na Resolução de Anáforas

O foco explícito é utilizado na resolução de anáforas pronominais e elipses. O $foco_{i-1}^{exp}$ da frase anterior é o mais forte candidato a antecedente numa frase subseqüente f_i. O uso do foco nestas condições assinala a tendência do transmissor em continuar falando sobre um mesmo indivíduo, permitindo assim introduzir mais informações sobre o mesmo. Vale a pena destacar que tanto os pronomes quanto as elipses *contêm* reduzido material informativo: posição sintática, número e gênero. O que torna mais presente a necessidade do foco.

Caso a utilização de $foco_{i-1}^{exp}$ para a resolução da anáfora ou elipse não seja possível (em virtude de alguma restrição de caráter semântico ou pragmático) serão usadas as outras entidades existentes na frase f_{i-1}, as quais estão representadas e ordenadas em LR_{i-1}^{exp}. Note que neste caso poderá haver uma *troca* de foco, sinalizando a mudança de atenção das entidades atuais em favor de novas entidades.

Assim, dada uma expressão anafórica[9] qualquer A_i^j numa frase f_i seu antecedente T_i^j é determinado por:

$$T_i^j = foco_{i-1}^{exp} \, sse \, K_{i-1} \cup \{A_i^j = foco_{i-1}^{exp}\} \nvDash \perp \tag{4.11}$$

onde K_{i-1} é o contexto resultante das $(i-1)$ frases anteriores.

Caso o antecedente T_i^j não seja $foco_{i-1}^{exp}$, então são utilizados os outros elementos de LR_{i-1}^{exp}:

[9] Resultante de um pronome ou elipse.

$$T_i^j = u_k \, sse \, u_k \in LR_{i-1}^{exp}, k = 2...t \wedge \neg u_l \in LR_{i-1}^{exp} \,|\, u_l \gg u_k \wedge K_{i-1} \cup \{A_i^j = u_k\} \nvDash \perp \quad (4.12)$$

onde u_k é um dos outros elementos de LR_{i-1}^{exp} (exceção feita ao $foco_{i-1}^{exp}$, k=1) e t é o número de elementos de LR_{i-1}^{exp}.

Usando as definições (4.11) e (4.12), a ordenação das entidades dada por (4.9) e a definição do cálculo do foco explícito da seção 4.3.5, os passos para a resolução de uma anáfora são:

1. Para a primeira frase f_1: $LR_0^{exp} = \phi$ e $foco_0^{exp} = nulo$ (qualquer pronome ou elipse A_1^j na primeira frase será considerado como não sendo anafórico, pois não existe nenhum antecedente possível[10]).

2. Repetir para todas as frases do discurso $f_i, i = 1...n$:

   (a) para a resolução de cada expressão anafórica A_i^j encontrada em f_i são aplicadas as definições (4.11) e (4.12),

   (b) resolvidas todas as possíveis anáforas A_i^j, cria-se o conjunto $Refs_i$ que contém todas as entidades introduzidas em f_i,

   (c) aplica-se sobre $Refs_i$ as regras de ordenação dadas por (4.9), determinando então LR_i^{exp} e $foco_i^{exp}$, elementos a serem utilizados na interpretação da próxima frase f_{i+1},

   (d) por fim, as relações de co-referência que ligam as anáforas de f_i aos seus antecedentes em K_{i-1} podem ser inseridas na interpretação em contexto K_i.

O foco implícito acompanha as entidades que foram co-referenciadas através da utilização de SNDs, sendo então propício para a resolução de anáforas nominais definidas, onde a expressão anafórica $\mathcal{A}$ e seu antecedente $\mathcal{T}$ não co-referenciam a mesma entidade, i.e. $\mathcal{A} = \mathcal{T}$, mas sim, como já visto é necessário introduzir a relação $\mathcal{R}(\mathcal{A}, \mathcal{T})$ sendo que possivelmente $\mathcal{T} = foco_{i-1}^{imp}$.

Tal como na utilização do foco explícito, caso o $foco_{i-1}^{imp}$ não possa ser utilizado na resolução da SND, será então utilizada a LR_{i-1}^{imp}, podendo neste caso haver uma troca de foco implícito, sinalizando uma mudança de assunto do discurso. A diferença é que

[10]Pode se pensar num contexto visual, mais este caso entraria no âmbito dos deíticos, os quais estão fora do escopo desta tese.

caso nenhum elemento de LR_{i-1}^{imp} seja o antecedente, é utilizada LR_{i-1}^{exp}. Persistindo a não resolução, $\mathcal{A}$ é acomodado (o referente é considerado como um indefinido).

Assim dado um SND A_i^j numa frase f_i, determina-se seu antecedente T_i^j através de:

$$T_i^j = foco_{i-1}^{imp}\, sse\, \exists R \,|\, K_{i-1} \cup \{R(A_i^j, foco_{i-1}^{imp})\} \nvdash \bot \tag{4.13}$$

onde K_{i-1} é o contexto resultante das $(i-1)$ frases anteriores e $\mathcal{R}$ é um das relações entre entidades apresentadas no capítulo 3.

Caso o antecedente T_i^j não seja o $foco_{i-1}^{imp}$, então são utilizados os outros elementos de LR_{i-1}^{imp}:

$$T_i^j = u_k\, sse\, u_k \in LR_{i-1}^{imp}, k = 2...t \wedge \neg u_l \in LR_{i-1}^{imp} | u_l \gg u_k \wedge K_{i-1} \cup \{R(A_i^j, u_k)\} \nvdash \bot \tag{4.14}$$

onde u_k é um dos outros elementos de LR_{i-1}^{imp} (exceção feita ao $foco_{i-1}^{imp}$, k=1) e t é o número de elementos de LR_{i-1}^{imp}.

Ainda, caso o antecedente T_i^j não seja nenhum dos outros elementos da LR_{i-1}^{imp}, então são utilizados os elementos da LR_{i-1}^{exp}:

$$T_i^j = u_k\, sse\, u_k \in LR_{i-1}^{exp}, k = 1...t \wedge \neg u_l \in LR_{i-1}^{exp} | u_l \gg u_k \wedge K_{i-1} \cup \{R(A_i^j, u_k)\} \nvdash \bot \tag{4.15}$$

Por fim, se nenhuma das entidades de LR_{i-1}^{exp} puder ser utilizada na resolução, então A_i^j é acomodada. Apesar de ter sido introduzida como um SND ela se comporta como um indefinido:

$$K_{i-1} \cup \{acomoda(A_i^j)\} \tag{4.16}$$

Usando as definições de (4.13) a (4.16), a ordenação das entidades dada por (4.9) e a definição do cálculo do foco implícito da seção 4.3.5, os passos para a resolução de uma SND são:

1. Para a primeira frase f_1: $LR_0^{imp} = \phi$, $foco_0^{imp} = nulo$, $LR_0^{exp} = \phi$ e $foco_0^{exp} = nulo$.
2. Repetir para todas as frases do discurso $f_i, i = 1...n$:

(a) para a resolução de cada expressão anafórica A_i^j encontrada em f_i são aplicadas as definições de (4.13) a (4.16). Note que para a frase f_1, $foco_1^{imp} = nulo$ e $LR_1^{imp} = \phi$,

(b) resolvidas todas as possíveis anáforas nominais definidas A_i^j, cria-se o conjunto $Refs_i^{imp}$ que contém seus respectivos antecedentes T_i^j,

(c) $Refs_i^{imp}$ é ordenado de acordo com a ordem de aparecimento de suas respectivas expressões anafóricas (eq. 4.8), determinando então LR_i^{imp} e $foco_i^{imp}$, elementos a serem utilizados na interpretação da próxima frase f_{i+1},

(d) por fim, os referentes introduzidos pelas expressões anafóricas e as relações entre estes e seus antecedentes são introduzidos na interpretação em contexto K_i.

Assim, a utilização dos focos e das listas de relevantes proporciona um duplo benefício na resolução de anáforas pronominais, nominais definidas e elipses: (1) reduz potencialmente o espaço de procura por um antecedente e (2) ordena as entidades no *subespaço de busca* de forma que aquelas melhor classificadas são as mais prováveis de serem antecedentes. Tudo isto proporciona uma otimização no processo de interpretação.

Mas note que isto somente é possível quando um provável antecedente estiver numa das listas em foco. Caso contrário a interpretação falha (no caso das elipses/pronomes) ou a entidade é acomodada no caso das SNDs. O que fazer nestes casos? Existem duas opções:

1. Considerar que a interpretação do discurso realmente falhou e parar a interpretação. Isto seria possível em casos em que o transmissor, intencionalmente, transmite discursos incoerentes.

2. Tentar continuar a interpretação (das expressões anafóricas), olhando para as entidades de outras frases que já não estão mais nas listas de relevantes nem em foco.

Ambos os casos, quando implementados da forma como colocado, não produzem bons resultados: para a interpretação seria o mesmo que deixar de ler um livro porque não fez uma determinada ligação. Por outro lado, sair olhando para todas as entidades anteriores é um processo computacional exaustivo e oneroso.

A solução é estruturar não só as entidades relevantes e os focos, mas também as frases que os contêm. Desta forma continua a se ter uma redução do espaço de busca,

aliada agora a uma busca ordenada de uma lista de relevantes já passada (não visível à interpretação atual).

Neste contexto, o uso tanto do foco explícito quanto do implícito é o acompanhamento das continuações e mudanças do centro de atenção em cada frase e/ou no conjunto, indicando como organizá-las em relação ao assunto atual do discurso. Assim, o foco explícito serve de medida para a coerência local, ou seja, entre duas frases consecutivas, e o foco implícito serve tanto para medir a coerência local quanto de um conjunto de frases que versam sobre um mesmo assunto, i.e, que tenham o mesmo foco implícito.

Usando a movimentação dos focos implícitos e explícitos é criada a estrutura que permite organizar a lista de relevantes de forma que quando há falha na interpretação de uma expressão anafórica entre uma frase e sua lista de relevantes, esta estrutura assume a função de determinação de qual lista de relevantes deve então ser utilizada. Esta estrutura é o tema da próxima seção.

4.4 Estrutura Nominal do Discurso

A estrutura nominal do discurso é resultado da organização das entidades existentes nas frases do discurso, em especial das mais salientes. O acompanhamento dos focos permite acompanhar a evolução dos assuntos no decorrer do discurso, refletindo parte da estrutura mental do transmissor em relação aos indivíduos ou entidades existentes no discurso [Kruijff-Korbayová e Steedman 2003].

Para se obter a estrutura nominal é feito o acompanhamento da movimentação das entidades salientes durante as frases de um discurso. O acompanhamento da saliência das entidades é a parte central de teorias tais como: Teoria da Centragem [Grosz, Joshi e Weinstein 1995] e Teoria do Foco [Sidner 1981], porém nenhuma delas estabelece um nível abstrato de agrupamento para este acompanhamento. A Teoria da Centragem usa a mudança de saliência para explicar a coerência entre duas frases subseqüentes. A Teoria do Foco usa a mudança de saliência para resolver anáforas pronominais. Nenhuma delas, ao contrário da proposta feita nesta tese, olha para a mudança de saliência como fonte de informação para acompanhar/estruturar as entidades utilizadas no discurso.

Outra grande diferença, relativamente à proposta desta tese, relaciona-se com os elementos salientes: enquanto nos casos anteriores os elementos salientes são apenas os que estão explícitos no discurso, aqui é considerada a existência de entidades que foram

explicitamente enunciadas e que, em frases subseqüentes, continuaram a ser referidas implicitamente (i.e. indiretamente). São entidade que ficam implícitas ao discurso e que, no mais das vezes, são introduzidas pelas anáforas nominais definidas.

A teoria aqui proposta baseia-se no acompanhamento dos focos do discurso (explícito e implícito) e no agrupamento de todo o material semântico existente na interpretação de uma frase – o chamado **segmento básico** – numa estrutura em árvore onde somente os nós mais à direita estão abertos para interpretação da próxima frase (fig. 9). Os nós internos são chamados **segmentos** e são compostos de material semântico herdado de seus nós filhos.

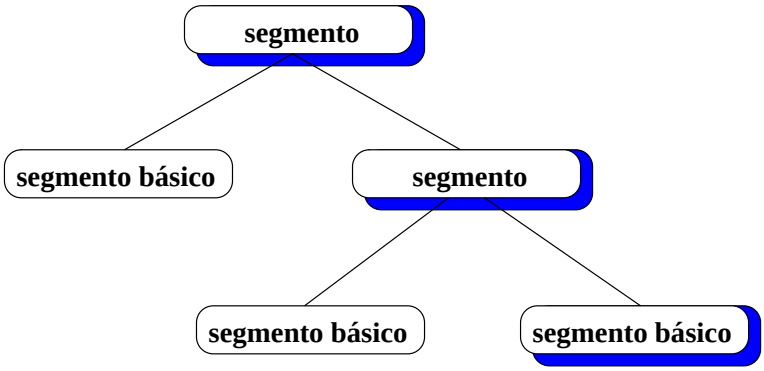


Figura 9: Segmentos da estrutura nominal do discurso.

A interpretação de uma nova frase, feita nos nós visíveis, seguirá a ordem do nó mais recente em direção ao nó *raiz* (fig. 10). A linha (1) corresponde à interpretação obtida quando se considera apenas as listas de relevantes entre as duas últimas frases. Caso a interpretação falhe, então a frase vai ser interpretada relativamente ao próximo segmento visível (linha 2). Caso ainda falhe, então é tentado o último segmento visível (linha 3). Caso este falhe, então a frase não pode ser interpretada nesta estrutura.

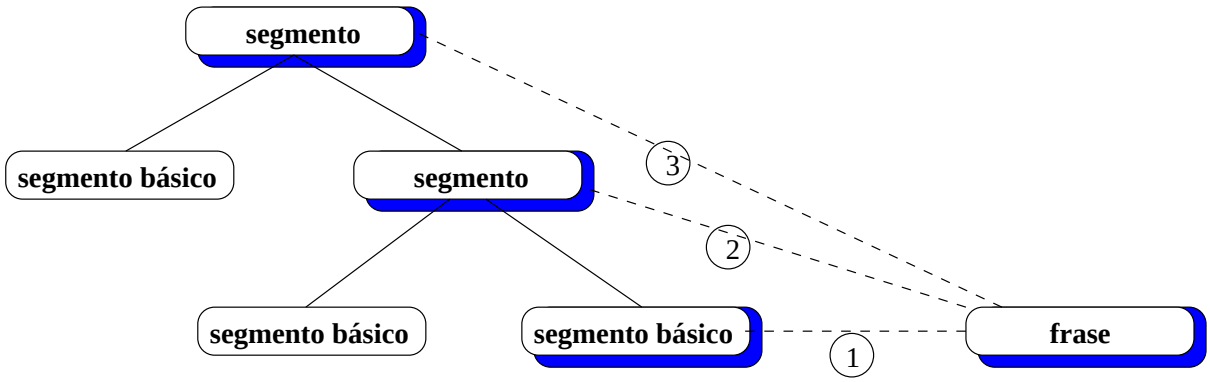


Figura 10: Ordem da interpretação de uma frase na estrutura nominal.

4.4.1 Segmento básico

Existem dois tipos de segmentos básicos: o primeiro originário da interpretação fora de contexto de uma frase e o segundo resultante da interpretação deste em contexto. O primeiro é denominado *segmento básico fora de contexto* e o segundo é denominado *segmento básico em contexto.*

O segmento básico fora de contexto é uma DRS [Kamp e Reyle 1993] onde os referentes representam entidades do discurso com suas respectivas condições e com as seguintes relações especiais denominadas **condições âncora**: *snd(x)* indicando um SND, *pro(x)* indicando um pronome e *eli(x)* indicando uma elipse. Por exemplo:

(4.17) Ele comprou flores.

que será representado como:

e, f
sing(e),masc(e),sujeito(e), pro(e), flor(f), plu(f),fem(f),objeto(f), comprar(e,f).

(4.18)

Já os segmentos básicos em contexto são DRS modificadas, cujos atributos e domínio de valor são os seguintes:

- Referente do segmento - s_i, uma constante diferente de qualquer outra existente. Seu valor identifica unicamente o segmento.
- Tipo do segmento - $tipo_i$: *básico.*
- Expressão lógica - $Conds_i$ usada para veicular as informações do discurso, provenientes da forma lógica da frase interpretada em contexto. Esta expressão contém também as relações resultantes da resolução das expressões anafóricas (e.g. *co-referência*, *parte de*, *membro de*, *subcategorizado por* e *acomodação),* bem como as condições âncora.
- Universo da DRS - LR_i^{exp}, lista ordenada dos referentes do discurso resultantes da interpretação fora de contexto da frase.

- Foco explícito - $foco_i^{exp}$, referente do discurso (ou valor *nulo*), indicando a entidade explícita saliente na frase.
- Lista de entidades relevantes implícitas - LR_i^{imp}, contendo os referentes do discurso resultantes da interpretação em contexto dos SNDs da frase.
- Foco implícito - $foco_i^{imp}$, referente do discurso (ou valor *nulo*), indicando a entidade implícita saliente na frase.

Esquematicamente estas informações estão organizadas de acordo com o diagrama (4.19):

s_i	$tipo_i : básico$
$foco_i^{exp}$ $foco_i^{imp}$	LR_i^{exp} LR_i^{imp}
$Conds_i$	

(4.19)

Tome o exemplo:

(4.20) a. Lucas foi à floricultura.

b. Ele comprou as flores.

A interpretação em contexto da frase (4.20b) terá como segmento a DRS (4.21):

s_b	$básico$
$foco_b^{exp} = l$ $foco_b^{imp} = f$	$LR_b^{exp} = [p, fs]$ $LR_b^{imp} = [f]$
$lucas(l), floricultura(f),$ $flores(fs), ele(p),$ $snd(fs).$	

(4.21)

4.4.2 Segmento composto

Os nós internos usados para a construção da árvore que representa a estrutura nominal do discurso são chamados **segmentos compostos** ou simplesmente **segmentos** (vide figura 10). Cada segmento representa um agrupamento de segmentos filhos, os quais podem ser segmentos básicos ou outros segmentos compostos.

Tal como um segmento básico, um segmento composto é uma lista de pares atributo-valor:

- Referente do segmento - s_i, expressão lógica - $Conds_i$, universo da DRS - LR_i^{exp}, foco explícito - $foco_i^{exp}$, lista de entidades relevantes implícitas - LR_i^{imp}, foco implícito - $foco_i^{imp}$, todos com definições idênticas às dos mesmos elementos num segmento básico.

- Tipo do segmento - $tipo_i$: $tipo_i \in \{elaboracao, mudanca\_assunto, mudanca\_topico, manutencao\_topico\}$, o qual indica como o segmento composto foi formado.

- Subsegmentos - uma lista (possivelmente vazia) ordenada de referentes de segmento. É a lista que dá a forma de árvore à estrutura nominal. Os segmentos da lista são os filhos, as folhas são sempre do tipo *básico*.

A lista de subsegmentos dá a forma de árvore à END, mas a inserção de novos nós na árvore não pode ser feita de maneira aleatória. Como já dito, somente os segmentos visíveis podem ser usados como referência.

Esquematicamente, um segmento composto é representado pelo diagrama (4.22):

s_i	$tipo_i$
$foco_i^{exp}$ $foco_i^{imp}$	LR_i^{exp} LR_i^{imp}
$Conds_i$	
sub-segmentos	

(4.22)

A criação de um determinado tipo de segmento depende exclusivamente da relação existente entre os focos de cada segmento-filho. Esta relação é determinada a partir da mudança ou manutenção do foco implícito, sinalizando uma mudança ou continuação do assunto. Por sua vez, a mudança ou manutenção do foco explícito sinaliza a mudança ou manutenção do tópico de uma frase (efeito local). Os tipos de segmentos existentes são:

- elaboração - um segmento do tipo elaboração indica que o assunto de seus segmentos-filhos é o mesmo e há uma elaboração sobre uma mesma entidade do discurso (tópico) nestes segmentos.

- mudança de assunto - este segmento indica que o discurso passa a dissertar sobre um novo assunto que não tem ligação nenhuma com o tópico anterior.

- mudança de tópico - este segmento indica que o discurso disserta sobre um tópico (entidade) diferente de um mesmo assunto.

- manutenção de tópico - neste segmento houve uma mudança de assunto, porém algumas entidades do assunto anterior continuarão a ser referenciadas no novo assunto. Esta relação lembra a "associação" em que as entidades de um determinado tema podem induzir um novo assunto.

Estes tipos de segmentos são obtidos aplicando as regras da tabela 5:

	$foco_{i-1}^{exp} = foco_i^{exp}$	$foco_{i-1}^{exp} \neq foco_i^{exp}$
$foco_{i-1}^{imp} = foco_i^{imp}$	**elaboração**	**mudança de tópico**
$foco_{i-1}^{imp} \neq foco_i^{imp}$	**manutenção do tópico**	**mudança de assunto**

Tabela 5: Relações entre $foco^{exp}$, $foco^{imp}$ e o tipo de segmento gerado.

Usando a movimentação ou continuidade dos focos expressos na tabela 5 é possível acompanhar a estrutura de assuntos transmitida num discurso. Cada elemento desta estrutura, em especial os segmentos visíveis, constitui um espaço ordenado de procura para a interpretação das expressões anafóricas da frase em interpretação.

Na próxima seção é definido o conteúdo semântico de cada segmento, o qual é determinado pela herança de atributos de seus segmentos-filhos e depende exclusivamente do tipo do segmento composto.

4.4.3 Criação de um segmento

Um segmento é criado quando da interpretação em contexto de uma frase qualquer f_i em relação a um único ponto da END. Este ponto é um nó da árvore. O novo segmento seg_{novo} é o resultado da criação de um novo nó na árvore, o qual vai herdar atributos do ponto de ancoragem (filho à esquerda, seg_2) e do segmento $básico_1$ resultante da interpretação fora de contexto f_i (figura 11).

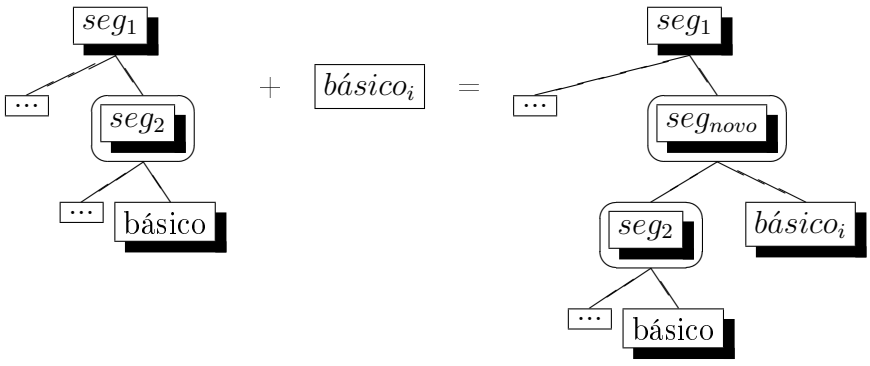


Figura 11: Composição de um novo segmento

Os atributos herdados constituem a única forma de inserção de material semântico num segmento composto. Devido a esta herança, um segmento composto tem duas funções bem definidas: (1) *resumir* as informações dos seus subsegmentos imediatos e (2) servir de *ponto de interpretação* para as próximas frases.

A determinação da herança depende exclusivamente da comparação entre os focos de dois segmentos, sejam eles básicos ou não. O resultado desta comparação dá origem a quatro tipos de segmentos: elaboração, mudança de assunto, mudança de tópico, manutenção de tópico.

4.4.3.1 Segmento do tipo elaboração

Um segmento novo do tipo elaboração, isto é, um segmento que é formado pela composição de outros dois subsegmentos onde ambos os focos implícitos e explícitos são iguais, ou caso o foco implícito do segmento mais antigo[11] seja *nulo*. A análise da herança deste tipo de segmentação é apresentada a seguir:

- Foco implícito do segmento composto - $foco_{seg}^{imp}$: focos implícitos iguais nos subsegmentos indicam que o transmissor está falando sobre um mesmo assunto e, portanto, a entidade *implícita* mais saliente é o foco implícito (não nulo) comum. O segmento composto, que é um segmento pai, vai representar o resumo de seus filhos, logo será herdado o assunto em comum $foco_{seg}^{imp}$, um foco implícito não nulo:

$$foco_{seg}^{imp} = foco_{filho1}^{imp} \text{ se } foco_{filho1}^{imp} \neq nulo \text{ senão } foco_{seg}^{imp} = foco_{filho2}^{imp} \tag{4.23}$$

- Lista de relevantes implícita do segmento composto - LR_{seg}^{imp}: com a continuação do assunto expressa pela continuação dos focos implícitos dos subsegmentos o $foco_{seg}^{imp}$, a lista LR_{seg}^{imp} (como todas as LRs) terá este elemento à cabeça:

$$LR_{seg}^{imp} = [foco_{seg}^{imp}] \tag{4.24}$$

- Foco explícito do segmento composto - $foco_{seg}^{exp}$: focos explícitos iguais nos subsegmentos indicam que o transmissor está falando sobre uma mesma entidade. $foco_{seg}^{exp}$ será esta entidade comum, representando um resumo do tópico de seus subsegmentos:

[11] Aquele segmento que é resultado da interpretação mais antiga, quando comparados dois a dois.

$$foco^{exp}_{seg} = foco^{exp}_{filho1} \tag{4.25}$$

- Lista de relevantes explícita do segmento composto - LR^{exp}_{seg}: a continuação do tópico, na forma da herança de $foco^{exp}_{seg}$, indica que este elemento é o *resumo*, não havendo necessidade da herança direta ou da combinação dos outros elementos das LRs de cada um dos subsegmentos. Como resultando LR^{exp}_{seg} é uma lista de um só elemento:

$$LR^{exp}_{seg} = [foco^{exp}_{seg}] \tag{4.26}$$

- A expressão lógica (condições da DRS) estará disponível conjuntamente com todos os referentes herdados (focos e LRs) pelo segmento composto.

Considere o seguinte exemplo:

(4.27) a. O ônibus chegou à rodoviária.

b. O motorista conversou com o cobrador.

c. (e) Φ saiu pela porta da frente.

A interpretação da frase (4.27b) tomando por base a frase (4.27a) produz o seguinte segmento básico:

s_b	*básico*
$foco^{exp} = m$	$LR^{exp} = [m, c]$
$foco^{imp} = o$	$LR^{imp} = [o]$
motorista(m), *snd(m),subcat_por(m,o)*, cobrador(c), *snd(c),subcat_por(c,o)*, conversar(o,c).	

(4.28)

Considerando a geração do segmento composto resultante da frase (4.27c) relativa ao segmento (4.28), tem-se a subárvore da figura 12:

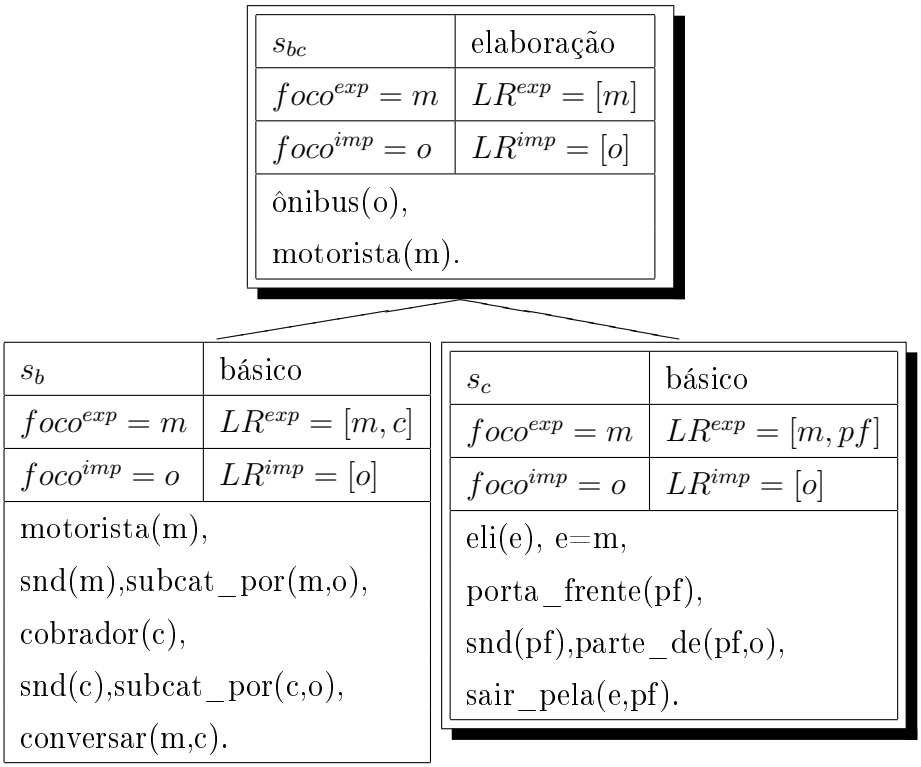


Figura 12: Subárvore resultante da interpretação das frases (4.27b) e (4.27c).

4.4.3.2 Segmento do tipo mudança de assunto

Um segmento novo do tipo mudança de assunto é formado pela composição de dois subsegmentos – denominados segmento *mais antigo* ou *mais à esquerda* e segmento *mais novo* ou *mais à direita* – cujos focos explícitos e focos implícitos são diferentes. Exceção feita quando *somente* o foco implícito do segmento *antigo* for *nulo*, caso em que os focos são considerados iguais. A análise da herança destes tipos de segmentação é apresentada a seguir:

- Foco implícito do segmento composto - $foco^{imp}_{seg}$: focos implícitos diferentes nos subsegmentos indicam que o transmissor mudou de assunto, caso em que começa um novo assunto (foco implícito). Isto significa que o assunto antigo deve ser *arquivado.* O resultado, em termos da END, é que o segmento antigo não deve ser esquecido, mas sim apenas a sua subárvore. Na END isto é obtido quando os atributos estão num segmento visível. Logo o foco implícito do segmento composto será:

$$foco^{imp}_{seg} = foco^{imp}_{filho_{antigo}} \tag{4.29}$$

- Lista de relevantes implícita do segmento composto - LR^{imp}_{seg}: visando dar uma maior amplitude à retomada de assunto posteriormente, a lista de relevantes implícita deve conter todos os elementos do segmento antigo, logo:

$$LR^{imp}_{seg} = LR^{imp}_{filho_{antigo}} \tag{4.30}$$

- Foco explícito do segmento composto - $foco^{exp}_{seg}$: focos explícitos diferentes nos subsegmentos indicam que o transmissor mudou de tópico (já havia mudado de assunto!) O segmento composto deve então herdar seus atributos do segmento mais antigo. Isto permite uma futura retomada do tópico:

$$foco^{exp}_{seg} = foco^{exp}_{filho_{antigo}} \tag{4.31}$$

- Lista de relevantes explícita do segmento composto - LR^{exp}_{seg}: utilizando o mesmo critério aplicado ao foco explícito, implica que a lista de relevantes explícita também deve ser herdada do segmento mais antigo:

$$LR^{exp}_{seg} = LR^{exp}_{filho_{antigo}} \tag{4.32}$$

- A expressão lógica estará visível, conjuntamente com todos os referentes herdados pelo segmento composto.

Considere o seguinte exemplo:

(4.33)

a. O ônibus chegou à rodoviária.

b. O motorista conversou com o cobrador.

c. Joana saiu de casa.

A interpretação da frase (4.33b), tomando por contexto a frase (4.33a), produz o segmento básico (4.28), explicado anteriormente. O segmento composto resulta da interpretação da frase (4.33c) relativa a este segmento. O resultado da interpretação é a subárvore da figura 13:

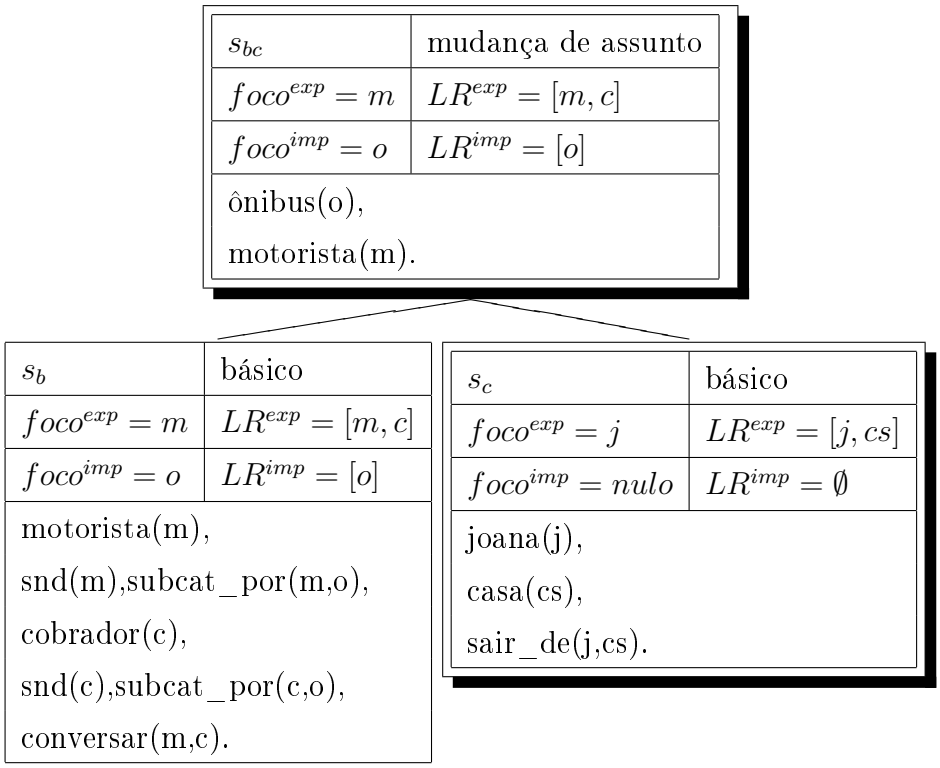


Figura 13: Subárvore resultante da interpretação das frases (4.33b) e (4.33c).

4.4.3.3 Segmento do tipo mudança de tópico

Um segmento novo do tipo mudança de tópico ocorre quando seus subsegmentos têm focos implícitos iguais (ou nulos quando for o segmento mais antigo) e focos explícitos diferentes. Isto caracteriza que o transmissor está detalhando as informações sobre entidades de um mesmo assunto. Ver abaixo o exemplo (4.38a). Neste contexto, o segmento composto herda as seguintes informações:

- Foco implícito do segmento composto - $foco^{imp}_{seg}$: focos implícitos iguais indicam continuação do assunto, logo o segmento composto herda o foco em comum (diferente de nulo):

$$foco^{imp}_{seg} = foco^{imp}_{filho} \tag{4.34}$$

- Lista de relevantes implícita do segmento composto - LR^{imp}_{seg}: como o assunto centra-se sobre uma determinada entidade e esta já é o foco implícito, logo não há necessidade da heranças dos elementos da LR implícita, à exceção de $foco^{imp}_{seg}$ que é a cabeça de LR^{imp}_{seg}:

$$LR_{seg}^{imp} = [foco_{filho_{antigo}}^{imp}] \tag{4.35}$$

- Foco explícito do segmento composto - $foco_{seg}^{exp}$: com a mudança de tópico (diferentes focos explícitos), significa que houve uma mudança de entidades no foco local. Para que estas entidades que ficaram no segmento não visível possam ser reutilizadas nas próximas interpretações, deve-se herdar as entidades explícitas, logo:

$$foco_{seg}^{exp} = foco_{filho_{antigo}}^{exp} \tag{4.36}$$

- Lista de relevantes explícita do segmento composto - LR_{seg}^{exp}: idem ao foco explícito.

$$LR_{seg}^{exp} = LR_{filho_{antigo}}^{exp} \tag{4.37}$$

- A expressão lógica e todos os referentes herdados pelo segmento composto estarão acessíveis.

Veja um exemplo:

(4.38)

a. O ônibus chegou à rodoviária.

b. O motorista brigou com o cobrador.

c. Os passageiros logo foram embora.

A interpretação da frase (4.38b) tomando por base a frase (4.38a) produz o seguinte segmento básico (4.28). A interpretação da frase (4.38c) relativa a este segmento produz a subárvore da figura 14:

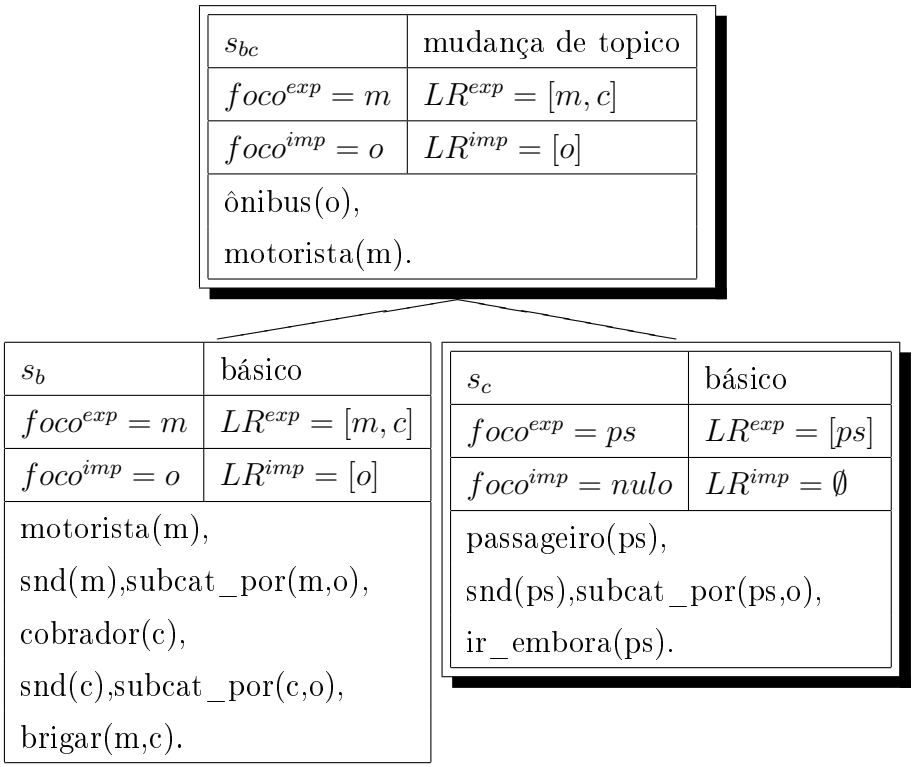


Figura 14: Subárvore resultante da interpretação das frases (4.38b) e (4.38c).

4.4.3.4 Segmento do tipo manutenção de tópico

Diferente dos tipos de segmentos anteriores onde o ponto de referência era o foco implícito[12], aqui não existe ponto de referência: o foco implícito controla a mudança de assunto e o foco explícito acompanha a introdução das informações sobre as entidades deste assunto. Desta maneira há uma *hierarquia* natural entre os focos de forma que um assunto é mais geral que um tópico.

Mas o que acontece quando o assunto muda, mas a entidade sobre a qual o transmissor centra sua atenção continua a mesma? Veja um exemplo:

(4.39) a. O ônibus chegou à rodoviária.

b. O motorista abriu as portas.

c. (e) Φ viu sua filha.

Note que o segmento básico resultante da interpretação em contexto da frase (4.39b) terá como focos: $foco_b^{imp} = onibus$, $foco_b^{exp} = motorista$. Usando este segmento como

[12]Note que quando o foco implícito era constante, havia a mudança ou não do tópico determinando a herança do novo segmento. Quando muito havia a variação conjunta de ambos os focos.

ponto de interpretação para a frase (4.39c), o segmento básico resultante terá como focos: $foco_b^{imp} = nulo$, $foco_b^{exp} = motorista$. Logo o foco implícito foi alterado, mas o foco explícito não. Porém o texto está coerente. Isto acontece porque ambos os focos contribuem para a coesão do texto. A não utilização de SNDs indica apenas que o texto a partir daquele ponto é mais coeso, não necessitando da utilização de recursos mais informativos tais como SNDs [Donnellan 1966].

Assim, o segmento composto não necessitará das informações das entidades explícitas, apenas herdará o assunto anterior possibilitando futuras referências via anáforas (segmento visíveis). O resultado é:

- Foco implícito do segmento composto - $foco_{seg}^{imp}$: será herdado o foco implícito do segmento mais antigo:

$$foco_{seg}^{imp} = foco_{filho_{antigo}}^{imp} \tag{4.40}$$

- Lista de relevantes implícita do segmento composto - LR_{seg}^{imp}: visando a dar uma maior amplitude à retomada de assunto posteriormente, a lista de relevantes implícita deve conter todos os elementos do segmento antigo:

$$LR_{seg}^{imp} = LR_{filho_{antigo}}^{imp} \tag{4.41}$$

- Foco explícito do segmento composto - $foco_{seg}^{exp}$ e lista de relevantes explícita do segmento composto - LR_{seg}^{exp}: não são herdados, pois não houve a mudança de tópico.

$$foco_{seg}^{exp} = nulo \tag{4.42}$$

$$LR_{seg}^{exp} = \emptyset \tag{4.43}$$

- Condições DRS - somente são herdadas aquelas cujos referentes foram introduzidos por entidades implícitas.

Construindo a herança para o exemplo (4.39), especificamente a interpretação da frase (4.39c) em relação ao segmento resultante da interpretação em contexto da frase (4.39b) resulta na subárvore da figura 15:

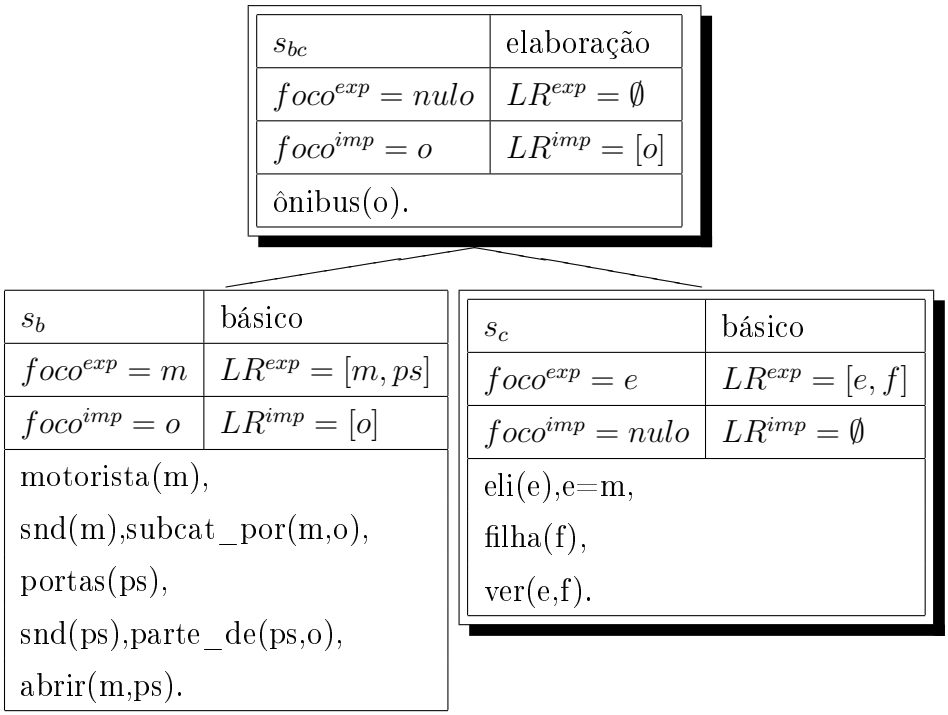


Figura 15: Subárvore resultante da interpretação das frases (4.39b) e (4.39c).

4.4.4 Reagrupamento de segmentos

Considerados os quatro tipos de segmentos que são compostos a partir da herança de atributos dos seus subsegmentos, agora são apresentados os critérios para reagrupamento destes segmentos em *macrosegmentos*. A obtenção de um macrosegmento é feita *offline*, ou seja, após a inserção do segmento resultante da interpretação da frase corrente na END e antes da interpretação da próxima frase.

Cada tipo de segmento define uma relação nominal entre os seus subsegmentos a partir do acompanhamento dos focos. Sendo o foco implícito um indicador do assunto e o foco explícito um indicador de tópico, todas as relações que mantenham um ou outro são indicadores de grupo. Usando estes critérios são criadas as seguintes regras para o reagrupamento de segmentos:

Regra 1 *Uma subárvore com segmentos visíveis, adjacentes e contíguos do tipo elaboração são agrupados num macrosegmento do tipo elaboração, onde um único segmento pai tem como subsegmentos os filhos não visíveis das elaborações originais. A ordem destes filhos no macrosegmento será dada pela profundidade do segmento filho na subárvore original.*

A herança de atributos do novo macrosegmento tipo *elaboração* é descrita a seguir:

- Os focos implícitos e explícitos do novo macrosegmento, respectivamente $foco^{imp}_{ms\_elab}$ e $foco^{exp}_{ms\_elab}$, são os focos do segmento do tipo *elaboração* de menor profundidade na subárvore ($foco^{imp}_{pai}$ e $foco^{exp}_{pai}$). Esta manutenção possibilita a continuação do assunto do discurso em torno de uma mesma entidade:

$$foco^{imp}_{ms\_elab} = foco^{imp}_{pai} \tag{4.44}$$

$$foco^{exp}_{ms\_elab} = foco^{exp}_{pai} \tag{4.45}$$

- A lista de relevantes implícita do macrosegmento $LR^{imp}_{ms\_elab}$ será composta apenas pelo foco implícito do segmento mais à esquerda, limitando assim a reinserção de assuntos marginais ao texto (entidades já tenham sido algures citadas no discurso):

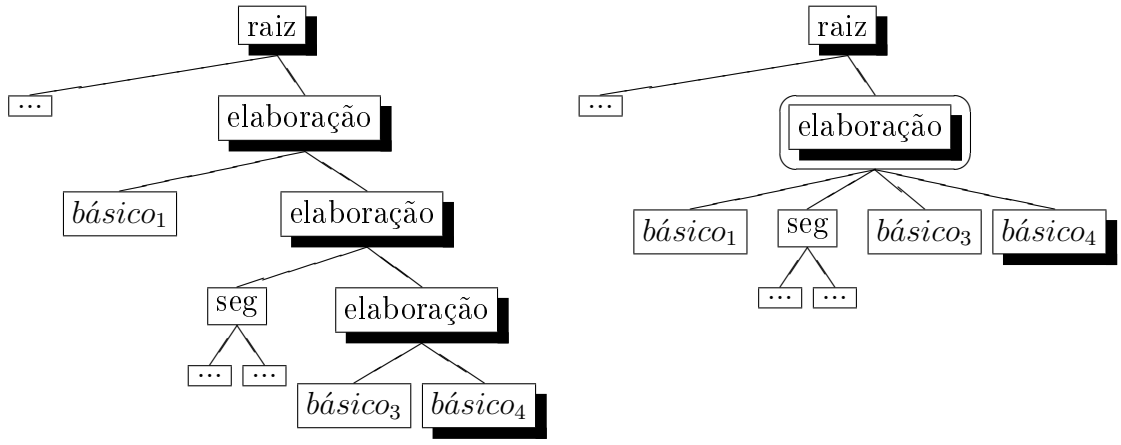
$$LR^{imp}_{ms\_elab} = [foco^{imp}_{ms\_elab}] \tag{4.46}$$

- A lista de relevantes explícita do macrosegmento $LR^{exp}_{ms\_elab}$ será a composição do foco explícito do segmento mais profundo, $foco^{exp}_{ms\_elab}$, com a lista de relevantes explícita do penúltimo segmento mais profundo, $LR^{exp}_{filho2_{ms\_elab}}$. Desta forma procura-se preservar não só o assunto do texto, mas também a possibilidade de recuperação local de entidades fora de foco:

$$LR^{exp}_{ms\_elab} = [foco^{exp}_{ms\_elab}\,,\, LR^{exp}_{filho2_{ms\_elab}}] \tag{4.47}$$

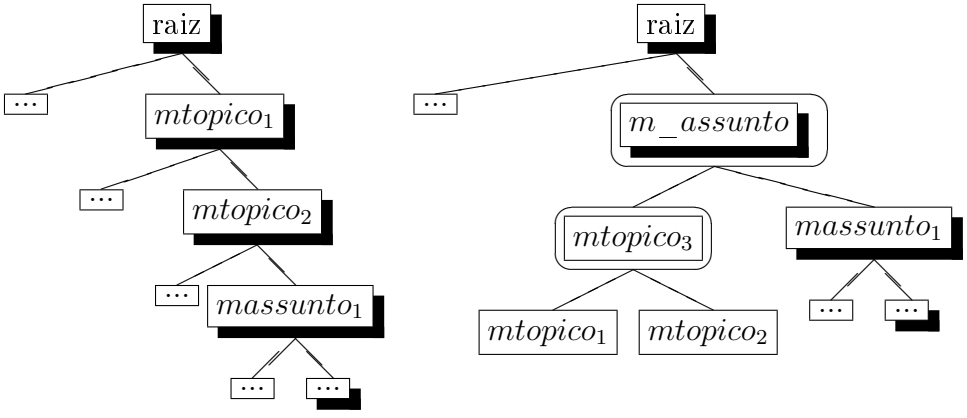
Um exemplo de aplicação da regra (1) é a subárvore da figura 16*(a)* a qual é reduzida à subárvore da figura 16(b).

Regra 2 *Uma subárvore com segmentos visíveis, adjacentes e contíguos do tipo* mudança de tópico *podem ser reagrupados em dois novos macrosegmentos quando o último segmento visível desta subárvore for seguido de um segmento visível do tipo* mudança de assunto. *O primeiro macrosegmento criado é do tipo* mudança de tópico *e agrupa todos os outros do mesmo tipo. O segundo macrosegmento tem dois filhos: o filho da esquerda é o macrosegmento criado anteriormente e o filho da direita é o segmento* mudança de assunto *que originalmente delimitava a subárvore.*



a) END sem agrupamento b) agrupamento de elaborações

Figura 16: Formação de um macrosegmento de elaborações.



a) END sem agrupamento b) agrupamento de mudança de tópico

Figura 17: Formação de um macrosegmento de mudança de tópico.

A figura 17 ilustra a aplicação da regra (2).

O macrosegmento do tipo *mudança de tópico* ($mtopico_3$ da figura 17) tem como herança de atributos:

- O foco implícito do macrosegmento $foco^{imp}_{ms\_mt}$ é o foco implícito do segmento de tipo *mudança de tópico* de maior profundidade na subárvore. Com isto o assunto do trecho de discurso coberto pela subárvore vai ser mantido:

$$foco^{imp}_{ms\_mt} = foco^{imp}_{ultimo\_filho} \tag{4.48}$$

- A lista de relevantes implícita do macrosegmento $LR^{imp}_{ms\_mt}$ é a lista resultante da união do $foco^{imp}_{ms\_mt}$ com todas as LRs implícitas dos segmentos de tipo *mudança de tópico* da subárvore. Os elementos repetidos de $LR^{imp}_{ms\_mt}$ são excluídos. À exceção da cabeça de $LR^{imp}_{ms\_mt}$ que é o $foco^{imp}_{ms\_mt}$, os outros elementos são ordenados na ordem inversamente proporcional à profundidade de seu segmento original ($\biguplus_{ordenada}$). Assim os assuntos que serão obscurecidos podem ser recuperados posteriormente:

$$LR^{imp}_{ms\_mt} = [foco^{imp}_{ms\_mt}, \biguplus_{ordenada} LR^{imp}_{segs\_filhos}] \tag{4.49}$$

- A lista de relevantes explícita $LR^{exp}_{ms\_mt}$ e o foco explícito $foco^{exp}_{ms\_mt}$ do macrosegmento são respectivamente: vazia e nulo. Com a mudança de assunto todas as entidades do assunto anterior devem ser desativadas:

$$LR^{exp}_{ms\_mt} = \emptyset \tag{4.50}$$

$$foco^{exp}_{ms\_mt} = nulo \tag{4.51}$$

O macrosegmento do tipo *mudança de assunto* (*massunto* da figura 17) tem a herança de um segmento deste tipo (seção 4.4.3.2):

- Ambos, a lista de relevantes implícita e o foco implícito são herdados do macrosegmento *mudança de tópico* criado anteriormente.

- Ambos, a lista de relevantes explícita e o foco explícito também são herdados do macrosegmento *mudança de tópico.*

4.4.5 A END e a interpretação de anáforas

A Estrutura Nominal do Discurso proporciona a organização do espaço de busca por um antecedente durante o processo de resolução de anáforas.

Considere um discurso D formado por n frases, tal que: $D = f_1, f_2, ..., f_{i-1}, f_i, ..., f_n$, os passos para a construção da END e a interpretação de uma anáfora são:

1. A END antes da frase f_1 é vazia.
2. Feita a interpretação fora de contexto de f_1, a representação semântica obtida (DRS) é então transformada num segmento básico, através da incorporação dos atributos: focos e listas de relevantes (vide seção 4.4.1). Este segmento passa a ser a END.
3. No restante das frases f_i,com $i \geq 2$, aplicam-se:
   (a) O segmento visível de maior profundidade é escolhido como ponto de interpretação PI na END.
   (b) É feita a interpretação fora de contexto da frase f_i, obtendo-se a interpretação $K_i^{parcial}$, onde os referentes introduzidos por pronomes, elipses e SNDs estão marcados, respectivamente, pelas condições semânticas: $pro(Ref)$ ou $eli(Ref)$ ou $snd(Ref)$.
   (c) A presença de pelo menos uma destas condições em $K_i^{parcial}$ indica a provável existência de entidades cujas referências são externas à própria frase. Neste caso o processo de interpretação anafórica deve ser disparado. Caso contrário $K_i^{parcial}$ é transformada em K_i (segmento básico) e inserida na END (vá para o passo f).
   (d) Ao se fixar um ponto de interpretação na END, o processo de escolha do antecedente T passa a ser o definido na seção 4.3.6. Assim, *todos* os elementos da lista de entidades anafóricas $LA_i = [a_i^1, a_i^2, ..., a_i^m]$ presentes em $K_i^{parcial}$ devem ser resolvidos no ponto PI usando este processo.
   (e) Caso pelo menos um dos elementos de LA_i não possa ser interpretado no ponto PI, faça:
      i. Se PI atual for o segmento de maior profundidade da END, então todas as entidades não resolvidas são armazenadas em LEN_i (Lista de Entidades Não Resolvidas) e todas as entidades resolvidas (com seus respectivos antecedentes e relações) são armazenadas em LER_i (Lista de Entidade Resolvidas).

ii. Caminha-se em profundidade inversa sobre END a partir de PI buscando o próximo segmento visível. Caso não exista um próximo PI (o atual já é a *raiz*) então o processo de interpretação falhou: os elementos em LER_i devem ser incorporados a K_i, apresentando as entidades passíveis de interpretação, e os elementos em LEN_i são acomodadas em K_i (são tratados como indefinidos).

(f) É feita a reconstrução da END baseada no ponto PI resultante dos passos anteriores e no segmento básico K_i resultante da interpretação de $K_i^{parcial}$ em PI. O processo de incorporação de K_i em PI é o descrito na seção 4.4.3.

(g) Os nós da END são então reagrupados de acordo com as regras descritas na seção 4.4.4. Este reagrupamento permite a otimização do processo de interpretação das próximas frases.

Por fim, veja um exemplo de aplicação da END na resolução de anáforas do texto a seguir:

(4.52) a. O ônibus chegou à rodoviária.

b. O motorista abriu as portas.

c. Os passageiros desceram pela porta de trás.

A interpretação fora de contexto da frase (4.52a) produz a seguinte DRS:

$$K_1^{parcial} = \begin{array}{|l|}\hline \text{o,r} \\ \hline \text{ônibus(o),snd(o),} \\ \text{rodoviária(r),snd(r),} \\ \text{chegar(o,r).} \\ \hline \end{array} \tag{4.53}$$

Como esta é a primeira frase, logo a END está vazia. Portanto não existe contexto para a interpretação da mesma, resultando na acomodação de todos os referentes que tenham condições anafóricas, no caso: $snd(o)$ e $snd(r)$. Então $K_1^{parcial}$ é transformada no segmento básico K_1, que também será a END.

$$K_1 = \begin{array}{|l|l|}\hline s_1 & básico \\ \hline foco_1^{exp} = o & LR_1^{exp} = [o, r] \\ foco_1^{imp} = nulo & LR_1^{imp} = \emptyset \\ \hline \multicolumn{2}{|l|}{\text{ônibus(o)},} \\ \multicolumn{2}{|l|}{snd(o), acomod(o),} \\ \multicolumn{2}{|l|}{\text{rodoviária(r)},} \\ \multicolumn{2}{|l|}{snd(r), acomod(r),} \\ \multicolumn{2}{|l|}{\text{chegar(o,r)}.} \\ \hline \end{array} \tag{4.54}$$

A interpretação fora de contexto da frase (4.52b) produz a seguinte DRS:

$$K_2^{parcial} = \begin{array}{|l|}\hline \text{m,ps} \\ \hline \text{motorista(m),snd(m)}, \\ \text{portas(ps),snd(ps)}, \\ \text{abrir(m,ps)}. \\ \hline \end{array} \tag{4.55}$$

A existência de condições *snd* em $K_2^{parcial}$ provoca a necessidade de interpretação anafórica. O ponto de interpretação será $PI = END = K_1$. É então aplicado o processo de seleção de antecedentes baseado nos focos (apresentado na seção 4.3.6) a cada uma das entidades anafóricas de $K_2^{parcial}$. Como o foco implícito de PI é nulo, será usada a lista de relevantes explícitas de PI. Para $snd(m)$ é selecionada a entidade melhor classificada em LR_1^{exp}, que é o ônibus[13] (o). O mesmo vai ocorrer com as portas. No primeiro caso o motorista é subcategorizado pela existência do ônibus. No caso das portas, estas são parte do ônibus. Como resultado, tem-se K_2:

$$K_2 = \begin{array}{|l|l|}\hline s_2 & básico \\ \hline foco^{exp} = m & LR^{exp} = [m, ps] \\ foco^{imp} = o & LR^{imp} = [o] \\ \hline \multicolumn{2}{|l|}{\text{motorista(m)},} \\ \multicolumn{2}{|l|}{snd(m), subcat\_por(m, o),} \\ \multicolumn{2}{|l|}{\text{portas(ps)},} \\ \multicolumn{2}{|l|}{snd(ps), parte\_de(ps, o),} \\ \multicolumn{2}{|l|}{\text{abrir(m,ps)}.} \\ \hline \end{array} \tag{4.56}$$

[13] Aqui não estão sendo consideradas as condições de validação da escolha. Como já visto *parcialmente* no capítulo 3, isto é uma questão de encontrar-se uma relação *válida* entre o motorista e o ônibus.

Neste momento esta observação é pertinente porque encontrado um antecedente e a relação R entre este e sua expressão anafórica A é necessário ainda que a inserção deste no contexto global mantenha o conjunto coerente. Uma relação válida somente entre a expressão anafórica e seu antecedente que seja incompatível com o restante da interpretação não deve ser considerada. Esta e outras questões relacionadas são abordadas no capítulo 5.

Em seguida a END deve ser atualizada com K_2 aplicada sobre PI:

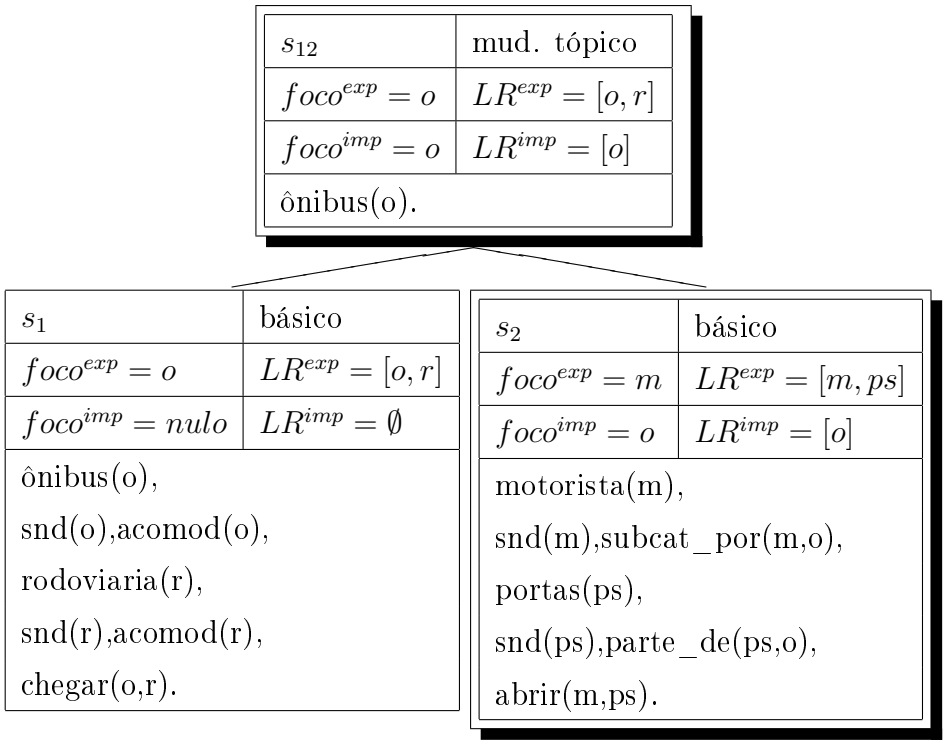


Figura 18: END resultante da interpretação das frases (4.52a) e (4.52b).

A interpretação fora de contexto da frase (4.52c) produz $K_3^{parcial}$:

$$K_3^{parcial} = \begin{array}{|l|}\hline \text{pg,pt} \\ \hline \text{passageiros(pg),snd(pg),} \\ \text{porta\_trás(pt),snd(pt),} \\ \text{descer(pg,pt).} \\ \hline \end{array} \tag{4.57}$$

A existência de condições *snd* em $K_3^{parcial}$ provoca a necessidade de interpretação anafórica. Os pontos de interpretação possíveis de acordo com a END da figura 18 são os segmentos visíveis: s_{12} e s_2. Este último é o escolhido por ser o de maior profundidade (mais recente), logo $PI = s_2$. Aplicando-se o processo de resolução[14] nas entidades anafóricas de $K_3^{parcial}$ tem-se K_3:

[14]Note que existe outra opção para a interpretação do SND *a porta de trás*. Em vez desta ser *parte do* ônibus pode ser *membro do* conjunto de portas introduzido na segunda frase. A solução seria gerar duas interpretações disjuntas. A solução definida neste capítulo não considera esta possibilidade. Porém a metodologia apresentada no capítulo 5 detalha esta solução.

$K_3 =$

s_3	*básico*
$foco^{exp} = pg$	$LR^{exp} = [pg, pt]$
$foco^{imp} = o$	$LR^{imp} = [o]$
passageiros(pg), *snd(pg),subcat_por(pg,o),* porta_trás(pt), *snd(pt),parte_ de(pt,o),* descer(pg,pt).	

(4.58)

Não houve falha de interpretação de $K_3^{parcial}$, portanto é no ponto de interpretação PI= s_2 que K_3 será inserida (figura 19):

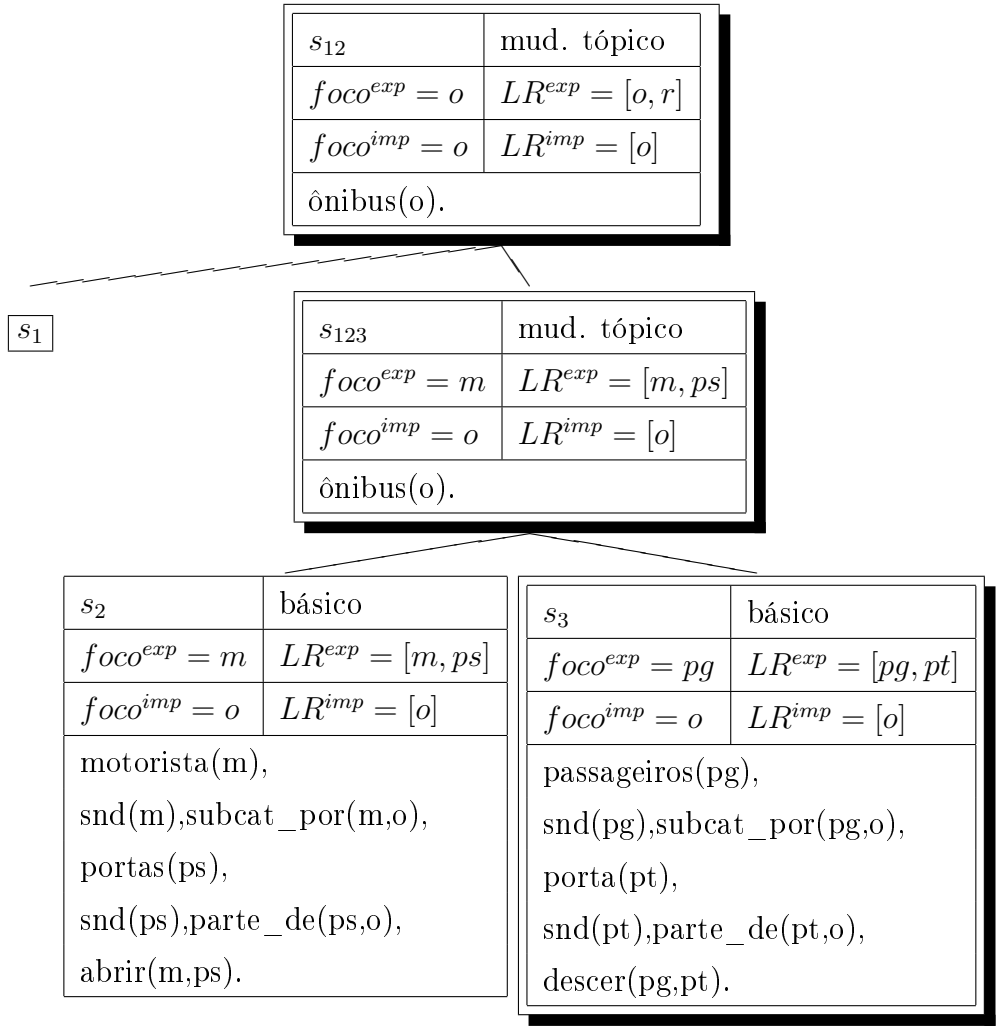


Figura 19: END resultante da interpretação das frases (4.52a) e (4.52b).

5 O protótipo

"*No universo nada se cria,*
nada se perde,
tudo se transforma."
Antonie Lavoisier

Neste capítulo é apresentado o protótipo implementado que permite a obtenção *conjunta* do antecedente $\mathcal{T}$ e da relação $\mathcal{R}$ da fórmula $R(\mathcal{A}, \mathcal{T})$. Para tal, são utilizadas a representação semântica e as regras pragmáticas introduzidas no capítulo 3 bem como a estrutura nominal do discurso descrita no capítulo 4, as quais são implementadas num ambiente de programação em lógica. O ambiente desenvolvido, além de gerar restrições na escolha de $\mathcal{R}$ dado um valor de $\mathcal{T}$ e vice-versa, também cria interpretações alternativas para o discurso baseadas nas frases que contenham entidades anafóricas.

5.1 Introdução

Considere um discurso D formado pelas frases $f_1, f_2, ..., f_i, ..., f_n$. A interpretação de cada frase f_i, fora de contexto, produz uma DRS D e pode incluir em suas condições as expressões lógicas: $snd(Ref)$, $pro(Ref)$ e $eli(Ref)$, indicando, respectivamente, que o referente Ref (pertencente ao universo de D) é um sintagma nominal definido, um pronome ou uma elipse. Em qualquer dos casos, a existência de tais condições em D indica que Ref representa uma entidade anafórica $\mathcal{A}$, cuja interpretação necessita da identificação de um antecedente $\mathcal{T}$ e também de uma relação $\mathcal{R}$ entre $\mathcal{A}$ e $\mathcal{T}$ que, no caso dos SNDs, dão origem à já apresentada equação: $\mathcal{R}(\mathcal{A}, \mathcal{T})$.

A identificação de $\mathcal{R}$ conhecido $\mathcal{T}$ foi feita no capítulo 3, onde foi definido um conjunto de regras pragmáticas que permite estabelecer relações entre $\mathcal{A}$ e $\mathcal{T}$, caso exista uma condição $snd(\mathcal{A})$ em D.

Já a identificação de $\mathcal{T}$ isoladamente, i.e., sem considerar uma possível relação $\mathcal{R}$, foi feita no capítulo 4, onde $\mathcal{T}$ é selecionado do conjunto das entidades previamente introduzidas no discurso, caso exista em D qualquer das condições $snd(\mathcal{A})$, $pro(\mathcal{A})$ ou $eli(\mathcal{A})$. Esta seleção é otimizada com o auxílio da Estrutura Nominal do Discurso, que limita o espaço de procura de $\mathcal{T}$ às entidades contidas num segmento visível e identificado como ponto de interpretação na END num dado instante. $\mathcal{T}$ é então selecionado seguindo os critérios de saliência apresentados na seção 4.3.6.

Finalmente, a identificação conjunta de $\mathcal{T}$ e $\mathcal{R}$ leva em consideração a interferência mútua entre ambos na interpretação de uma frase f_i, tendo como referência um segmento visível PI na END: sem $\mathcal{T}$ não é possível identificar uma relação $\mathcal{R}$ e, caso o $\mathcal{T}$ identificado não possa estabelecer uma relação com $\mathcal{A}$, então é necessário identificar outro $\mathcal{T}_1$ em PI. No caso de não haver mais nenhum antecedente possível em PI, será escolhido um próximo segmento visível na END e repetido o processo. Caso PI já seja o segmento pai então é procurado o segmento visível de maior profundidade na END que contenha o menor número de anáforas não resolvidas. As anáforas não resolvidas neste ponto são acomodadas (i.e. inseridas) na expressão lógica do segmento gerado.

A implementação que permite obter $\mathcal{T}$ e $\mathcal{R}$ é dividida em duas fases:

- Primeiramente, a interpretação pragmática de uma frase é feita provando por abdução a fórmula lógica desta na base de conhecimento com os fatos extraídos do discurso [Hobbs et al. 1993] e presentes apenas nos segmentos vi-

síveis da END (interpretação nominal). O esquema abdutivo de Eshghi e Kowalski [Eshghi e Kowalski 1989] é aqui utilizado na identificação das relações.

- A seguir, a representação semântica da frase interpretada é adicionada aos fatos extraídos do discurso como um todo, resultando numa nova *base de conhecimento*. Aqui é utilizado o sistema de remoção de contradições para programas em lógica estendida proposto por Pereira et al [Pereira, Damásio e Alferes 1993], permitindo assim tirar proveito dos métodos de derivação existentes em programa em lógica estendido e obter uma semântica bem definida para esta base de conhecimento.

Por fim, a implementação descrita acima foi feita utilizando o REVISE [Damásio, Nejdl e Pereira 1994], um sistema de programação em lógica para a revisão de bases de conhecimento construído em Prolog. O apêndice B apresenta a utilização do REVISE.

5.2 A especificação do sistema

A interpretação pragmática das entidades anafóricas de uma frase é feita de modo a ancorá-la no contexto representado pela estrutura nominal do discurso. A entrada para a interpretação é a fórmula lógica que representa a frase fora de contexto. O resultado da interpretação é composto por uma expressão a ser adicionada à representação semântica e uma nova END atualizada com o segmento que representa a nova frase.

A interpretação pragmática de uma frase é feita provando por abdução [Hobbs et al. 1993] a fórmula lógica da frase numa base de conhecimento com os fatos extraídos do texto. A expressão inferida é adicionada à base de conhecimento que contém os fatos do texto e torna a fórmula lógica da frase uma conseqüência dessa base. Essa expressão é a justificação da fórmula lógica da frase.

São utilizadas duas bases de conhecimento: kb_{texto} e kb_{int}. kb_{texto} é a base de conhecimentos com os fatos do texto e kb_{int} é a base usada para a interpretação das possíveis anáforas de uma frase.

A base de conhecimentos kb_{texto} contém todos os fatos extraídos do texto, o que abrange a representação lógica das frases, sua ancoragem na END e as regras que permitem modelar conhecimento dependente ou independente do domínio da aplicação.

A END fornece o contexto para a interpretação de uma frase em relação às anteriormente interpretadas. Ela é utilizada pela base de conhecimentos kb_{int}. kb_{int} é empregada

na interpretação das anáforas de uma frase. Nela se aplica a abdução para provar um determinado SND segundo as regras pragmáticas da seção 3.3.

A utilização de kb_{int} também permite tornar mais leve o processo de interpretação de anáforas pois ao restringir os pontos de interpretação (PI) a apenas os segmentos visíveis, reduz-se o espaço de busca por antecedentes e possíveis relações, controlando o caráter explosivo da abdução.

5.2.1 O processo de interpretação de uma frase

A interpretação de uma frase f_i qualquer, considerando o contexto dado pela END, é feita da seguinte forma:

1. Provar os predicados $snd(Ref)$, $pro(Ref)$ e $eli(Ref)$ de f_i na kb_{int}.

   Na fórmula lógica da frase fora de contexto nem os antecedentes nem as relações estão identificadas. A frase será ancorada em relação a um segmento visível da END. Um conjunto de literais é abduzido e a conjunção destes literais é adicionada à fórmula lógica da frase como explicação para a interpretação das anáforas.

2. Inserir o segmento da nova frase na END e aplicar a reconstrução caso possível.

   Após obtida uma interpretação, acompanhada da ancoragem de um novo segmento na END (vide seção 4.4.3), pode-se então aplicar as regras para reagrupamento de segmentos (seção 4.4.4). Isto elimina as cadeias de segmentos repetidos, tornando a estrutura mais compacta, i.e., com uma profundidade menor.

3. Adicionar a representação semântica da frase na kb_{texto} e verificar se a kb_{texto} atualizada é consistente.

   Se a END é consistente, a expressão lógica que representa a frase, atualizada com as relações entre entidades (os literais abduzidos), é adicionada à kb_{texto}. Como a kb_{texto} possui axiomas independentes do domínio da aplicação e restrições de integridade que não estão na kb_{int}, a atualização da kb_{texto} pode gerar uma base que não contenha modelos preferidos. Se este for o caso, a interpretação é descartada, pois a existência de modelos preferidos é condição para que a interpretação da nova frase seja válida.

A implementação da kb_{texto} utiliza o sistema para remoção de contradições para programas em lógica estendidos proposto por Pereira et al. [Pereira, Damásio e Alferes 1993].

5.2.2 Sistema para remoção de contradições

O sistema utiliza a remoção de contradições de Pereira et al. [Pereira, Damásio e Alferes 1993] para programas em lógica estendidos com negação explícita. Isto é feito porque se adota a semântica bem fundada para programas em lógica estendidos, o que pode levar à inexistência de modelos. O sistema de revisões funciona relaxando as suposições de mundo fechado para remover as contradições que implicam a inexistência de modelos consistentes.

Um programa em lógica estendido [Pereira, Damásio e Alferes 1993] é um conjunto de regras e de restrições de integridade da forma:

$$H \leftarrow B_1, ..., B_n, \, not\, C_1, ..., \, not\, C_m \;\; (m \geq 0, \, n \geq 0) \tag{5.1}$$

onde H, $B_1, ..., B_n$, $C_1, ..., C_m$ são literais objetivo, i. e., um átomo A ou sua negação explícita $\neg A$. Nas restrições de integridade, H é $\bot$ (contradição). Como resultado a restrição de integridade da equação (5.1) também pode ser escrita como:

$$\leftarrow B_1, ..., B_n, \, not\, C_1, ..., \, not\, C_m \;\; (m \geq 0, \, n \geq 0) \tag{5.2}$$

Um sistema de remoção de contradições é um conjunto (P, I, R), onde P é um programa em lógica, I é um conjunto de restrições de integridade e R é um conjunto de revisíveis.

Os revisíveis de um programa P são os elementos de um subconjunto do conjunto de todos os literais na forma $not\, L$ que não possuem regras para L em P, ou seja, verdadeiros por CWA (Closed World Assumption).

Um programa P é contraditório se e somente se não existe uma $WFSX(P)$, onde $WFSX(P)$ é a semântica bem fundada para programas em lógica com negação explícita [Pereira e Alferes 1992].

Já o conjunto (P, I, R) é contraditório se e somente se:

- P é contraditório, ou
- existe alguma restrição de integridade da forma $\leftarrow L_1, ..., L_n$ e $\{L_1, ..., L_n\} \in WFSX(P)$

$P \cup S$ é revisão de (P, I, R) se e somente se:

- $S \subset R$, e

- $(P \cup S,\ I,\ R)$ não é contraditório.

S é revisão mínima de $(P \cup S,\ I,\ R)$ se e somente se:

- $P \cup S$ é revisão de (P, I, R), e
- $\nexists S' \subset S.\ P \cup S'$ é revisão.

Este sistema é utilizado para a implementação da kb_{texto} porque permite lidar de forma simples com o problema dos modelos preferidos não serem únicos: para um dado par $\mathcal{A}$ e $\mathcal{T}$ pode haver mais de uma relação $\mathcal{R}$ possível.

5.2.3 A base de conhecimentos com os fatos do texto (kb_{texto})

A estrutura interna da base de conhecimentos kb_{texto} é composta por três teorias, seguindo o modelo originalmente elaborado por Rodrigues [Rodrigues 1995] para a interpretação temporal do texto:

- TC, é a teoria que define o cenário.

  A teoria do cenário contém a representação em lógica dos fatos do texto. É a união do resultado da interpretação das frases.

  Por exemplo, a frase (5.3) pode ser representada[1] por (5.4).

  (5.3) Lucas comprou as flores.

$$\begin{aligned} &\mathit{indivíduo}(u_1),\ \mathit{nome}(\mathit{lucas},\ u_1),\ \mathit{lexico}(u_1, \mathit{sing}, \mathit{masc}, n),\ \mathit{sujeito}(u_1), \\ &\mathit{entidade}(u_2),\ \mathit{nome}(\mathit{flor}, u_2),\ \mathit{lexico}(u_2, \mathit{plu}, \mathit{fem}, n),\ \mathit{objeto}(u_2),\ \mathit{snd}(u_2). \end{aligned} \tag{5.4}$$

- TD, a teoria dependente do domínio.

  Contém regras que permitem relacionar as entidades entre si e caracterizá-las segundo o domínio em pauta no discurso. Por exemplo, a regra (5.5) estabelece que a entidade rodoviária não está relacionada com a entidade motorista.

[1] Desconsiderando a interpretação do tempo verbal.

$$nao\_relacionado(U_1, U_2) \leftarrow entidade(U_1), entidade(U_2), nome(rodoviaria, U_1), nome(motorista, U_2). \quad (5.5)$$

- TI, a teoria independente do domínio.

  A TI contém regras que são válidas para qualquer aplicação, independente do domínio. Por exemplo, a regra (5.6) diz que duas entidades diferentes U_1 e U_2 não podem estar relacionadas por *parte_de* se o tamanho de U_1 for maior que o tamanho de U_2.

$$\neg parte\_de(U_1, U_2) \leftarrow entidade(U_1),\ entidade(U_2),\ U_1 \neq U_2,\ tamanho(U_1) > tamanho(U_2). \quad (5.6)$$

A kb_{texto} foi especificada de forma que sua tradução para o programa em lógica fosse tão simples quanto possível. É utilizada uma teoria de tipos para o tratamento das variáveis, as quais são transformadas em literais [Rodrigues 1995]. As condições para que uma interpretação seja válida como programa em lógica são escritas como restrições de integridade. O conjunto dos revisíveis é formado pelas relações tratadas: *parte_de, membro_de* e *subcategorizado_por,* além da pseudo relação *acomodação* utilizada como última opção para a interpretação de um SND.

5.2.4 A base de conhecimentos para interpretação das entidades anafóricas de uma frase (kb_{int})

Em kb_{int}, a interpretação é feita de forma pragmática através da prova da fórmula lógica da frase numa base de conhecimentos com os fatos do texto. Foi utilizado o esquema abdutivo de Eshghi e Kowalski [Eshghi e Kowalski 1989] para a especificação, pois esse esquema permite a escrita das regras em programação em lógica com negação por falha e foi facilmente implementado.

Na interpretação das anáforas nominais definidas, quando se *observa* a existência de uma condição $snd(Ref)$, a explicação é que existe um antecedente T (num segmento visível) relacionado com a entidade representada por Ref. Considerando a END, a prova de um $snd(Ref)$ somente é possível se forem inferidos três componentes:

1. O ponto de interpretação PI o qual é determinado pela END.

2. O antecedente T, inferido a partir dos focos e das listas de entidades relevantes existentes em PI.

3. A relação R, abduzida a partir das regras pragmáticas representadas sob a forma de restrições de integridade.

Devido a existência de múltiplos PI, T e R, só são consideradas como justificações as soluções abdutivas que verificam as restrições de integridade e que forem básicas e minimais. Uma justificação é **básica** se nenhum dos fatos na justificação pode ser explicado pela teoria e é **minimal** se não existe nenhuma justificação que seja subconjunto desta.

5.3 A implementação

A implementação foi feita dentro do ambiente em programação em lógica REVISE (versão 2.3) e implementada em SWI-Prolog (versão 5.5.3) num ambiente GNU/Linux. A grande vantagem da implementação neste conjunto (REVISE/Prolog) é que ambos são declarativos, permitindo assim que a especificação das regras pragmáticas e da representação semântica das frases se confunda com a sua implementação.

A figura 20 apresenta a arquitetura geral da implementação, onde:

1. A entrada do sistema são frases de um texto, uma por vez.

2. A frase é a entrada para a interpretação fora de contexto, que tem como saída a representação semântica da frase. Cabe salientar que apesar dessa etapa constar na arquitetura, ela não foi implementada, pois envolve a análise léxica e sintática de frases irrestritas, o que está fora do escopo da tese. Assim, para o sistema, as frases *têm que ser fornecidas* já no formato semântico de uma DRS (i.e. referentes e condições). Como exemplo "Lucas comprou flores" é colocado no sistema como: $ref(1)$, $ref(2)$, $lucas(1)$, $flor(2)$, $masc(1)$, $sing(1)$, $fem(2)$, $plu(2)$. Cada indivíduo introduz um referente $ref(x)$, onde x é um número inteiro que representa o referente, e uma ou mais condições que utilizem este referente. Obrigatoriamente deverão ser introduzidas as condições dadas pela informação léxica associadas a cada indivíduo: número, gênero e grau, para além das condições *especiais* $snd(x)$, $pro(x)$ e $eli(x)$, quando aplicável.

3. A interpretação fora de contexto vai alimentar a interpretação em contexto, onde todas as condições especiais devem ser provadas na base de conhecimento kb_{int}, a

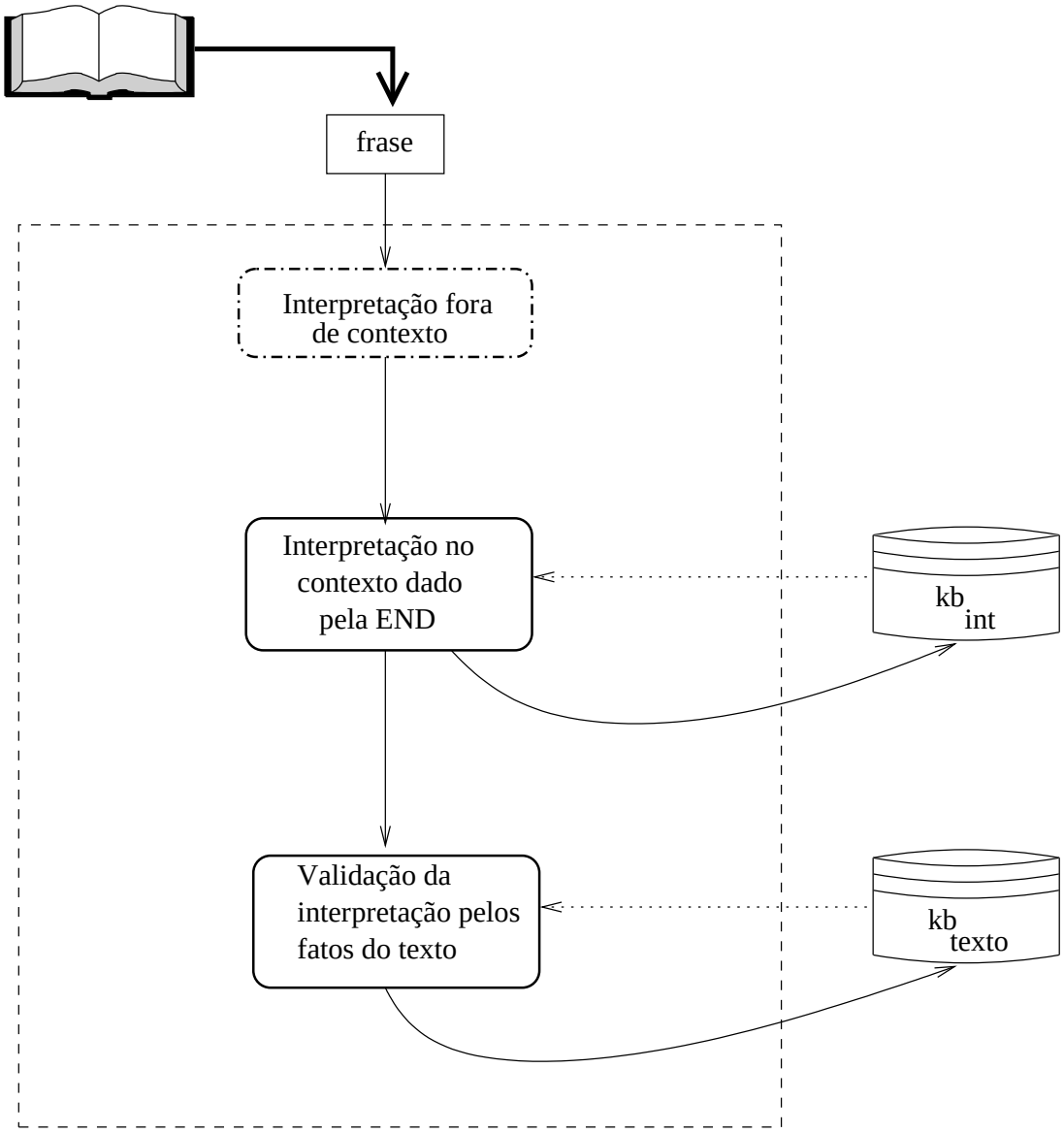


Figura 20: Visão geral da implementação.

qual contém todos os fatos representados nos segmentos visíveis da END. Internamente, kb_{int} está estruturada em segmentos, visto que o ponto de interpretação (PI) também deve ser identificado. A prova de um $snd(x)$ pode gerar mais de uma solução, resultando em mais de um modelo para kb_{int}. Isto ocorre quando existe mais de uma relação possível para um dado par $\mathcal{T}$ e $\mathcal{A}$ ou quando existe mais de um PI que permite a prova de todas as condições especiais existentes na representação da frase. A existência de diversos candidatos a antecedente não introduz soluções alternativas, pois na implementação é considerado que apenas o primeiro antecedente válido é utilizado. A saída desta fase é um conjunto de soluções individuais, onde cada solução é constituída por um referente indicando o segmento visível onde a frase foi interpretada e um conjunto de elementos $(\mathcal{A}, \mathcal{T}, \mathcal{R})$ onde $\mathcal{A}$ é o referente da expressão anafórica, $\mathcal{T}$ é um referente dentro do segmento e $\mathcal{R}$ é uma das relações: *parte_de, membro_de, coreferencia e subcategorizado_por* ou a pseudo relação *acomodação.*

4. A entrada para a fase de validação em kb_{texto} são as soluções da fase anterior. Cada solução é então testada: seus literais são inseridos em kb_{texto}. Caso esta continue coerente então a solução é válida. Soluções que produzam uma base de conhecimento inválida são descartadas. As soluções restantes são classificadas por ordem de preferência (modelos preferidos), armazenadas em kb_{texto} e sincronizadas com kb_{int}.

A seguir é apresentada a implementação das bases de conhecimento kb_{int} e kb_{texto}.

5.3.1 A implementação da kb_{int}

A implementação da kb_{int} é feita no sistema de remoção de contradições apresentado na seção 5.2.2. A tradução para o sistema de remoção de contradições é direta, mantendo-se o programa em lógica e as restrições de integridade. Entretanto, o conjunto dos literais abdutíveis pode ter de ser alterado, pois, no sistema de remoção de contradições, os literais abdutíveis não podem aparecer como cabeça de regras no programa, o que também é proposto por Kakas et al. [Kakas, Kowalski e Toni 1992] para que as soluções sejam todas básicas.

O conjunto dos revisíveis é constituído pelos literais da forma:

- $R'(E_i, E_j)$, onde $R' \in \{parte\_de', membro\_de', subcategorizado\_por', coreferencia'\}$,

E_i é uma entidade pertencente ao PI e E_j é um referente pertencente à representação semântica da frase corrente.

- $R'(E_j)$, onde $R' \in \{acomodacao'\}$ e E_j é um referente pertencente à representação semântica da frase corrente.
- $ponto\_interpretacao(S_i)$, sendo S_i um referente para o ponto de interpretação onde o conjunto de anáforas da frase corrente é interpretada. Note que não existe a regra para $ponto\_interpretacao'(S_i)$, pois $ponto\_interpretacao(S_i)$ nunca aparece como cabeça de regras no programa em lógica.

Pelo fato de as regras terem sido reescritas, o programa contém as seguintes regras de conversão:

$$\begin{aligned} &R(E_i, E_j) \leftarrow R'(E_i, E_j),\ com \\ &\qquad R \in \{parte\_de, membro\_de, subcategorizado\_por, coreferencia\}\ e \\ &\qquad R' \in \{parte\_de', membro\_de', subcategorizado\_por', coreferencia'\}. \\ &R(E_j) \leftarrow R'(E_j),\ com\ R \in \{acomodacao\}\ e\ R' \in \{acomodacao'\}. \end{aligned} \tag{5.7}$$

A especificação e a implementação do programa em lógica não diferem. O programa em lógica é constituído pelas três teorias utilizadas na kb_{texto}: TI, a teoria independente do domínio; TC, a teoria do cenário; e TD, a teoria dependente do domínio.

5.3.1.1 As regras de TI

As regras de TI permitem a inferência da relação R e do ponto de interpretação de uma frase no contexto dado pela END. As regras são da forma:

$$\begin{aligned} parte\_de(E_1, E_2) \leftarrow\ & visível(S_j, E_2), \\ & entidade(E_1),\ entidade(E_2), \\ & not\,(E_1 = E_2), \\ & sing(E_2), \\ & not\,plu(E_1), \\ & ponto\_interpretacao(S_j). \end{aligned} \tag{5.8}$$

A regra (5.8) indica que os seguintes fatos podem ser a justificativa (explicação) para uma entidade anafórica E_1 frente a um antecedente E_2:

- A entidade E_2 pertence a um segmento visível.
- E_1 é ancorado em relação a um segmento visível S_j se as entidades E_1 e E_2 forem ser diferentes.
- E_2 está no singular e E_1 não é plural.
- A relação existente entre as duas entidades vai ser $parte\_de$, quando considerado o segmento S_j como ponto de interpretação.

Existe uma regra como a (5.8) para cada interpretação prevista pela teoria, de acordo com as regras da seção 3.4.1.

5.3.1.2 As regras de TC

São obtidas a partir da estrutura nominal do discurso para a interpretação da frase e devem conter:

- todos os literais da expressão lógica existente nos **segmentos visíveis.**
- todos os literais da forma $visível(s_i)$, com s_i sendo o atributo **referente** de um segmento visível.
- todos os literais do atributo **expressão** do segmento da frase que vai ser interpretada.

5.3.1.3 As regras de TD

As regras de TD modelam o conhecimento geral sobre as entidades no domínio do discurso. Por exemplo, podem descrever as condições gerais para que uma entidade não esteja relacionada com outras (e.g. tamanhos incompatíveis, locais diferentes, tempos diferentes, etc). Um exemplo de regra é:

$$\begin{aligned} gen\_membro\_de(E_1, E_2) \leftarrow\ & lapis(E_1), \\ & material\_escolar(E_2), \\ & coletivo(E_2). \end{aligned} \tag{5.9}$$

A regra 5.9 introduz as condições gerais para que um lápis seja membro do conjunto de objetos que compõem o material escolar: (1) as duas entidades devem estar presentes no discurso e a entidade "material escolar" deve ser vista como uma entidade coletiva.

5.3.1.4 As restrições de integridade

Derivadas do conjunto de revisíveis, as restrições de integridade são:

$$\begin{aligned}
&\leftarrow membro_de(E_1, E_2),\ not\, gen_membro_de(E_1, E_2).\\
&\leftarrow parte_de(E_1, E_2),\ not\, gen_parte_de(E_1, E_2).\\
&\leftarrow subcategorizado_por(E_1, E_2),\ not\, gen_subcategorizado_por(E_1, E_2).\\
&\leftarrow coreferencia(E_1, E_2),\ not\, gen_coreferencia(E_1, E_2).
\end{aligned} \tag{5.10}$$

As restrições de integridade em (5.10) equivalem aos axiomas independentes de domínio da kb_{texto}, apresentados na próxima seção.

5.3.2 A implementação da kb_{texto}

A implementação da kb_{texto} é constituída pelo conjunto (P_{txt}, IC_{txt}, R_{txt}). A tradução da kb_{texto} é relacionada com a própria base da seguinte forma:

- Existe um modelo preferido para a kb_{texto} se e somente se existe uma revisão mínima, S, para a tradução da kb_{texto} e $S \cup TC$ é consistente.
- Existe um modelo preferido para a kb_{texto} que é modelo para um objetivo G se e somente se existe uma revisão mínima, S, para (P_{txt}, $IC_{txt} \cup \{\leftarrow not\, G\}$, R_{txt}) e $S \cup TC$ é consistente.
- $S \cup TC$ é consistente se o conjunto de pares de entidades (anáfora e antecedente) que estão em $S \cup TC$ tem uma interpretação que obedece a todas as relações em $S \cup TC$.

Deste modo a implementação e a especificação da kb_{texto} são relacionadas, em particular os modelos preferidos. Na especificação da kb_{texto}, os modelos preferidos são aqueles em que só existem relações entre entidades anafóricas que não geram contradições em TC (i.e. relações anormais).

Os modelos preferidos são definidos para determinar quando a representação de um conjunto de frases de um texto é consistente, além de permitir uma semântica computa-

cionalmente construtiva, evitando a proliferação de entidades, uma vez que são minimais em relação ao número de entidades.

A tradução para o sistema de remoção de contradições é obtida através da tradução das condições para que uma interpretação seja válida, da TC, da TD e da TI e através da definição do conjunto de revisíveis e de alguns predicados auxiliares. Como a especificação e a implementação são bastante próximas, apenas a última será apresentada.

5.3.2.1 Condições para que uma interpretação seja válida

A tradução das condições para que a interpretação seja válida são:

1. Num modelo preferido cada entidade anafórica tem um e somente um antecedente associado.
   $\leftarrow$ $entidade(E_1)$, $entidade(E_2)$, $entidade(E_3)$,
   $snd(E_1)$, $R(E_1, E_2)$, $not\ R(E_1, E_3)$.

2. Num modelo preferido uma entidade tem uma e somente uma relação com seu antecedente.
   $\leftarrow$ $entidade(E_1)$, $entidade(E_2)$, $snd(E_1)$,
   $R_a(E_1, E_2)$,
   $not\ R_b(E_1, E_2)$.

3. Num modelo preferido uma entidade, supostamente anafórica, não pode ter uma relação com seu antecedente e ser ao mesmo tempo acomodada.
   $\leftarrow$ $entidade(E_1)$, $entidade(E_2)$, $snd(E_1)$,
   $R(E_1, E_2)$,
   $not\ acomodacao(E_1)$.
   ou
   $\leftarrow$ $entidade(E_1)$, $entidade(E_2)$, $snd(E_1)$,
   $acomodacao(E_1)$
   $not\ R(E_1, E_2)$.

4. Num modelo preferido cada entidade tem pelo menos uma proposição léxica associada.

   Esta condição é sempre garantida na tradução de frases em língua natural para expressões da TC.

5. Num modelo preferido sempre que existe uma entidade descrita com suas proposições léxicas, deve-se expandi-la em termos associados.

Sempre que existe uma expressão na forma $lexico(ref, num, gen, grau)$ em TC, deve-se colocar as regras (5.11) em P_{txt}.

$$\begin{array}{l} num(ref). \\ gen(ref). \\ grau(ref). \end{array} \tag{5.11}$$

6. Num modelo preferido a extensão do predicado *membro_de* está contida na extensão do predicado *gen_membro_de*.

$\leftarrow entidade(E_1),\ entidade(E_2),\ membro\_de(E_1,\ E_2),\ not\ gen\_membro\_de(E_1,\ E_2).$

7. Num modelo preferido a extensão do predicado *parte_de* está contida na extensão do predicado *gen_parte_de*.

$\leftarrow entidade(E_1),\ entidade(E_2),\ parte\_de(E_1,\ E_2),\ not\ gen\_parte\_de(E_1,\ E_2).$

8. Num modelo preferido a extensão do predicado *subcategorizado_por* está contida na extensão do predicado *gen_subcategorizado_por*.

$$\begin{array}{ll} \leftarrow & entidade(E_1),\ entidade(E_2), \\ & subcategorizado\_por(E_1,\ E_2),\ not\ gen\_subcategorizado\_por(E_1,\ E_2). \end{array}$$

9. Num modelo preferido a extensão do predicado *coreferencia* está contida na extensão do predicado *gen_coreferencia*.

$\leftarrow entidade(E_1),\ entidade(E_2),\ coreferencia(E_1,\ E_2),\ not\ gen\_coreferencia(E_1,\ E_2).$

10. Num modelo preferido uma relação de *coreferencia* aplicada entre um antecedente e sua anáfora, estabelece restrições sobre a relação entre estas entidades com uma terceira.

$$R_a = R_b \leftarrow entidade(E_1),\ entidade(E_2),\ coreferencia(E_1,\ E_2),\ entidade(E_3),\ R_a(E_1, E_3),\ R_b(E_2, E_3).$$

5.3.2.2 Tradução da teoria do cenário (TC)

TC é a teoria que descreve um cenário, composta pela união do resultado da interpretação das frases de um texto. TC é uma conjunção de símbolos só com constantes. A sua implementação é feita adicionando a teoria de tipos à especificação.

Por exemplo, a tradução da expressão que representa a frase (5.12) na especificação da TC é (5.13) e na implementação é descrita em (5.14).

(5.12) Lucas comprou as flores.

$$lucas(e_1),\ flores(e_2),\ snd(e_2). \tag{5.13}$$

é traduzida em:

$$\begin{array}{l} entidade(e_1). \\ entidade(e_2). \\ lucas(e_1). \\ flor(e_2). \\ snd(e_2). \\ acomodacao(e_2). \\ sing(e_1).masc(e_1). \\ plu(e_2).fem(e_2). \end{array} \tag{5.14}$$

A tradução é feita da seguinte forma:

1. Para cada constante c em TC deve existir a regra

   $entidade(c)$, se $c \in E_c$(conjunto de entidades).
   $sing(c)\, ou\, plu(c)$, $masc(c)\, ou\, fem(c)$, .

2. Para cada termo da forma $snd(e)$ em TC, o programa deve conter as regras :

   $snd(e)$.
   $acomodacao(e)$.
   ou

$R(e, t)$.
onde R é uma relação e t um antecedente.

3. Os outros termos em TC são simplesmente adicionados.

5.3.2.3 A teoria dependente do domínio - TD

A teoria dependente do domínio é um conjunto de fatos que especificam: (1) as condições necessárias para que uma relação possa ser estabelecida e (2) em que condições uma relação é anormal.

$$\begin{aligned} gen\_parte\_de(E_1, E_2) \leftarrow\ & motor(E_1), \\ & carro(E_2), not\, plu(E_2). \end{aligned} \tag{5.15}$$

$$\begin{aligned} anormal(parte\_de, E_1, E_2) \leftarrow\ & motor(E_1), \\ & carro(E_2), \\ & tamanho(E_1) > tamanho(E_2). \end{aligned} \tag{5.16}$$

A tradução desta teoria e da TI para a implementação é realizada através da inclusão da teoria de tipos e da substituição de variáveis.

5.3.2.4 A teoria independente do domínio - TI

TI contém o conhecimento geral que não depende do domínio da aplicação.

As regras apresentadas no início desta seção também são independentes do domínio da aplicação, mas estão separadas da TI porque a linguagem para a especificação de TI não possui expressividade para representar tais regras. Assim, tais regras são especificadas por meio de axiomas.

5.3.2.5 Revisíveis

O conjunto de revisíveis é definido por: $\{parte\_de'(E_1, E_2)$, $membro\_de'(E_1, E_2)$, $subcategorizado\_por(E_1, E_2)$, $coreferencia(E_1, E_2)\}$, com E_1 e E_2 duas entidades.

É necessário usar as relações R' em lugar das relações R, uma vez que os revisíveis não podem ser cabeça de nenhuma regra.

5.3.2.6 Outras regras

O programa deve ainda conter a definição de =:

$$A = A.$$
$$A = B \leftarrow B = A.$$
$$A = B \leftarrow A = C, C = B.$$
$$A = B \leftarrow A =' B$$

Finalmente, para cada predicado P do programa, exceto os predicados $entidade/1$, $snd/1$, $pro/1$ e $eli/1$, deve-se juntar a sua definição:

$$P(X_1, ..., X_n) \leftarrow X_1 = X'_1, ..., X_n = X'_n, P(X'_1, ..., X'_n).$$

5.4 Avaliação do protótipo

O protótipo foi implementado em Prolog. Os tempos de processamento apresentados na tabela 6 são os resultados obtidos executando o protótipo num PC Athlon (i686), 700 MHz, 256MB de RAM, kernel Linux versão 2.6.10 (distribuição Fedora 2), SWI-Prolog versão 5.5.2 (*Multi-threaded*) e REVISE versão 2.3.

Nos testes do mecanismo proposto neste capítulo usaram-se textos cujas frases já estavam representadas na forma de uma DRS, i.e., um conjunto de referentes e as condições a eles associadas. Cada frase é uma interpretação fora de contexto estão marcadas as entidades *possivelmente* anafóricas (via $snd, pro\,e\,eli$). Veja o exemplo:

(5.17) O ônibus chegou à rodoviária.

o qual é representado como:

$$onibus(1). \quad sing(1). \quad masc(1).$$
$$snd(1).$$
$$rodoviaria(2). \quad sing(2). \quad fem(2).$$
$$snd(2).$$

Esta é a entrada para o sistema. Que a partir da interpretação do programa em lógica acima vai acrescentar ainda os seguintes fatos:

$$acomoda(1).$$
$$acomoda(2).$$
$$visivel(3).$$
$$seg(3, [1,2], 1, [1,2], nulo, vazio).$$

indicando que as entidades representadas pelos referentes 1 e 2 foram acomodadas. Elas fazem parte de um segmento identificado pelo referente 3 e representado por $seg(ref, universo, foco^{imp}, LR^{imp}, foco^{exp}, LR^{exp})$, onde LR é a lista de relevantes.

Considerando o contexto gerado pela frase anterior é feita a interpretação da frase (5.18).

(5.18) O motorista abriu as portas.

cuja representação fora de contexto é:

$$\begin{array}{lll} motorista(4). & sing(4). & masc(4). \\ snd(4). & & \\ portas(5). & plu(5). & fem(5). \\ snd(5). & & \end{array} \tag{5.19}$$

A interpretação em contexto de (5.19) insere os seguintes literais:

$$subcategorizado\_por(4, 1).$$
$$parte\_de(5, 1).$$
$$visivel(6).$$
$$seg(6, [4,5], 1, [1], 4, [4,5]).$$

A complexidade da interpretação das entidades anafóricas de uma frase de um texto depende da interpretação em contexto feita em kb_{int} e da prova de consistência feita em kb_{texto}. Considere o seguinte texto:

(5.20) O ônibus chegou à rodoviária.

O motorista abriu as portas.

Os passageiros desceram pela porta de trás.

O motorista saiu pela porta de frente.

Ele foi até o escritório da empresa.

O gerente estava lhe esperando.

Nele o processo de interpretação das entidades anafóricas é dividido em:

1. Provar os predicados anafóricos *snd*, *pro e eli* da frase na kb_{int}.
2. Este passo é proporcional aos seguintes fatores:
   - Número de segmentos visíveis na estrutura nominal do discurso.
   - Número de regras na kb_{int} que podem ser usadas para provar os predicados anafóricos.
   - Número de restrições de integridade e complexidade da sua verificação.
3. Inserir o segmento da nova frase na END.
4. O tempo de inserção do novo segmento na END é constante.
5. O tempo de reorganização da END varia de acordo com o número de segmentos visíveis.
6. Verificar se a nova kb_{texto} é consistente.
   - Verificar se existe uma revisão mínima para o programa que constitui a kb_{texto}.
   - A complexidade desta verificação é sobretudo devido à complexidade de verificar se alguma restrição de integridade é violada.
   - Verificar se o conjunto relações pragmáticas inferidas são consistentes (as que estão no atributo expressão de algum segmento visível mais as revistas para a kb_{texto} verificar as restrições de integridade).

A complexidade de fazer uma prova na base de conhecimentos kb_{int} é menor que a complexidade de determinar se a base de conhecimentos kb_{texto} é consistente, pois a kb_{int} não tem que verificar tantas restrições de integridade como a kb_{texto}. É a verificação das restrições de integridade da kb_{texto} que torna o processo de verificação de consistência tão pesado.

Na tabela 6 apresentam-se os tempos para o processamento das frases do texto (5.20).

Frase	prova na kb_{int}	consistência kb_{texto}	n^o interações calculadas	n^o interações válidas	total
3	2.3s	4.6s	5	2	6.9s
4	3.1s	6.4s	6	2	9.5s
5	3.5s	8.7s	6	3	12.2s
6	5.9s	14.0s	10	5	19.9s
					48.5s

Tabela 6: Tempos para a interpretação de textos.

O item *prova na* kb_{int} exibe os tempos para cálculo de todas as provas dos predicados anafóricos da frase para cada END usada como contexto. No caso da frase 2 este é o tempo para calcular as soluções com a única estrutura, o segmento da frase 1; no caso da frase 3, este tempo é a média do tempo do cálculo das soluções obtidas com a primeira END e do cálculo das soluções obtidas com a segunda estrutura das primeiras duas frases.

O item *consistência da* kb_{texto} é o tempo necessário para encontrar o conjunto das revisões mínimas da kb_{texto}. Este cálculo é feito para determinar se a interpretação da frase é consistente com as frases anteriores.

A conclusão a que se chegou é que a utilização de duas bases de conhecimento no processo de interpretação permitiu melhorar a forma como são verificadas as restrições de integridade e assim diminuir o número de interpretações possíveis para um conjunto de frases.

6 *Considerações finais*

"*Fora da caridade não há salvação.*"
Allan Kardec

Neste capítulo são apresentadas algumas conclusões sobre o trabalho realizado e propostas algumas continuações possíveis para o trabalho.

Nesta tese foi proposta e caracterizada uma metodologia computacional para a interpretação pragmática das anáforas nominais definidas que possibilita também a interpretação de anáforas pronominais e elipses do sujeito. Esta metodologia é um componente do processamento automatizado de textos.

O objetivo da construção deste componente foi obter uma base de conhecimentos onde as entidades anafóricas estivessem resolvidas: localizado o antecedente e, caso necessário, identificadas as relações entre entidades. Estas relações estão presentes no texto de forma implícita, podendo ser identificadas através da interpretação pragmática da informação léxica das entidades relacionadas. Isto foi resumido na equação $\mathcal{R}(\mathcal{A}, \mathcal{T})$, onde dado $\mathcal{A}$, identificam-se $\mathcal{T}$ e $\mathcal{R}$.

A identificação de um antecedente e a interpretação pragmática da informação léxica acrescentam à fórmula lógica de cada frase as relações anafóricas implícitas $\mathcal{R}$ entre a expressão anafórica de uma frase e antecedentes existentes em outras frases anteriores.

Para além da interpretação pragmática, foi necessário a criação de uma estrutura do discurso que fosse capaz de fornecer para cada frase a parte do contexto de interpretação constituída pelas entidades mais salientes no discurso - a Estrutura Nominal do Discurso. A função primária da END foi permitir a restrição do espaço de busca por antecedentes $\mathcal{T}$.

Com relação à identificação da saliência das entidades, foram definidos dois focos do discurso - o foco explícito e o foco implícito. O acompanhamento dos focos permitiu a criação da END e as regras pelas quais um foco pode ser determinado permitiu a ordenação da busca por um antecedente $\mathcal{T}$.

A interpretação pragmática das entidades anafóricas da frase foi feita por abdução seguindo a proposta de [Hobbs et al. 1993]. O caráter explosivo do processo abdutivo é controlado pelo contexto dado pela estrutura nominal do discurso, onde só se procuram as ligações das entidades anafóricas às entidades mais salientes.

Mesmo assim, o autor desta tese pensa ser possível substituir a abdução por outros mecanismos de inferência. A abdução introduz restrições no tempo de computação. Sem ela o processo pode ser consideravelmente reduzido. Ela seria substituída por mecanismos de identificação de relações baseados em CORPUS.

O que, a nosso ver, não é possível substituir é o caráter não monotônico necessário ao raciocínio. Mesmo que uma relação possa ser *deduzida* de uma base de conhecimentos (supondo que todas as premissas lá estejam) ainda é possível que uma relação anteriormente

estabelecida seja revogada quando mais informação for veiculada pelo discurso.

A abdução é um processo *bem humano* e desta forma facilitou a resolução de anáforas nominais definidas no que tange a *encontrar uma hipótese dada uma observação.*

Com relação aos testes realizados, eles foram limitados por dois fatores:

1. falta de dicionários incorporados à maquinaria, em especial a um dicionário de coletivos. Isto permitiria um melhor resultado na identificação das relações através das regras pragmáticas (cap. 3).
2. falta de analisador léxico e sintático incorporado a um conversor semântico para textos irrestritos (CORPORA), o qual seria a entrada para o protótipo do cap. 5.

O autor tem como objetivo contíguo ao fechamento desta tese trabalhar os seguintes tópicos:

- Em textos irrestritos, é necessário verificar a ocorrência de casos em que há a troca de foco implícito mas o foco explícito continua o mesmo. Isto pode ser testado em textos já marcados (CORPORA).
- As relações tratadas nesta tese são de especificação, i.e., as entidades introduzidas pelas expressões anafóricas são mais específicas que seus antecedentes. É necessário estender tais relações para o tratamento de generalizações e estudar qual o comportamento dos focos nestes casos. Neste contexto pretende-se utilizar a taxinomia proposta por Strand [Strand 1996].
- Verificar, em textos irrestritos, a possibilidade da ancoragem parcial. Por exemplo, quando parte das entidades anafóricas de uma frase é ancorada num determinado ponto de interpretação e outra parte num outro ponto. Qual o impacto desta possibilidade na estrutura?
- Utilizar o trabalho desenvolvido em aplicações na área de recuperação de informação, onde três possibilidades são vislumbradas:

    1. geração de resumo dos documentos: a END traduz o acompanhamento dos focos do discurso, herdando valores e resumindo segmentos, o que pode ser usado na geração de um resumo do documento.

2. utilizar a metodologia numa máquina de busca: tanto a geração de um documento virtual onde as anáforas estejam resolvidas quanto a utilização direta da END como índice para máquinas de busca. No apêndice A estas propostas são melhor detalhadas.
3. utilizar a metodologia na classificação de documentos: derivado do proposto no apêndice A, onde são atribuídos pesos à END, é possível criar uma classificação de coleções de documentos.

Por fim, integrar as entidade introduzidas pela interpretação temporal [Rodrigues 1995] e verificar como elas podem introduzir restrições à END e vice versa. Este trabalho faria a integração da estrutura nominal do discurso com a estrutura temporal do discurso.

Referências

[Abbott 1993] ABBOTT, B. A pragmatic account of the definiteness effect in existencial sentences. *Journal of Pragmatics*, v. 19, p. 39–55, 1993.

[Allen 1995] ALLEN, J. F. (Ed.). *Intentions in Communication*. 2. ed. [S.l.]: Benjamin/Cummings Publishing, 1995.

[Alshawi 1990] ALSHAWI, H. Resolving quasi logical forms. *Computational Linguistics*, v. 16, n. 3, p. 133–144, 1990.

[Ariel 1996] ARIEL, M. Referring expressions and the +/- coreference distinction. In: FRETHEIM, T.; GUNDEL, J. K. (Ed.). *Reference and Referent Accessibility*. [S.l.]: John Benjamins Publishing Company, 1996. p. 13–25.

[Asher 1993] ASHER, N. *Reference to Abstract Objects in Discourse*. [S.l.]: Kluwer Academic Publishers, 1993.

[Baeza-Yates e Ribeiro-Neto 1999] BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern information retrieval*. Wokingham, UK: Addison-Wesley, 1999.

[Beaver 2004] BEAVER, D. I. The optimization of discourse anaphora. *Linguistics and Philosophy*, v. 27, p. 3–56, 2004.

[Blutner 2000] BLUTNER, R. Some aspects of optimality in natural language interpretation. *Journal of Semantics*, v. 17, p. 189–217, 2000.

[Bos 2003] BOS, J. Bridging as coercive accommodation. *Computational Linguistics*, v. 29, n. 2, p. 179–210, 2003.

[Bos, Buitelaar e Mineur 1995] BOS, J.; BUITELAAR, P.; MINEUR, A.-M. Bridging as coercive accommodation. In: ET.AL., S. M. (Ed.). *Proceedings of the Workshop on Computational Logic for Natural Language Processing*. South Queensferry, Scotland: [s.n.], 1995. Disponível em: <citeseer.ist.psu.edu/bos95bridging.html>.

[Brennan, Friedman e Pollard 1987] BRENNAN, S. E.; FRIEDMAN, M. W.; POLLARD, C. J. A centering approach to pronouns. In: *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 1987. p. 155–162.

[Brewka, Dix e Konolige 1997] BREWKA, G.; DIX, J.; KONOLIGE, K. *Nonmonotonic Reasoning: An Overview*. Stanford, CA: CLSI Publications, 1997. (Lecture Notes Number 73).

[Carter 1987] CARTER, D. *Interpreting Anaphors in Natural Language Texts*. [S.l.]: Ellis Horwood Books, 1987.

[Chafe 1996] CHAFE, W. Inferring identifiability and acessibility. In: FRETHEIM, T.; GUNDEL, J. K. (Ed.). *Reference and Referent Accessibility*. [S.l.]: John Benjamins Publishing Company, 1996.

[Clark 1977] CLARK, H. Bridging. In: JOHNSON-LAIRD, P. N.; WASON, P. C. (Ed.). *Thinking: Readings in Cognitive Science*. [S.l.]: Cambridge University Press, 1977.

[Cohen 2000] COHEN, A. The king of france is, in fact, bald. *Natural Language Semantics*, n. 8, p. 291–295, 2000.

[Cohen e Erteschik-Shir 2002] COHEN, A.; ERTESCHIK-SHIR, N. Topic, focus, and the interpretation of bare plurals. *Natural Language Semantics*, n. 10, p. 125–165, 2002.

[Cooper 1993] COOPER, R. Generalized quantifers and resource situations. In: ACZEL, P. et al. (Ed.). *Situation Theory and its Applications*. [S.l.]: CSLI and University of Chicago, 1993. v. 3, p. 191–212.

[Cormack 1992] CORMACK, S. *Focus and Discourse Representation Theory*. Tese (Doutorado) — University of Edinburgh, 1992.

[Dagan e Itai 1990] DAGAN, I.; ITAI, A. Automatic acquisition of constraints for the resolution of anaphora and syntactic ambiguities. In: *Proceedings of the International Conference on Computational Linguistics*. [S.l.: s.n.], 1990. v. 3, p. 330–332.

[Dahl e Fraurud 1996] DAHL, Ö.; FRAURUD, K. Animacy in grammar and discourse. In: FRETHEIM, T.; GUNDEL, J. K. (Ed.). *Reference and Referent Accessibility*. [S.l.]: John Benjamins Publishing Company, 1996.

[Damásio, Nejdl e Pereira 1994] DAMáSIO, C. V.; NEJDL, W.; PEREIRA, L. M. REVISE: An extended logic programming system for revising knowledge bases. In: *Knowledge Representation and Reasoning*. [S.l.]: Morgan Kaufmann, 1994.

[Damásio, Pereira e Schroeder 1996] DAMáSIO, C. V.; PEREIRA, L. M.; SCHROEDER, M. Revise progress report. In: *Proc. of the Workshop on Automated Reasoning: Bridging the Gap between Theory and Practice*. Brighton: [s.n.], 1996.

[Donnellan 1966] DONNELLAN, K. S. Reference and definite descriptions. In: *The Philosophical Review*. [S.l.: s.n.], 1966. p. 281–304. Reprinted in *R*eadings in the Philosophy of Language by Peter Ludlow (ed), MIT Press.

[Dowty, Wall e Peters 1981] DOWTY, D. R.; WALL, R. E.; PETERS, S. *Introduction to Montague Semantics*. [S.l.]: D. Reidel Publishing, 1981.

[Eshghi e Kowalski 1989] ESHGHI, K.; KOWALSKI, R. Abduction compared with negation by failure. In: *Proceedings of 6th International Conference on Logic Programming*. [S.l.: s.n.], 1989. p. 245–255.

[Filho e Freitas 2003] FILHO, A. M. C.; FREITAS, S. A. A. d. Interpretação do futuro do pretérito em narrativas. In: *Anais do 1º workshop em Tecnologia da Informação e da Linguagem Humana, TIL'2003*. São Carlos - SP, Brasil: [s.n.], 2003. Disponível em: <nilc.icmc.sc.usp.br/til2003>.

[Fraurud 1990] FRAURUD, K. Definiteness and the processing of np's in natural discourse. *Journal of Semantics*, v. 7, n. 4, p. 395–433, 1990.

[Freitas 1993] FREITAS, S. A. A. d. *Deíticos e Anáforas Pronominais*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, Porto Alegre - RS, Brasil, 1993.

[Freitas e Lopes 1994] FREITAS, S. A. A. de; LOPES, J. G. P. Discourse segmentation: Extending the centering theory. In: *XI Simpósio Brasileiro de Inteligência Artificial*. UFCE - Fortaleza - CE: [s.n.], 1994.

[Freitas e Lopes 1996] FREITAS, S. A. A. de; LOPES, J. G. P. Solving the reference to mixable entities. In: *Proceedings of the Indirect Anaphora Workshop*. University of Lancaster, Lancaster, UK: [s.n.], 1996.

[Freitas e Lopes 1998] FREITAS, S. A. A. de; LOPES, J. G. P. *An Abductive Approach to the Definite Noun Anaphora Problem*. DI/FCT, Lisboa, Portugal, 1998.

[Freitas, Lopes e Menezes 2004] FREITAS, S. A. A. de; LOPES, J. G. P.; MENEZES, C. da S. Abducing definte descriptions relations. In: *Anais do XXIV Congresso da Sociedade Brasileira de Computação*. Salvador - BA, Brasil: [s.n.], 2004.

[Gardent 2002] GARDENT, C. Generating minimal definite descriptions. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2002. p. 96–103.

[Garrod, Freudenthal e Boyle 1994] GARROD, S. C.; FREUDENTHAL, D.; BOYLE, E. The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language*, n. 33, p. 39–68, 1994.

[Groenendijk e Stokhof 1991] GROENENDIJK, J.; STOKHOF, M. Dynamic predicate logic. *Linguistics and Philosophy*, v. 14, p. 39–100, 1991.

[Grosz e Sidner 1986] GROSZ, B.; SIDNER, C. L. Attention, intentions and the structure of the discourse. *Computational Linguistics*, v. 12, n. 3, p. 175–204, 1986.

[Grosz e Sidner 1990] GROSZ, B.; SIDNER, C. L. Plans for discourse. In: COHEN, P.; MORGAN; POLLACK, M. (Ed.). *Intentions in Communication*. [S.l.: s.n.], 1990. p. 417–443.

[Grosz e Sidner 1998] GROSZ, B.; SIDNER, C. L. Lost intuitions and forgotten intentions. In: WALKER, M. A.; JOSHI, A. K.; PRINCE, E. F. (Ed.). *Centering in Discourse*. [S.l.]: Oxford University Press, 1998. p. 39–51.

[Grosz 1977] GROSZ, B. J. *The Representation and Use of Focus in a System for Understanding Dialogs*. SRI International, Menlo Park, California, 1977.

[Grosz, Joshi e Weinstein 1983] GROSZ, B. J.; JOSHI, A. K.; WEINSTEIN, S. Providing a unified account of definite noun phrases in discourse. In: *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*. MIT, USA: [s.n.], 1983. p. 44–50.

[Grosz, Joshi e Weinstein 1995] GROSZ, B. J.; JOSHI, A. K.; WEINSTEIN, S. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, v. 21, n. 2, p. 203–225, 1995.

[Gruber 1976] GRUBER, J. *Lexical Structure in Syntax and Semantics*. [S.l.]: North Holland Publishers, 1976.

[Gundel 1994] GUNDEL, J. K. On different kinds of focus. In: BOSCH, P.; SANDT, R. van der (Ed.). *Focus in Natural Language Processing*. Heidelberg, Germany: IBM, 1994, (Working Papers of the Institute for Logic and Linguistics, v. 3). p. 457–466.

[Gundel, Hegarty e Borthen 2003] GUNDEL, J. K.; HEGARTY, M.; BORTHEN, K. Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information*, v. 12, p. 281–299, 2003.

[Hahn, Markert e Strube 1996] HAHN, U.; MARKERT, K.; STRUBE, M. A conceptual reasoning approach to the resolution of textual ellipses. In: *Proceedings of 12th European Conference on Artificial Intelligence, ECAI'96*. Budapest: [s.n.], 1996. p. 572–576.

[Hahn e Strube 1996] HAHN, U.; STRUBE, M. Incremental centering and center ambiguity. In: *Proceedings of 18th Annual Conference of the Cognitive Science Society, CogSci'96*. San Diego, USA: [s.n.], 1996. p. 568–573.

[Hahn e Strube 1996] HAHN, U.; STRUBE, M. Parsetalk about functional anaphora. In: *Proceedings of International Conference on Artificial Intelligence, AI'96*. Toronto, Canada: [s.n.], 1996. p. 133–145.

[Hahn, Strube e Markert 1996] HAHN, U.; STRUBE, M.; MARKERT, K. Bridging textual ellipsis. In: *Proceedings of the 16th International Conference on Computational Linguistics*. [S.l.: s.n.], 1996. p. 496–501.

[Hajičová, Skoumalová e Sgall 1995] HAJIČOVá, E.; SKOUMALOVá, H.; SGALL, P. An automatic procedure for topic-focus identification. *Computational Linguistics*, v. 21, n. 1, p. 81–94, 1995.

[Hawkins 1978] HAWKINS, J. A. *Definiteness and Indefiniteness: a study in reference and Grammaticality prediction*. [S.l.]: Croom Helm, 1978.

[Heim 1982] HEIM, I. *The Semantics of Definite and Indefinite Noun Phrases*. Tese (Doutorado) — University of Massachussets, 1982.

[Hirst 1981] HIRST, G. *Anaphora in Natural Language Understanding: A Survey*. [S.l.]: Spring-Verlag, 1981. (Lecture Notes in Computer Science, v. 119).

[Hobbs 1979] HOBBS, J. R. Coherence and coreference. *Cognitive Science*, v. 3, n. 1, p. 67–89, 1979.

[Hobbs 1985] HOBBS, J. R. *On the Coherence and Structure of Discourse*. CLSI - Stanford University, 1985. Relatório Técnico nº CSLI-85-37.

[Hobbs 1993] HOBBS, J. R. Intention, information, and structure in discourse: A first draft. In: *Proceedings of the NATO Advanced Research Workshop: Burning Issues in Discourse*. Maratea, Italy: [s.n.], 1993. p. 41–66.

[Hobbs et al. 1993] HOBBS, J. R. et al. Interpretation as abduction. *Artificial Intelligence*, v. 63, p. 69–142, 1993.

[Hovy 1990] HOVY, E. H. Parsimonious and profligate approaches to the question of discourse structure relations. In: *Proceedings of the 5th International Workshop on Natural Language Generation*. Pittsburgh: [s.n.], 1990. p. 170–176.

[Huang 1994] HUANG, Y. *The Syntax and Pragmatics of Anaphora*. [S.l.]: Cambridge University Press, 1994. (Cambridge Studies in Linguistics).

[Huang 2000] HUANG, Y. Discourse anaphora: Four theoretical models. *Journal of Pragmatics*, v. 32, p. 151–176, 2000.

[Jr. e Duffy 2001] JR., C. C.; DUFFY, S. A. Sentence and text comprehension: Roles of linguistic structure. *Annual Reviews in Psycholinguistic*, v. 52, p. 167–196, 2001.

[Kakas, Kowalski e Toni 1992] KAKAS, A.; KOWALSKI, R.; TONI, F. Abdutive logic programming. *Journal of Logic Computational*, v. 2, n. 6, p. 719–770, 1992.

[Kameyama 1997] KAMEYAMA, M. Intrasentential centering: A case study. Available as cmp-lg/9707005 at Computational Linguistics E-print. 1997.

[Kameyama, Passanneau e Poesio 1993] KAMEYAMA, M.; PASSANNEAU, R. J.; POESIO, M. Temporal centering. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Ohio State University – Columbus, Ohio, USA: [s.n.], 1993. p. 70–77.

[Kamp e Reyle 1993] KAMP, H.; REYLE, U. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1993.

[Kehler 1993] KEHLER, A. A discourse copying algorithm for ellipsis and anaphora resolution. In: *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*. OTS - Utrecht University - Utrecht, The Netherlands: [s.n.], 1993. p. 203–212.

[Kehler 1993] KEHLER, A. The effect of establishing coherence in ellipsis and anaphora resolution. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Ohio State University – Columbus, Ohio, USA: [s.n.], 1993. p. 62–69.

[Kehler 1993] KEHLER, A. Intrasentential constraints on intersentential anaphora in centering theory. In: *Proceedings of the Workshop on Centering Theory in Natural Language Occurring Discourse*. University of Pennsylvania: [s.n.], 1993.

[Kehler 2000] KEHLER, A. Coherence and the resolution of ellipsis. *Linguistics and Philosophy*, v. 23, p. 533–575, 2000.

[Komagata 2003] KOMAGATA, N. Information structure in subordinate and subordinate-like clauses. *Journal of Logic, Language and Information*, v. 12, p. 301–318, 2003.

[Kruijff-Korbayová e Steedman 2003] KRUIJFF-KORBAYOVá, I.; STEEDMAN, M. Discouse and information structure. *Journal of Logic, Language and Information*, n. 12, p. 249–259, 2003.

[Langacker 1966] LANGACKER, R. W. On pronominalization and the chain of command. In: REIBEL, D.; SHANE, S. (Ed.). *Modern Studies in English*. [S.l.]: Prentice Hall, 1966. p. 160–186.

[Levine, Guzmán e Klin 2000] LEVINE, W. H.; GUZMáN, A. E.; KLIN, C. M. When anaphor resolution fails. *Journal of Memory and Language*, n. 43, p. 594–617, 2000.

[Lopes e Freitas 1994] LOPES, J. G. P.; FREITAS, S. A. A. de. Improving centering to support discourse segmentation. In: BOSCH, P.; SANDT, R. van der (Ed.). *Focus in Natural Language Processing*. Heidelberg, Germany: IBM, 1994, (Working Papers of the Institute for Logic and Linguistics, v. 3). p. 533–542.

[Mann e Thompson 1987] MANN, W. C.; THOMPSON, S. A. *Rhetorical Structure Theory: A Theory of Text Organization*. ISI Reprint Series, 1987. Relatório Técnico nº ISI/RS-87-190.

[Markert, Strube e Hahn 1996] MARKERT, K.; STRUBE, M.; HAHN, U. Inferential realization constraints on functional anaphora in the centering model. In: *Proc. of 18th Annual Conference of the Cognitive Science Society, CogSci'96*. San Diego, USA: [s.n.], 1996. p. 609–614.

[Moore e Pollack 1992] MOORE, J. D.; POLLACK, M. E. A problem for rst: The need for multi-level discourse analysis. *Computational Linguistics*, v. 18, n. 4, p. 537–544, 1992.

[Passonneau e Litman 1997] PASSONNEAU, R. J.; LITMAN, D. J. Discourse segmentation by human and automated means. *Computational Linguistics*, v. 23, n. 1, p. 103–139, 1997. Disponível em: <http://www.aclweb.org/anthology/J97-1005.pdf>.

[Paulo 2002] PAULO, N. de S. *Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo*. 2002. Via internet. Disponível em: <http://acdc.linguateca.pt/cetenfolha/>.

[Pereira e Alferes 1992] PEREIRA, L. M.; ALFERES, J. Well founded semantics for logic programs with explicit negation. In: B.NEWMANN (Ed.). *Proceedings of the 10th European Conference on Artificial Inteligence*. [S.l.: s.n.], 1992. p. 102–106.

[Pereira, Damásio e Alferes 1993] PEREIRA, L. M.; DAMáSIO, C. V.; ALFERES, J. J. Diagnosis and debugging as contradiction removal. In: *Proceedings of the second international workshop on Logic Programming and Non-Monotonic Reasonig*. [S.l.]: MIT Press, 1993. p. 316–330.

[Polanyi 1988] POLANYI, L. A formal model of the structure of discourse. *Journal of Pragmatics*, n. 12, p. 601–638, 1988.

[Polanyi e Berg 1996] POLANYI, L.; BERG, M. van den. Discourse structure and discourse interpretation. In: DEKKER, P.; STOKHOF, M. (Ed.). *Tenth Amsterdam Colloquium*. Department of Philosophy, University of Amsterdam: [s.n.], 1996. p. 113–131. Disponível em: <citeseer.nj.nec.com/polanyi96discourse.html>.

[Polanyi, Berg e Ahn 2003] POLANYI, L.; BERG, M. van den; AHN, D. Discourse structure and sentential information structure. *Journal of Logic, Language and Information*, v. 12, p. 337–350, 2003.

[Prince 1981] PRINCE, E. F. Toward a taxonomy of given-new information. In: COLE, P. (Ed.). *Radical Pragmatics*. New York: Academic Press, 1981. p. 223–256.

[Prince 1992] PRINCE, E. F. Description. In: THOMPSON, S.; MANN, W. (Ed.). *Discourse description: diverse analyses of a fund raising text*. Philadelphia/Amsterdam: John Benjamins Publishing Company, 1992. p. 295–325.

[Reinhart 1976] REINHART, T. *The Syntactic Domain of Anaphora*. Tese (Doutorado) — MIT, Cambridge, MA, USA, 1976.

[Reinhart 1981] REINHART, T. Definite np anaphora an c-command domains. *Linguistic Inquiry*, v. 12, n. 4, p. 605–635, 1981.

[Reinhart 1983] REINHART, T. *Anaphora and Semantic Representation*. [S.l.]: Croom Helm, London, 1983.

[Rodrigues 1995] RODRIGUES, I. P. *Processamento de texto: Interpretação temporal*. Tese (Doutorado) — Universidade Nova de Lisboa, 1995.

[Rodrigues e Lopes 1992] RODRIGUES, I. P.; LOPES, J. G. P. A system for text temporal information retrieval. In: BOULAY, B. du; SGUREV, V. (Ed.). *Artificial Intelligence V: methodology, systems, applications*. Cambridge, MA: North-Holland, 1992. p. 181–190.

[Rodrigues e Lopes 1992] RODRIGUES, I. P.; LOPES, J. G. P. Temporal structure of discourse. In: *Proceedings of the 13th International Conference on Computational Linguistics*. Nantes, France: [s.n.], 1992. p. 331–337.

[Rodrigues e Lopes 1993] RODRIGUES, I. P.; LOPES, J. G. P. Building the text temporal structure. In: FILGUEIRAS, M.; DAMAS, L. (Ed.). *Progress in Artificial Intelligence - 6th Portuguese Conference on AI, EPIA'93*. [S.l.: s.n.], 1993, (Lecture Notes in Artificial Intelligence, number 727). p. 45–60.

[Rodrigues e Lopes 1994] RODRIGUES, I. P.; LOPES, J. G. P. Temporal information retrieval from text. In: MARTIN-VIDE, C. (Ed.). *Current Issues in Mathematical Linguistics*. [S.l.]: North-Holland, 1994. p. 279–288.

[Rodrigues e Lopes 1995] RODRIGUES, I. P.; LOPES, J. G. P. Representing events and states for information retrieval from texts. In: *Proceedings of the 5th International Workshop on Natural Language Understanding and Logic Programming*. Lisbon - Portugal: Springer-Verlag, 1995. p. 39–54.

[Russell 1919] RUSSELL, B. Description. In: *Introduction to Mathematical Philosophy*. [S.l.: s.n.], 1919. Reprinted in *R*eadings in the Philosophy of Language by Peter Ludlow (ed), MIT Press.

[Sandt 1992] SANDT, R. van der. Presupposition projection as anaphor resolution. *Journal of Semantics*, v. 19, p. 333–377, 1992.

[Sanford e Garrod 1981] SANFORD, A. J.; GARROD, S. C. *Understanding written language: Explorations of comprehension beyond the sentence*. Chichester, England: John Wiley and Sons, 1981.

[Scliar-Cabral 2002] SCLIAR-CABRAL, L. Referência: Qual a referência e como evocá-la? *DELTA*, n. 18, p. 57–85, 2002.

[Seville e Ramsay 1999] SEVILLE, H.; RAMSAY, A. Reference-based discourse structure for reference resolution. In: *Proceedings of the Workshop on The Relation of Discourse/Dialogue Structure and Reference on the 37th ACL*. University of Maryland, College Park, Maryland, USA: [s.n.], 1999. p. 90–99.

[Sibun 1993] SIBUN, P. Domain structure, rhetorical structure, and text structure. In: *Proceedings of the Workshop on Intentionality and Structure in Discourse Relations on the 31st ACL*. Ohio State University – Columbus, Ohio, USA: [s.n.], 1993. p. 118–121.

[Sidner 1979] SIDNER, C. L. *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. Tese (Doutorado) — MIT, Cambridge, MA, USA, 1979.

[Sidner 1981] SIDNER, C. L. Focusing for interpretation of pronouns. *American Journal for Computational Linguistics*, v. 7, n. 4, p. 217–231, 1981.

[Spenader 2003] SPENADER, J. Factive pressuppositions, accommodation and information structure. *Journal of Logic, Language and Information*, v. 12, p. 351–368, 2003.

[Strand 1996] STRAND, K. A taxonomy of linking relations. In: *Proceedings of the Indi-Ana Workshop*. University of Lancaster, UK: [s.n.], 1996.

[Strawson 1950] STRAWSON, P. F. On referring. v. 59, p. 320–344, 1950. Reprinted in *R*eadings in the Philosophy of Language by Peter Ludlow (ed), MIT Press.

[Strohner et al. 2000] STROHNER, H. et al. Discourse focus and conceptual relations in resolving referential ambiguity. *Journal of Psycholinguistic Research*, v. 29, n. 5, p. 497–516, 2000.

[Strube e Hahn 1996] STRUBE, M.; HAHN, U. Functional centering. In: *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, USA: [s.n.], 1996. p. 270–277.

[Strzalkowski 1999] STRZALKOWSKI, T. (Ed.). *Natural Language Information Retrieval*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1999.

[Vieira e Poesio 2000] VIEIRA, R.; POESIO, M. An empirically-based system for processing definite descriptions. *Computational Linguistics*, v. 26, n. 4, p. 539–593, 2000.

[Walker, Lida e Cote 1994] WALKER, M.; LIDA, M.; COTE, S. Japanese discourse and the process of centering. *Computational Linguistics*, v. 20, n. 2, p. 193–232, 1994.

[Walker 1989] WALKER, M. A. Evaluating discourse processing algorithms. In: *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 1989. p. 251–261.

[Walker 1996] WALKER, M. A. Limited attention and discourse structure. *Computational Linguistics*, v. 22, n. 2, p. 255–264, 1996.

APÊNDICE A – Pesquisa de informações em documento

Este anexo descreve as possíveis utilizações da metodologia desenvolvida nesta tese, em especial da END, na pesquisa de informações em documentos digitais. Olhando para a estrutura nominal, nota-se que a árvore resultante expressa a forma com que os assuntos se movimentaram no decorrer de um discurso. Até o momento, essa movimentação, que dá origem à END, foi utilizada *exclusivamente* para resolver anáforas. Porém, a END resultante da interpretação de um discurso ou documento é uma hierarquização das suas entidades de forma a que as mais salientes estejam mais visíveis na árvore final.

Assim, considerando um conjunto de documentos digitais $\mathcal{DS} = d_1, d_2, ...d_i, ..., d_n$, as metodologias para recuperação de informação [Baeza-Yates e Ribeiro-Neto 1999, Strzalkowski 1999] podem utilizar o processo desenvolvido nesta tese de duas formas:

1. cada documento d_i passa pelo processo de interpretação de anáforas, o qual produz um documento virtual dv_i. No documento virtual todas as anáforas resolvidas estão substituídas pela descrição completa dos seus antecedentes. Logo a seguir, dv_i é entrada para uma máquina de indexação. Após o tratamento de todos os documentos de $\mathcal{DS}$, o índice gerado pode ser usado para pesquisa.

2. cada documento d_i passa pelo processo de interpretação de anáforas e produz o documento virtual dv_i e a END e_i. A seguir d_i, dv_i e e_i são entrada para uma máquina de indexação, a qual vai gerar três índices distintos, um para cada entrada. Após o tratamento de todos os documentos de $\mathcal{DS}$, os índices gerados podem ser utilizados por uma metamáquina de busca (MMB).

A seguir são descritas estas duas formas de utilização da END.

A.1 Documento Virtual

Neste tipo de utilização, a metodologia descrita nesta tese é aplicada sobre um documento puro d, pertencente à coleção $\mathcal{DS}$, gerando um documento virtual dv onde os SNDs e pronomes foram substituídos pelos termos léxicos de seus antecedentes. Por exemplo: "O Lucas foi à padaria. Ele estava com fome" em d vai ser substituído por "O Lucas foi à padaria. O Lucas estava com fome" em dv.

Uma máquina de indexação de palavras/termos usa dv como entrada. O processo é repetido para todos os documentos de $\mathcal{DS}$ gerando um índice único I_{DS} que é utilizado na pesquisa de documentos (figura 21).

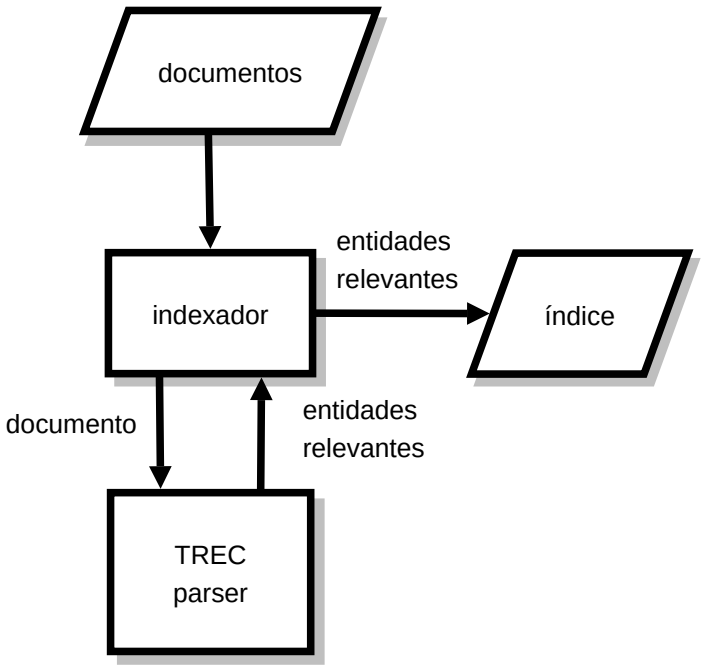


Figura 21: Arquitetura geral do indexador de documentos.

Dada uma sentença de busca q constituída de uma série de termos $t_1, t_2, ..., t_i, ..., t_n$ e operações lógicas AND e OR[1] entre estes termos, o processo de localização da coleção de documentos CD_q que satisfaz q é definido a seguir:

1. Repetir para $0 < i \leq n$:

   (a) Pesquisar o termo t_i em I_{CD}. O resultado é a lista CD_{t_i}, ou seja, o subconjunto dos documentos de $\mathcal{DS}$ onde figuram o termo t_i.

[1]Por simplificação, não estão sendo consideradas outras operações tais como XOR e NOT.

2. Entre todos termos t_i de q existe uma operação possível. Se nada for dito em contrário, todas as operações são *ANDs lógicos*. A expressão q infixa deve ser pós-fixada, gerando a pilha qp.

3. Obter CD_q a partir do cálculo da expressão dada por qp, considerando que:

   (a) a operação AND entre duas coleções é a intersecção dos seus documentos,

   (b) a operação OR entre duas coleções é a união de seus documentos.

A seguir é feita uma classificação dos documentos de CD_q para melhorar a exposição dos resultados ao usuário. Esta ordenação considera critérios tais como [Baeza-Yates e Ribeiro-Neto 1999]: freqüência de ocorrência dos termos, proximidade dos termos do início do documento, localização de campos especiais etc.

A utilização do processo descrito nesta tese permite a melhoria de conteúdo do documento virtual. Com a resolução das anáforas, os termos contidos em cada frase do documento gerado expõem entidades e ligações que estavam obscuras no documento original. O resultado é que a indexação vai agora captar tais termos, possibilitando uma busca mais refinada dos documentos. Porém, o grande impacto dá-se no processo de classificação de CD_q, onde haverá uma modificação da freqüência dos termos e na posição dos mesmos nos documentos. Assim, a classificação dos documentos é melhor do que quando usado o documento puro.

A utilização deste tipo de metodologia permite uma fácil integração com os mecanismos de indexação/busca já existentes. Porém, ficam algumas perguntas em aberto: no caso das anáforas nominais definidas o que deve ser colocado no documento virtual, visto que neste caso a expressão anafórica e o antecedente não são os mesmos? Como colocar as relações identificadas no documento virtual?

Diferente das co-referências onde somente há substituições, neste caso seriam necessárias inserções léxicas no texto. Tais inserções devem ser feitas através da geração de frases, sem a qual pode haver modificação do conteúdo semântico do documento. Isto é um trabalho dobrado: gerar texto é tão complexo quanto interpretá-lo.

Por fim, distoando do contexto dos documentos virtuais, a Estrutura Nominal do Discurso, gerada pela interpretação das anáforas e representando o acompanhamento das entidades mais salientes do discurso, é um resumo do documento. Como tal ela pode ser usada como fonte de busca, constituindo assim um índice à parte.

A.2 A Metamáquina de Busca

A metamáquina de busca é um modelo no qual podem existir quantos índices forem necessários (figura 22). Cada documento da coleção vai alimentar analisadores individuais tais como: interpretação de anáforas, extração da END e de outras estruturas, lematização, árvore sintática etc. A saída de cada um destes analisadores pode ser tanto um documento virtual quanto um outro tipo de representação, por exemplo, uma DRS, uma árvore de derivação sintática ou uma estrutura qualquer. Cada um destes resultados é indexado por um determinado tipo de índice (e.g. árvores balanceadas, *suffix arrays* etc).

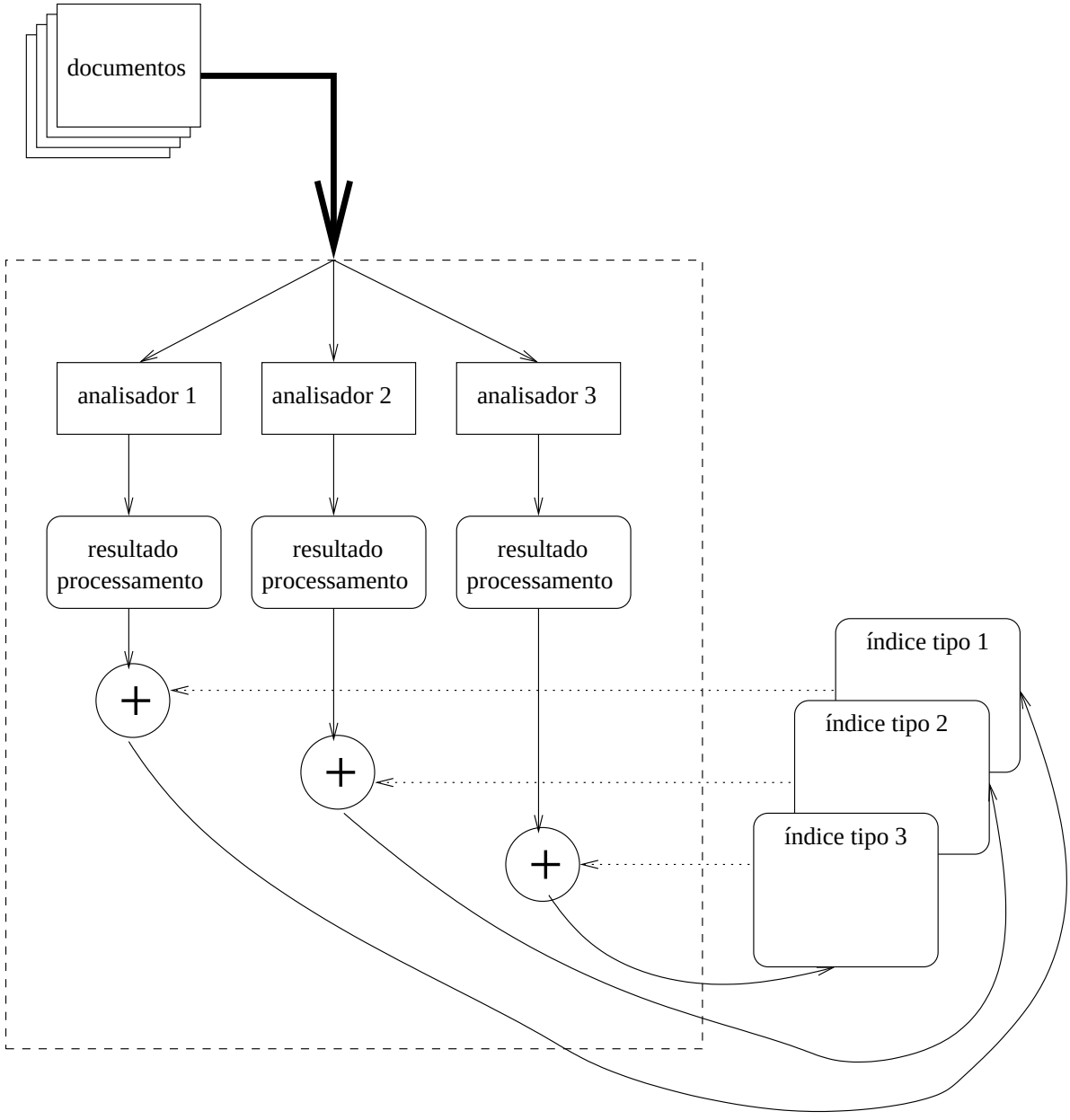


Figura 22: Indexação numa metamáquina de busca.

No processo de busca, a MMB vai pesquisar uma dada sentença q em todos os índices (figura 23). Quando possível a sentença q é adaptada para o tipo de informação armazenada num determinado índice, por exemplo: se um determinado índice representa documentos lematizados então deve-se lematizar q, se o índice armazena árvores de derivação sintática então deve-se atribuir categorias sintáticas aos termos de q. O resultado da busca são n coleções[2] de documentos que satisfazem q.

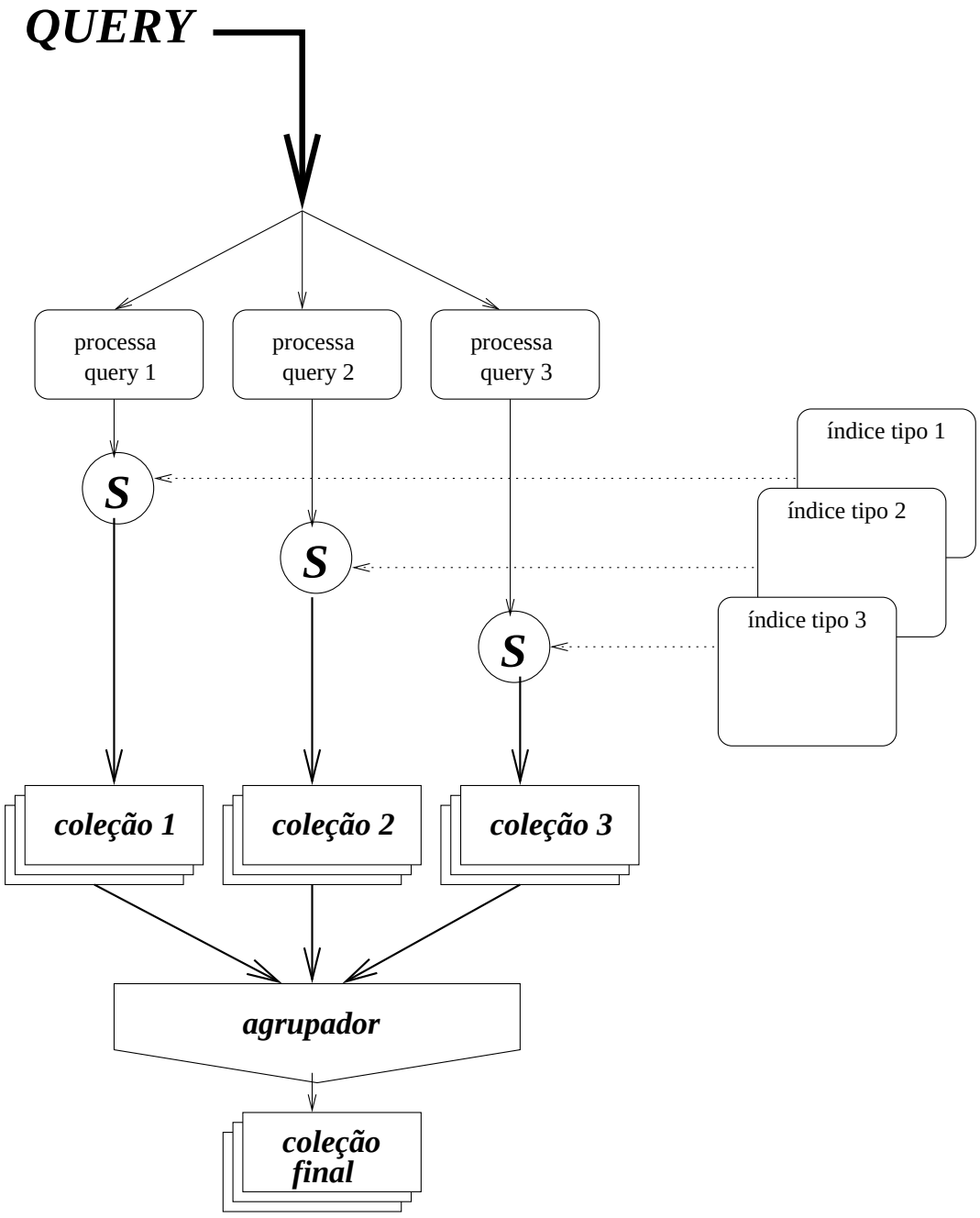


Figura 23: Busca numa MMB.

[2] Tantas quantas forem o número de índices utilizados.

As coleções individuais (*e.g. coleção 1, coleção 2 e coleção 3)* devem ser agrupadas num único conjunto de documentos, o qual será apresentado ao usuário. O processo de agrupamento é dividido em duas fases:

1. agrupamento das coleções individuais - são identificados quais documentos estão presentes em todas as coleções. Em princípio, a intersecção dos documentos das diversas coleções será o conjunto final. Porém é possível estabelecer pesos, de acordo com o perfil do usuário[3], para a existência de documentos em determinados índices. Pode então ocorrer que um documento não faça parte da intersecção das coleções, mas mesmo assim esteja presente na coleção final.

2. classificação da coleção final - é ordenado o conjunto resultante da fase anterior de forma a apresentar primeiramente os documentos mais *expressivos* para uma dada sentença q. O processo de ordenação considera os critérios utilizados pelos *documentos virtuais* [Baeza-Yates e Ribeiro-Neto 1999], os quais são aplicados a cada documento e balanceados pelo tipo de informação representada em cada índice. É possível modificar os pesos do balanceamento de acordo com o perfil do usuário.

A END é implementada na MMB através da criação de um índice próprio (END-MMB), o qual vai armazenar a estrutura em árvore da END, com cada nó contendo apenas os focos implícitos e explícitos. Originalmente, uma END tem o formato apresentado na figura 24a. Note que o último segmento visível (SV), que seria a representação da última frase interpretada, foi considerado como sendo um segmento (S). É feita uma transformação nesta árvore: todos os segmentos visíveis passam a ser a raiz da árvore de seus subsegmentos e cada SV, por sua vez, é elemento de uma lista (figura 24b).

Note que a figura 24b representa uma seqüência de subárvores ordenadas pelos SVs. Cada SV representa um assunto que ficou visível após a interpretação do documento. Mais ainda, a lista de SVs constitui a forma pela qual os assuntos foram sendo conduzidos pelo transmissor. Assim, um assunto no início da lista tem um peso maior do que um assunto no final da lista.

Olhando para cada SV individualmente, vê-se que este tem uma subárvore agregada representando a forma pela qual o assunto do SV (i.e. focos) foi desenvolvido no decorrer de um documento. Este desenvolvimento é também estruturado na forma de árvore e considera que os nós pais são mais relevantes que nós filhos. Portanto nós mais próximos da raiz assinalam assuntos mais relevantes para um dado SV.

[3] Aqui não é tratada a questão da identificação dos perfis de usuário.

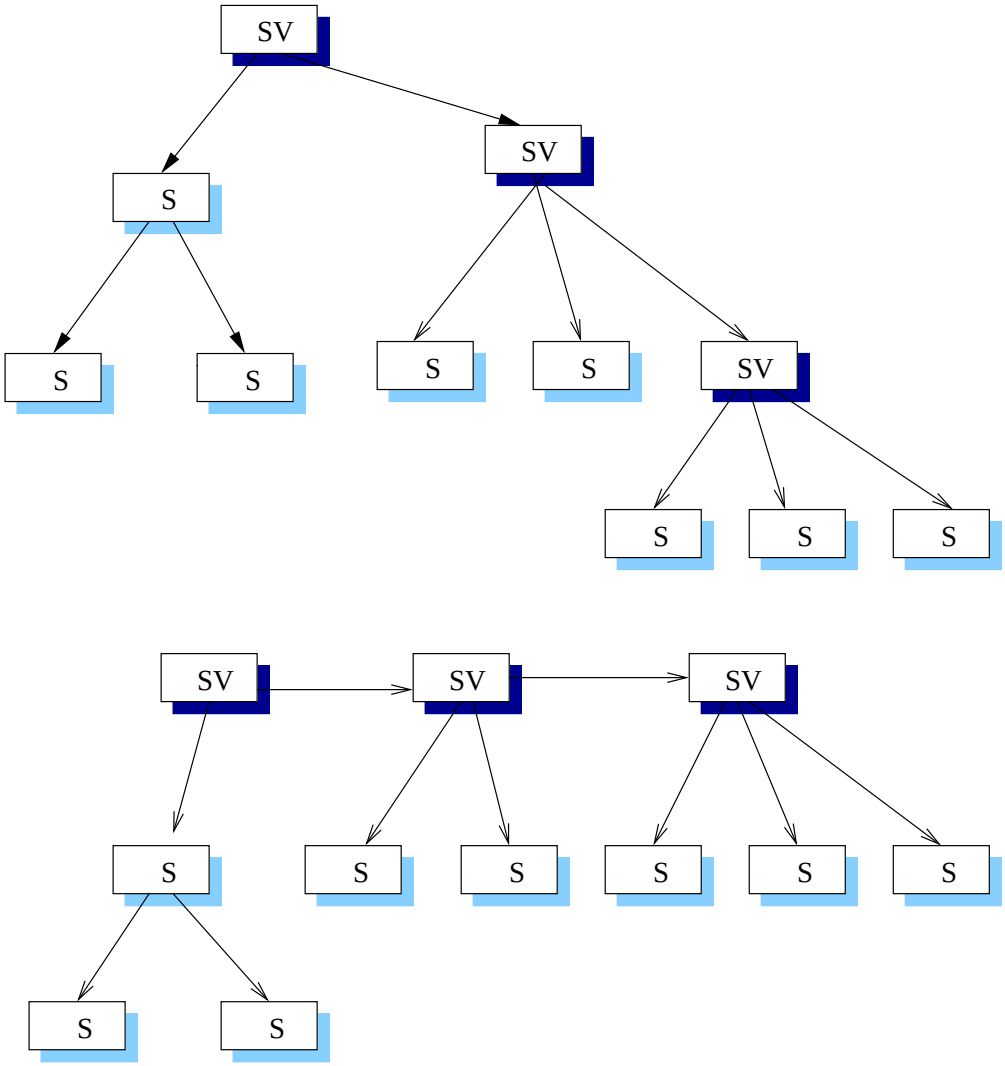


Figura 24: Transformando uma END num índice END-MMB.

Esta nova visão sobre a END pode ser usada como índice de busca e classificação de documentos. Para tal considera-se que somente os focos vão ser indexados, mantendo para cada documento a sua estrutura de seqüência de árvores (figura 24b).

Dada um sentença q qualquer, a busca é feita localizando-se cada termo t_i de q no índice END. Vale ressaltar que nem todos os termos t_i serão localizados, pois o índice END representa apenas os assuntos descartando o restante das entidades e eventualidades. Porém quando um termo t_i for encontrado num SV de um documento é sinal este trata do assunto e deve ser selecionado.

Localizado um conjunto de documentos $\mathcal{CD}_q = d_1, d_2, ..., d_i, ..., d_m$ para uma dada q deve-se então fazer a sua classificação. O mecanismo de classificação calcula para cada documento d_i um determinado valor de relevância VR_{d_i}. Valores maiores para VR_{d_i} significam documentos mais relevantes. O cálculo de VR_{d_i} leva em consideração as posições onde os termos q foram localizados na END-MMB (figura 25).

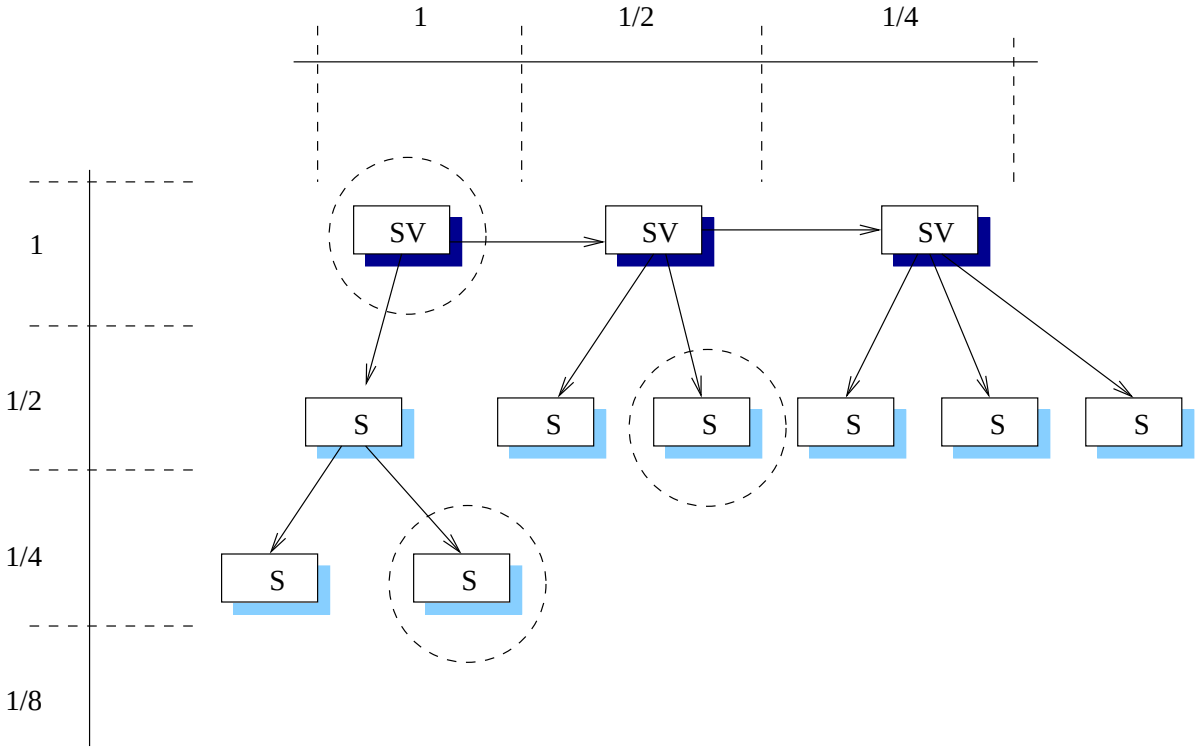


Figura 25: Pesos para o cálculo do valor de relevância.

A figura 25 mostra a relação entre a profundidade em que um termo foi encontrado e a influência que o mesmo deve ter no cálculo de VR_{d_i} para um dado documento d_i. Quanto mais próximo da raiz tiver sido localizado um termo maior deve ser VR_{d_i} (i.e. menor profundidade). Quanto mais próximo da primeira subárvore (i.e. mais à esquerda)

estiver a subárvore que contenha um termo localizado maior deve ser VR_{d_i}.

Levando em conta estes critérios são atribuídos:

- para cada nível de profundidade, a fração $\frac{1}{2^p}$ onde $p = 0, 1, 2, ...(mp - 1)$, mp é a máxima profundidade da árvore (0 para a raiz e crescendo com a profundidade) e
- para cada nível de deslocamento horizontal, a fração $\frac{1}{2^v}$ onde $v = 0, 1, 2, ...(mv - 1)$, mv é o número de segmentos visíveis existentes na END-MMB de d_i (considerando 0 para a subárvore mais à esquerda e incrementando o valor para a direita).

Dado o conjunto de termos $TL_{d_i} = [t_1, t_2, ..., t_j, ..., t_k] \,|\, t_j \in q$ localizados na END-MMB do documento d_i e as coordenadas v_j, p_j, respectivamente deslocamento e profundidade do termo t_j na árvore (figura 25), é calculado o valor de relevância VRq_{d_i} do documento d_i dado q:

$$VRq_{d_i} = \prod_{v=0}^{(mv-1)} (\frac{1}{2^j} VRq_{d_i}^{SV_v}) \tag{A.1}$$

onde $VRq_{d_i}^{SV_v}$ mede o grau de relevância de cada SV_v em d_i. Os valores de $VRq_{d_i}^{SV_v}$ não podem ser nulos, fato que acontece quando nenhum dos termos de q são localizados na árvore cujo pai é SV_v. De acordo com a equação A.1 quanto mais um SV_v estiver à esquerda, maior a sua relevância para o documento (v menor). Por sua vez, $VRq_{d_i}^{SV_v}$ é calculado de acordo com a profundidade que os termos t_j são encontrados na sua subárvore. Quanto mais termos forem encontrados melhor para a afirmação do assunto em SV_v, quanto mais profundo estiver um termo menor a sua importância para SV_v. Isto é traduzido na equação A.2.

$$VRq_{d_i}^{SV_v} = \sum_{p=0}^{(mp-1)} [\frac{1}{2^p} f(t_j, v, p)] \tag{A.2}$$

onde a função $f(t_j, v, p)$ retorna 1 caso o termo t_j seja encontrado nas coordenadas (v, p) e retorna 0 (zero) caso o termo não seja encontrado nestas coordenadas.

A.3 Considerações finais

Fica claro que a metodologia desenvolvida neste trabalho pode ser empregada na busca de informação. Isto pode ser feito tanto no tratamento de documentos virtuais

através da interpretação das anáforas, quanto na geração de um índice, via END, para uma metamáquina de busca.

Ambas as formas de busca apresentam vantagens e desvantagens: enquanto o índice único é adequado à busca irrestrita para grandes coleções de texto pois tem um peso computacional baixo, a metamáquina pode fornecer buscas mais refinadas a um custo computacional maior, porém perfeitamente justificáveis na busca de documentos corporativos (intranet).

Por fim, apesar de não explorado até aqui, a técnica de utilização da END na MMB pode ser estendida para a recuperação de informação (geração de resumos) e para a classificação de coleções de documentos, através da medição de similaridade entre árvores END.

APÊNDICE B – Utilização do REVISE

Este anexo apresenta a utilização do REVISE [Damásio, Pereira e Schroeder 1996] na implementação do protótipo. O sistema foi originalmente implementado em Sicstus Prolog. Foi feita uma adaptação para o SWI Prolog (licença LGPL). Porém a operação é a mesma do original:

1. Carrega-se o interpretador Prolog
   $ pl

2. Carrega-se o módulo principal do revise
   ?- [boot].

3. A seguir o conjunto de regras *revise* são carregadas
   ?- read_file2(´pragmatic.rules´).

4. Os resultados da interpretação é obtida usando-se o predicado *findall*
   ?- findall(B,solution(B),S).

5. Cada elemento da lista de soluções S é um par de listas contendo, respectivamente, os *revisíveis* que devem ser tornados verdadeiros e os que devem ser tornados falsos para que a base de dados fique consistente.

As regras do arquivo *pragmatic.rules* são:

- O conjunto de revisíveis

:- revisable(member_of1(A,T)).
:- revisable(part_of1(A,T)).
:- revisable(subcat_of1(A,T)).
:- revisable(coref1(A,T)).
:- revisable(accomod1(A)).
:- revisable(visivel(S)).

- e o conjunto de regras no formato REVISE

  mesmo(E,E).

  snd(A) <- antecedente_snd(T),
  not mesmo(A, T),
  member_of1(A,T).

  Note que *mesmo*(E, E). é um predicado Prolog *puro e snd*(A) *é declarado como uma regra REVISE através do símbolo ´<-´.*

As frases de um texto são um conjunto de fatos:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% texto: o autocarro chegou à estação.
%       o condutor abriu as portas.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
onibus(i1). sing(i1). masc(i1). snd(i1).
rodoviaria(i2). sing(i2). fem(i2). snd(i2).
motorista(i3). sing(i3). masc(i3). snd(i3). animado(i3).
portas(i4). fem(i4). plu(i4). snd(i4).

visivel(s1).
foco_exp(s1,i1). lre_exp(s1,[i1,i2]).
foco_imp(s1,null). lre_imp(s1,[]).
% para explicar porque snd(i3) e snd(i4)
<- not snd(i3).
<- not snd(i4).
```