Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Análise de Discurso: Teorias Discursivas e Aplicações em Processamento de Línguas Naturais

Thiago Alexandre Salgueiro Pardo
Maria das Graças Volpe Nunes

NILC-TR-03-06

Maio 2003

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Este relatório apresenta uma introdução à área de Análise de Discurso, fazendo uma revisão bibliográfica das principais teorias discursivas utilizadas em Processamento de Línguas Naturais (PLN), abrangendo os níveis da retórica, da semântica e das intenções. São apresentados, também, os principais pontos dos trabalhos que tentaram estabelecer possíveis mapeamentos entre as teorias, mostrando como estes níveis podem co-existir no discurso. Por fim, são relatadas algumas pesquisas em PLN que fizeram uso de análise discursiva.

Este trabalho conta com o apoio financeiro da FAPESP e do CNPq.

ÍNDICE

1. Introdução

As pesquisas em Lingüística e Lingüística Computacional têm comprovado há tempos que um texto não é constituído por uma simples seqüência de sentenças, mas que possui uma estrutura altamente elaborada. Essa estrutura é composta por componentes lingüísticos e por informação de natureza extralingüística que influencia o processo de escrita do texto e até mesmo o modo como ele é interpretado pelos leitores.

Na área de Processamento de Línguas Naturais (PLN), sistemas como tradutores automáticos, corretores gramaticais e sumarizadores de textos, fazem uso, em sua maioria, de recursos computacionais como etiquetadores morfossintáticos (*taggers*) e analisadores sintáticos (*parsers*). Considerando toda a riqueza da língua, o conhecimento fornecido por *taggers* e *parsers* é pouco diante de todo o conhecimento contido em um texto. Considere o trecho de texto abaixo:

Embora você não goste, trabalhar é importante. O trabalho enobrece o homem.

Uma análise morfossintática destas duas sentenças reconheceria, basicamente, suas palavras e as classes gramaticais a que pertencem. Uma análise sintática produziria as árvores sintáticas das duas sentenças, indicando as funções sintáticas de suas partes (sujeito, predicado, objetos, etc.). No nível semântico, estabelecer-se-ia, usualmente, o significado de cada uma das duas sentenças em algum formalismo específico. A partir deste nível, isto é, a semântica, torna-se cada vez mais complexo o tratamento computacional, dada a natureza das informações tratadas e à ambigüidade inerente à língua.

Até então, as sentenças foram consideradas como unidades isoladas. É possível notar, entretanto, que há mais informação neste trecho de texto do que a produzida por uma análise morfossintática, sintática ou mesmo semântica. Por exemplo, é possível reconhecer (i) que há uma relação de oposição entre as cláusulas "Embora você não goste" e "trabalhar é importante" e (ii) que uma maior importância é atribuída à cláusula "trabalhar é importante" dentro da relação de oposição estabelecida, assim como o fato de que a segunda sentença (ou terceira cláusula) "O trabalho enobrece o homem" justifica porque trabalhar é importante. Cláusula, neste contexto, é uma unidade mínima de significado que possui papel funcional no texto (Dik, 1997)[1]. Nessa própria definição, o termo "papel funcional" refere-se ao fato de que (iii) um texto, assim como suas partes, deve satisfazer um objetivo comunicativo, a função para a qual o texto foi produzido (nos termos de Halliday, 1985), ou seja, as intenções subjacentes ao texto. No exemplo anterior, uma possível intenção seria "persuadir uma pessoa a trabalhar", derivada das intenções da primeira e da segunda cláusula, que poderiam ser, respectivamente, "reconhecer o fato da pessoa não gostar de trabalhar" e "ressaltar a importância

[1] De acordo com a literatura especializada, os termos "cláusula", "segmento textual", "segmento discursivo", "trecho de texto" e "proposição", entre outros, possuem diferentes significados. Entretanto, no contexto das pesquisas de análise automática de discurso, esses termos têm sido usados de forma indiscriminada. Neste relatório, para simplicidade, os termos serão usados de forma intercambiável. É importante esclarecer, entretanto, que, quando se trabalha no nível do discurso, como é o caso aqui, fala-se em proposições, ou segmentos discursivos, que correspondem ao conteúdo de uma cláusula, de um segmento textual qualquer, de uma sentença, ou mesmo de um trecho de texto, dependendo do contexto que se discute.

do trabalho". Ainda, é possível perceber que a intenção da segunda cláusula é dominante em relação à intenção da primeira para a satisfação da intenção primária de "persuadir uma pessoa a trabalhar".

Como pode ser notado, há relações intra e intersentenciais em um texto de várias naturezas. São estas relações que atribuem sentido a um texto, tornando-o coerente e caracterizando-o, de fato, como um texto (Koch e Travaglia, 1990; Koch, 1998). Apesar da importância destas relações, poucos sistemas computacionais são capazes de reconhecê-las, dada a subjetividade, complexidade e ambigüidade inerente a este nível de análise.

Em PLN, a relação entre complexidade e os níveis de análise comumente distinguidos pode ser esquematizada como mostra a Figura 1. Quanto mais se sobe em direção aos níveis da Pragmática e do Discurso, mais complexos são a modelagem e o tratamento computacional. As relações discriminadas anteriormente encontram-se nos níveis mais complexos e abstratos, isto é, nos níveis da Semântica e da Pragmática e do Discurso. Análises desta natureza, capazes de reconhecer estas relações, fazem parte da área conhecida como "Análise de Discurso".

Uma ferramenta capaz de extrair as relações semânticas e pragmáticas/discursivas de um texto, montando sua estrutura discursiva, seria de grande utilidade para o PLN, podendo auxiliar, por exemplo, em problemas de Resolução de Anáforas, Tradução Automática e Sumarização de Textos.

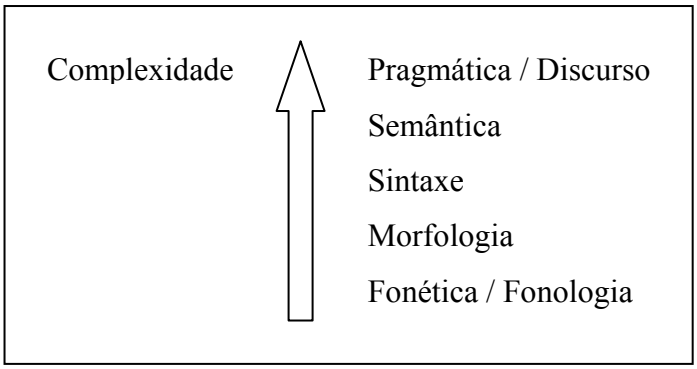

Figura 1 – Complexidade & níveis de conhecimento

Para a língua portuguesa, sabe-se de apenas alguns trabalhos que abordam o tratamento (computacional) discursivo como é comentado aqui. Rino (1996a), por exemplo, define uma modelagem discursiva no contexto da sumarização automática que é aplicável para o português (Rino, 1996b; Pardo, 2002). Ela sugere que a informação a ser selecionada para se formar um sumário deve ser deduzida com base semântica e intencional. Feltrim et. al. (2000, 2001) investigam a estruturação esquemática de textos científicos em português com o intuito de construir uma ferramenta de auxílio à escrita. Para o inglês, por sua vez, muitos trabalhos se destacam na área de processamento de discurso. Vários deles são comentados nas próximas seções.

A revisão apresentada neste relatório é parte da investigação para o desenvolvimento de um analisador discursivo para o português do Brasil, chamado DiZer (*DIscourse AnalyZER for BRazilian Portuguese*), cuja função é produzir a estrutura discursiva de um texto dado como entrada. Esta estrutura discursiva é formada pelas estruturas retórica e semântica e por relações intencionais. As teorias

discursivas correspondentes a estes níveis de análise são apresentadas na próxima seção. No nível retórico, será apresentada a RST (*Rhetorical Structure Theory*) (Mann and Thompson, 1987); no nível semântico, serão explicadas as relações semânticas de Jordan (1992); por fim, no nível intencional, serão introduzidas as relações intencionais da GSDT (*Grosz and Sidner Discourse Model*) (Grosz and Sidner, 1986). A Seção 3 descreve os principais pontos de alguns trabalhos que implementaram analisadores discursivos, mais especificamente, analisadores retóricos. A Seção 4 relata algumas pesquisas recentes em PLN que fazem uso de análise discursiva. A Seção 5 apresenta algumas conclusões.

2. Teorias Discursivas

2.1. Retórica

A retórica é o meio pelo qual um texto é organizado para satisfazer o objetivo comunicativo que o escritor do texto pretendia atingir ao escrevê-lo[2]. São as relações retóricas em um texto que delineiam como seu conteúdo é relacionado e como e em que medida cada segmento textual contribui para a satisfação das intenções do escritor. Diz-se, portanto, que as relações retóricas são responsáveis por atribuir coerência a um texto, relacionando suas partes e atribuindo-lhe sentido. Por essa razão, elas também são comumente chamadas de *relações de coerência*.

A RST (Mann and Thompson, 1987) estabelece relações retóricas que, de acordo com os autores, podem ser aplicadas a uma grande gama de textos sem necessitarem de modificações ou inclusões de outras relações[3]. As relações são mostradas na primeira coluna da Tabela 1 (em inglês, como na obra de referência).

Tabela 1 – Relações retóricas da RST (Mann and Thompson, 1987)

Relações	**Multinuclear**	**Tipo das Relações**
Antithesis	Não	*Presentational*
Background	Não	*Presentational*
Circumstance	Não	*Subject Matter*
Concession	Não	*Presentational*
Condition	Não	*Subject Matter*
Elaboration	Não	*Subject Matter*
Enablement	Não	*Presentational*
Evaluation	Não	*Subject Matter*
Evidence	Não	*Presentational*
Interpretation	Não	*Subject Matter*
Justify	Não	*Presentational*
Means	Não	*Subject Matter*
Motivation	Não	*Presentational*
Non-volitional Cause	Não	*Subject Matter*
Non-volitional Result	Não	*Subject Matter*

[2] Um escritor, ao produzir seu texto, pode ter mais de um objetivo comunicativo a satisfazer. Entretanto, um deles costuma ser proeminente, sendo este considerado pelas pesquisas em análise discursiva o objetivo comunicativo do discurso.

[3] Mann e Thompson ressaltam, entretanto, que alguns tipos de textos não são analisáveis por relações retóricas, citando como exemplo os textos jurídicos.

Otherwise	Não	*Subject Matter*
Purpose	Não	*Subject Matter*
Restatement	Não	*Subject Matter*
Solutionhood	Não	*Subject Matter*
Summary	Não	*Subject Matter*
Volitional Cause	Não	*Subject Matter*
Volitional Result	Não	*Subject Matter*
Contrast	Sim	*Subject Matter*
Joint	Sim	*Subject Matter*
List	Sim	*Subject Matter*
Sequence	Sim	*Subject Matter*

Em casos padrões, essas relações se estabelecem entre dois trechos de texto, geralmente adjacentes, sendo um nuclear (N) e outro complementar (S – "satélite"), indicando, respectivamente, a informação principal para a satisfação da intenção subjacente à relação e uma informação adicional, a qual influencia de alguma forma a interpretação que o leitor faz da informação nuclear. Quando ambas as informações relacionadas são igualmente importantes, diz-se que se tem uma relação multinuclear, isto é, com mais de um núcleo e nenhum satélite. Por exemplo, no trecho de texto "Embora você não goste, trabalhar é importante.", a primeira cláusula é o satélite e a segunda é o núcleo da relação retórica *concession* que existe entre as cláusulas. Por sua vez, no trecho de texto "O garoto chegou da escola e fez sua lição de casa. Depois, foi brincar com os amigos.", há uma relação *sequence* (indicando uma "seqüência" de eventos) entre as cláusulas "O garoto chegou da escola", "e fez sua lição de casa." e "Depois, foi brincar com os amigos.", sendo que todas são consideradas núcleos da relação, pois possuem a mesma importância. A segunda coluna da Tabela 1 identifica as relações multinucleares.

A definição de cada relação pressupõe a existência de quatro tipos de informação necessários para determinar sua ocorrência entre dois trechos de texto:

- Restrições sobre o núcleo (N);
- Restrições sobre o satélite (S);
- Restrições sobre a combinação do núcleo e do satélite (N+S);
- Efeito (ou intenção do escritor): especificação do efeito que a relação em questão causa no leitor, quando este interpreta um texto, ou do efeito pretendido pelo escritor, quando este seleciona tal relação para estruturar seu texto.

Como exemplo, as Figuras 2, 3, 4 e 5 mostram as definições das relações retóricas *concession*, *background*, *evidence* e *elaboration*.

Nome da relação: *concession*

Restrições sobre N: o escritor julga N válido.
Restrições sobre S: o escritor não afirma que S pode não ser válido.
Restrições sobre N+S: o escritor mostra uma incompatibilidade aparente ou em potencial entre N e S; o reconhecimento da compatibilidade entre N e S melhora a aceitação de N pelo leitor.
Efeito: o leitor aceita melhor N

Figura 2 – Definição da relação retórica *concession*

Nome da relação: *background*
Restrições sobre N: o leitor não compreenderá suficientemente N antes de ler S. **Restrições sobre S**: sem restrições. **Restrições sobre N+S**: S aumenta a habilidade do leitor em compreender algum elemento em N. **Efeito**: a habilidade do leitor para compreender N aumenta.

Figura 3 – Definição da relação retórica *background*

Nome da relação: *evidence*
Restrições sobre N: o leitor poderia não acreditar em N de forma satisfatória para o escritor. **Restrições sobre S**: o leitor acredita em S ou o achará válido. **Restrições sobre N+S**: a compreensão de S pelo leitor aumenta sua convicção em N. **Efeito**: a convicção do leitor em N aumenta.

Figura 4 – Definição da relação retórica *evidence*

Nome da relação: *elaboration*
Restrições sobre N: sem restrições. **Restrições sobre S**: sem restrições. **Restrições sobre N+S**: S apresenta detalhes adicionais sobre a situação ou algum elemento de N. **Efeito**: o leitor reconhece S como apresentando detalhes adicionais sobre N.

Figura 5 – Definição da relação retórica *elaboration*

Como exemplo de análise, a Figura 7 mostra a análise retórica do texto da Figura 6 (cujos segmentos foram numerados para referência), fazendo uso de algumas das relações ilustradas anteriormente, onde os arcos com os rótulos N e S indicam, respectivamente, os núcleos e satélites das relações. Esse texto é do gênero científico, do domínio da Computação, e é parte da introdução de uma dissertação de Mestrado extraída do corpus do NILC[4] (Núcleo Interinstitucional de Lingüística Computacional) (Pinheiro e Aluísio, 2003; Kuhn et al., 2000).

[1] A representação de grandes dicionários de língua natural, principalmente nos casos em que se trabalha com vários milhões (ou dezenas de milhões) de palavras, é um interessante problema computacional a ser tratado dentro da área de Processamento de Língua Natural. [2] Autômatos finitos, largamente usados na construção de compiladores, são excelentes estruturas para representação desses dicionários, [3] permitindo acesso direto aos às palavras e seus possíveis atributos.

[4] Um dicionário contendo mais de 430.000 palavras da língua portuguesa sem atributos, cuja representação em formato texto ocupa mais de 4.5Mb, pode ser convertido em um autômato compactado de apenas 218Kb.

Figura 6 – Texto 1

[4] O NILC é formado por pesquisadores em Lingüística Computacional da USP/São Carlos, da UFSCar e da UNESP/Araraquara. (www.nilc.icmc.usp.br)

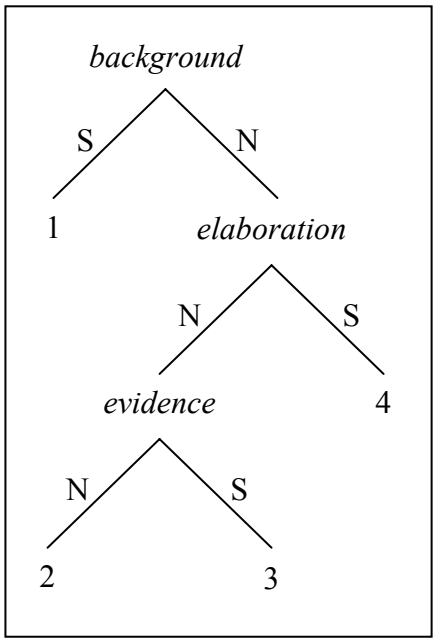

Figura 7 – Estrutura retórica do Texto 1

Dependendo dos objetivos de pesquisa, as relações retóricas da RST podem ser organizadas de várias maneiras, observando-se, por exemplo, a participação do escritor na relação, a participação do leitor na relação, a ordem mais comum de realização do núcleo e do satélite no texto para cada relação (que pode ser núcleo antes do satélite ou satélite antes do núcleo), etc. A terceira coluna da Tabela 1 faz uma divisão em termos do efeito que as relações causam, que é a organização mais interessante de se analisar. Relações *presentational*, também chamadas de "intencionais" ou "pragmáticas", são aquelas que interferem em alguma inclinação do leitor, podendo aumentar seu desejo ou sua habilidade de realizar uma ação, crenças, aceitações e convicções; relações *subject matter*, também chamadas de "semânticas" ou "informativas", são aquelas cujo efeito pretendido é o de que o leitor apenas reconheça a existência da relação entre dois trechos de texto, não interferindo em suas inclinações pessoais. A relação *elaboration* é um exemplo de relação *subject matter*, pois seu efeito pretendido é que o leitor reconheça que o satélite apresenta detalhes adicionais sobre o núcleo; a relação *evidence*, por sua vez, é uma relação *presentational* porque aumenta a convicção do leitor no núcleo da relação.

As relações retóricas *subject matter*, ou semânticas, não são iguais às relações semânticas de Jordan (1992) (introduzidas na próxima subseção), pois as primeiras, apesar de não interferirem nas inclinações pessoais do leitor, ainda estão ligadas às intenções do escritor do texto por meio da determinação do que é nuclear ou não em um texto, ou seja, qual segmento é mais importante para que a relação atinja seu efeito pretendido (ou seja, sua intenção). As diferenças entre os tipos de relações usadas serão mais bem explicadas na próxima subseção.

A RST foi aplicada para várias finalidades dentro das áreas da interpretação e da geração de textos em PLN, abrindo as portas para um estudo mais sistemático da estrutura textual. Entretanto, a RST esbarra em problemas como a ambigüidade retórica e a falta de uma formalização robusta da teoria, além da complicada interação com as intenções subjacentes aos textos: a) a ambigüidade retórica é um

problema inerente à língua natural; b) a falta de uma formalização robusta diz respeito à dificuldade de se escolher uma única relação retórica para relacionar trechos de texto baseando-se somente nas definições das relações fornecidas por Mann e Thompson; c) as intenções permeiam a escolha de uma relação retórica, contudo, não são explicitadas pelas relações.

Vários autores modificaram, complementaram ou discutiram a especificação da RST, buscando uma maior clareza das relações para fins computacionais. A dificuldade em se definir relações dessa natureza reside no caráter interpretativo e subjetivo da língua, posto que remete à apreensão do significado. Ainda, tais relações devem supostamente cobrir todas as possíveis relações existentes no universo comunicativo. As pesquisas de Marcu (1997, 2000a), em especial, contribuíram bastante para a resolução dos problemas da RST, como será comentado na próxima seção.

2.2. Semântica

De acordo com Jordan (1992), uma relação semântica constitui uma "noção semântica textual de conexão binária entre quaisquer duas partes de um texto". Como as partes conectadas são geralmente cláusulas, as relações semânticas são também chamadas de relações clausais ou relações interclausais.

As relações semânticas de Jordan são, na realidade, um amálgama das relações propostas em vários outros trabalhos importantes em PLN, destacando-se os trabalhos de Winter (1968, 1971, 1974, 1976, 1977, 1979, 1982), Hoey (1979, 1983a, 1983b), Hoey e Winter (1986) e do próprio Jordan (1978, 1985a, 1985b, 1988, 1989). A lista completa de relações semânticas propostas por Jordan é mostrada na primeira coluna da Tabela 2. A segunda coluna indica a tipologia das relações definida pelo próprio autor.

Tabela 2 – Relações semânticas de Jordan (1992)

Relações	**Tipos das Relações**
Identification	*Detail*
Classification	
Specification	
Appearance	
Characteristics	
Function	
Material	
Parts	
Active	*General*
Passive	
Agent	
Source	
Assessment	*Logical*
Basis	
Cause	
Effect	
Emotive Effect	
Purpose	
Means	

Problem	
Solution	
Possibility	*Modal*
Capability	
Correctness	
Propriety	
Necessity	
Need	
Completion	
Achievement	
Future	
Intention	
Mandate	
Authority	
Determination	
Permission	
Obligation	
Willingness	
Desire	
Time	*Time*
Before	
After	
Simultaneous	
Inverted time	
Elaboration	*Text manipulation*
Summary	
Repetition	
Paraphrase	
Forecast	
Transition	
Collateral inversion	*Special*
Concession	
Compatibility	
Contrast	
Comparison	
Conditionals	
Document structures	
Hypothetical-Real	
Transition couplets	
Accompaniment	*Other*
Circumstance	
Inverted circumstance	
Connection	
Enablement	
Example	
Extent	
Location	

Inverted Location	
Manner	
True	

Segundo Jordan, estas relações capturam a forma como os conhecimentos contidos em um texto se relacionam, sendo completamente desvinculadas das intenções do escritor. De fato, esta característica das relações semânticas é o que as diferenciam das relações retóricas, principalmente das relações retóricas *subject matter*, ou semânticas, da RST. Neste caso específico, apesar das relações *subject matter* e das relações de Jordan estabelecerem relações entre os conteúdos proposicionais de trechos de textos, não interferindo nas inclinações pessoais do leitor, as relações retóricas *subject matter* identificam o que é nuclear ou não para a satisfação do efeito pretendido (a intenção subjacente), enquanto as relações semânticas "puras" de Jordan, não (como sugerem, também, Moser e Moore, 1996). Por exemplo, para o trecho de texto "Um incêndio destruiu várias casas. Algumas pessoas foram para o hospital.", tanto a relação retórica *non-volitional cause* quanto a relação *non-volitional result* poderiam ser utilizadas para relacionar as duas sentenças: se o trecho mais importante para a satisfação do objetivo comunicativo do escritor do texto for o primeiro (que fala do incêndio), a relação *non-volitional result* deve ser usada; caso contrário, se o trecho mais importante para a satisfação do objetivo comunicativo do escritor for o segundo (que fala das pessoas que foram para o hospital), a relação *non-volitional cause* deve ser usada. As relações semânticas de Jordan, por sua vez, apenas estabelecem a relação existente entre dois conteúdos proposicionais, não tendo a função de indicar o que é mais importante ou não para qualquer que seja o efeito pretendido pelo escritor do texto. Para o exemplo anterior, haveria apenas uma relação semântica de causa/efeito entre as duas sentenças, indicando que o incêndio causou o fato de algumas pessoas terem ido para o hospital, não atribuindo, assim, maior importância a nenhuma das sentenças.

Assim como na RST, entre os problemas encontrados na determinação das relações semânticas em um texto, há a ambigüidade inerente a este nível de análise, ou seja, quando mais de uma relação é possível entre duas cláusulas (vide o grande elenco de relações da Tabela 2 e a tênue diferença entre algumas delas). Jordan afirma que, nestes casos, é ingenuidade tentar definir uma única relação. Ele sugere que todas as relações possíveis sejam incorporadas à estrutura discursiva do texto sob análise.

A Tabela 3 mostra um subconjunto das relações semânticas de Jordan utilizado por Rino (1996a) em seu modelo de discurso. Por ser bem menor e mais genérico, é possível se determinar uma única relação para conectar as partes de um texto.

Tabela 3 – Relações semânticas de Rino (1996a)

Relações
Enable
Rationale
Proof
Cause
List
Contrast
Attribute

Detail
Exemplify
Evaluation
Reason
Sequence
Background

Como exemplo de análise semântica, a Figura 8 mostra a estrutura semântica do Texto 1 (Figura 6) utilizando o subconjunto de relações de Rino. Esta estrutura é interpretada da seguinte forma: o segmento 1 contextualiza (relação *background*) os segmentos 2-4 (isto é, os segmentos de 2 a 4); o segmento 4 exemplifica (relação *exemplify*) o que é apresentado nos segmentos 2-3; o segmento 3 dá provas (relação *proof*) do que é dito no segmento 2.

Para o texto considerado e para a interpretação apreendida, pode-se notar que a estrutura semântica é isomórfica à estrutura retórica, isto é, suas "árvores" possuem o mesmo formato. Porém, a isomorfia não é regra geral. Em muitos casos, as estruturas semânticas e retóricas são diferentes.

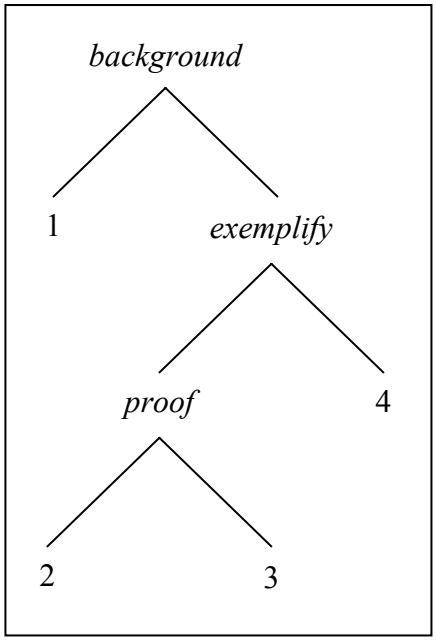

Figura 8 – Estrutura semântica do Texto 1

2.3. Intenções

Segundo Grosz e Sidner (1986), todo discurso é essencialmente produzido com a finalidade de satisfazer uma ou mais intenções. São as intenções que individualizam e tornam coerente o discurso.

De acordo com a GSDT, quando um escritor escreve seu texto, ele produz e estrutura o conteúdo do texto em função de suas intenções. Nestes termos, há dois tipos de intenções: a intenção primária do discurso e as intenções dos segmentos do discurso, as quais devem contribuir para a satisfação da intenção primária. Por exemplo, para o texto da Figura 6, a intenção primária pode ser "convencer o leitor de que autômatos finitos são excelentes estruturas para a representação de grandes dicionários de língua natural". Por sua vez, a intenção do segmento 1 pode

ser "fazer com que o leitor saiba que a representação de grandes dicionários de língua natural é um interessante problema computacional a ser tratado", que, assim como as intenções dos outros segmentos, contribui para a satisfação da intenção primária. É o reconhecimento destas intenções que permite ao leitor recuperar o que o escritor pretendia comunicar com o texto.

Como as intenções possíveis em um discurso são infinitas (tanto as primárias quanto as dos segmentos), a GSDT organiza o discurso por meio de relações de contribuição e satisfação entre as intenções, que, por sua vez, são finitas. Estas relações, chamadas de relações intencionais, são duas:

- *dominance* (DOM): se a intenção de um segmento Y contribui para a satisfação da intenção de um segmento X, então se diz que a intenção de X domina (*dominates*) a intenção de Y, ou seja, DOM(X,Y);
- *satisfaction-precedence* (SP): se a intenção de um segmento X deve ser satisfeita antes da intenção de um segmento Y, então se diz que a satisfação da intenção de X deve preceder (*satisfaction-precedes*) a satisfação da intenção de Y, ou seja, SP(X,Y).

Como exemplo, para o texto da Figura 6, as seguintes relações valem: DOM([2-4],1), DOM([2-3],4) e DOM(2,3), nas quais [X-Y] indica o segmento discursivo formado pelos segmentos textuais de X a Y. Não há, neste discurso, nenhuma relação SP "significativa" (nos termos da GSDT), as quais são muito comuns em discursos que apresentam seqüências de eventos.

A GSDT também define duas relações informativas que, como as relações semânticas descritas na subseção anterior, não se estabelecem entre intenções, mas entre o conteúdo proposicional dos segmentos. Elas são:

- *supports* (SUP): se a crença do leitor no conteúdo de um segmento Y fornece subsídios para que o leitor creia no conteúdo de um segmento X, então se diz que o conteúdo de Y sustenta (*supports*) o conteúdo de X, ou seja, SUP(Y,X);
- *generates* (GEN): se a execução de uma ação descrita no segmento Y contribui para a execução de uma ação descrita no segmento X, então se diz que Y gera (*generates*) X, ou seja, GEN(Y,X).

Grosz e Sidner determinaram, ainda, uma correspondência direta entre a relação DOM e as relações SUP e GEN. Para X e Y quaisquer, tem-se as seguintes correspondências:

- no caso de crença, DOM(X,Y) $\Leftrightarrow$ SUP(Y,X);
- no caso de ação, DOM(X,Y) $\Leftrightarrow$ GEN(Y,X).

nas quais o símbolo $\Leftrightarrow$ representa uma implicação de dois sentidos. Em outras palavras, caso se verifique que o discurso trata de crenças (ações), é possível determinar DOM em função de SUP (GEN), assim como SUP (GEN) em função de DOM.

Há uma forte correlação entre a GSDT e a RST. Como foi dito anteriormente, as relações retóricas da RST são determinadas na análise de um texto em função das intenções percebidas. Entretanto, tais intenções não são feitas explícitas nas definições das relações, mesmo porque são infinitas. Como a GSDT modela justamente as intenções do discurso em termos das relações intencionais, diz-se que ela complementa a RST e vice-versa. A próxima subseção relata alguns trabalhos

que evidenciaram tal relacionamento, assim como o relacionamento entre o nível semântico e o retórico.

2.4. Mapeamento entre Relações Discursivas

Como se pôde notar nas subseções anteriores, é possível analisar um discurso em vários níveis, com diferentes perspectivas e objetivos em mente. Na literatura, muito se tem discutido como tais níveis podem co-existir em um discurso, como um nível influencia o outro e se há ou não mapeamentos possíveis entre eles. A seguir, são relatadas as principais pesquisas que estabeleceram possíveis relacionamentos e mapeamentos entre os níveis discursivos. A subseção 2.4.1 mostra a relação entre as relações retóricas e as intenções, enquanto a subseção 2.4.2 mostra a relação entre as relações retóricas e as semânticas. Os problemas encontrados e as soluções sugeridas na literatura são discutidos.

2.4.1. Retórica e Intenções

Pesquisas recentes na área de Análise de Discurso têm mostrado e concordado com o fato de que a retórica é a forma de expressão das intenções no discurso (Maier e Hovy, 1991; Maybury, 1992; Dale, 1993; Hovy, 1993; Maier, 1993; Moore e Paris, 1993; Moore, 1995; Moser e Moore, 1996; Rino, 1996a; Marcu, 1999, 2000b; Pardo (2002); etc.). Assim, quando um escritor produz um texto, ele tem em mente um objetivo comunicativo, uma intenção, a atingir. Por meio das relações retóricas, ele organiza e estrutura o conteúdo textual de forma que sua intenção seja satisfeita. Ressalta-se, porém, que este mapeamento de intenções em relações retóricas não é simples, pois (a) uma mesma intenção pode ser realizada por diferentes estratégias retóricas e (b) uma mesma estratégia retórica pode servir para a realização de diferentes intenções. Como exemplo desse relacionamento complexo, a Tabela 4 mostra parte do mapeamento identificado por Moore e Paris entre intenções e relações retóricas para a aplicação que desenvolveram, que consistia em um sistema de diálogo para gerar explicações. Por exemplo, a intenção de "capacitar o leitor a identificar algo" (segunda linha da tabela) pode ser realizada pelas relações retóricas *circumstance*, *condition*, *contrast*, etc. Por outro lado, a relação retórica *contrast* pode ser a realização da intenção de "capacitar o leitor a identificar algo" e "fazer com que o leitor acredite em uma proposição".

Tabela 4 – Mapeamento de intenções em relações retóricas de Moore e Paris (1993)

Intenções	**Relações Retóricas**
▪ persuadir o leitor sobre uma proposição ▪ persuadir o leitor a realizar uma ação ▪ tornar o leitor competente para compreender algo ▪ tornar o leitor competente para realizar uma ação	*evidence* *motivation* *background* *enablement*
▪ capacitar o leitor a identificar algo	*circumstance* *condition* *contrast* *elaboration* *purpose* *sequence*
▪ fazer com que o leitor acredite em uma proposição	*contrast* *elaboration*

Por serem as teorias discursivas mais representativas sobre intenções e retórica, respectivamente, a GSDT e a RST têm sido objetos de estudo de muitas pesquisas. As relações da RST, em especial, têm implícitas no campo "Efeito" de suas definições as intenções que pretendem atingir. Por exemplo, a relação *concession* pretende que o leitor aumente sua convicção no conteúdo do segmento nuclear, qualquer que seja este conteúdo. A GSDT, reconhecendo a infinidade de intenções existentes, propôs a estruturação discursiva com base na relação entre as intenções do discurso, que, por sua vez, são finitas.

Alguns trabalhos tentaram unir as duas teorias e seus pressupostos teóricos. Apesar das diferenças evidentes entre a RST e a GSDT, que vão desde o que se considera como unidade elementar de análise (unidades mínimas de significado *vs* uma ou mais unidades mínimas seqüenciais que satisfazem uma intenção) até a própria estruturação (retórica *vs* intencional), similaridades foram encontradas, permitindo o mapeamento (às vezes parcial) de um nível no outro.

Moser e Moore (1996) foram as primeiras a reconhecer a correspondência entre os conceitos de nuclearidade da RST e de dominância da GSDT. Elas verificaram que em uma relação retórica padrão, isto é, com um núcleo e um satélite, a intenção subjacente ao núcleo domina (DOM) a intenção subjacente ao satélite (o caminho inverso também vale, ou seja, quando um segmento domina outro, pode-se dizer que o primeiro será o núcleo de uma relação retórica e o segundo o satélite). As relações de dominância dadas como exemplo na subseção 2.3 para o texto da Figura 6 foram derivadas desta forma a partir da estrutura retórica da Figura 7 e, como se pode verificar, estão corretas.

No caso de relações multinucleares, onde não há um satélite para ser dominado pelo(s) núcleo(s), Moser e Moore levantam a hipótese de que, talvez, a RST e a GSDT possuam pressupostos teóricos incompatíveis neste ponto. Justamente nesta questão, Marcu (1999) estendeu o trabalho de Moser e Moore afirmando que, em relações multinucleares, não há dominância entre os segmentos, mas "pode" haver precedência (SP). Posteriormente, Marcu (2000b) formalizou esse mapeamento entre relações retóricas e intencionais. Além disso, mostrou que é possível derivar a intenção primária de um discurso por meio de sua estrutura retórica. Segundo ele, a intenção primária é dada pelo segmento mais nuclear da estrutura retórica em conjunto com a relação retórica correspondente. Na estrutura da Figura 7, por exemplo, a intenção primária seria dada pela combinação do segmento 2, que é o segmento mais nuclear da estrutura, com a relação retórica *evidence*, resultando na intenção primária "aumentar a convicção do leitor de que autômatos finitos são excelentes estruturas para a representação de dicionários de língua natural", na qual o fato de "aumentar a convicção do leitor em algo" vem do campo "Efeito" da definição da relação *evidence*, enquanto o fato de que "autômatos finitos são excelentes estruturas para a representação de dicionários de língua natural" vem do conteúdo proposicional do segmento 2.

Da mesma forma que as relações intencionais podem ser determinadas a partir da estrutura retórica, Marcu enfatiza que as relações intencionais também podem ser usadas para restringir as possíveis estruturas retóricas de um texto em um processo automático de análise, caso estas relações estejam disponíveis de antemão, o que, normalmente, não se observa em problemas reais de PLN.

Moser e Moore e Marcu partiram dos pressupostos teóricos da RST e da GSDT para determinar o relacionamento entre retórica e intenções. Em uma linha oposta (empírica), Rino (1996a), por meio de análise de corpora, determinou um

mapeamento de relações semânticas e intencionais em relações retóricas. Esse mapeamento (aprimorado por Pardo, 2002) é mostrado na Tabela 5[5].

Tabela 5 – Mapeamento de Rino (1996a)

Caso	**Relações semânticas**	**Relações retóricas[6]**	**Relações intencionais**
1	*enable*(Y,X)	*purpose1*(Y,X) *means*(X,Y)	SP(X,Y) DOM(Y,X)
2	*rationale*(Y,X)	*purpose2*(X,Y) *justify1*(X,Y)	SUP(Y,X) DOM(X,Y) not SP(X,Y)
3	*proof*(X,Y)	*evidence*(X,Y) *justify2*(X,Y)	SP(Y,X) DOM(X,Y) SUP(Y,X)
4	*cause*(X,Y)	*nonvolres*(Y,X) *nonvolcause*(X,Y)	SP(Y,X) DOM(X,Y) GEN(Y,X)
5	*simSem*(X,Y)	*list*(X,Y)	DOM(X,Y) DOM(Y,X)
6	*difSem*(X,Y)	*contrast*(X,Y)	DOM(X,Y) DOM(Y,X)
7	*attribute*(X,Y) *detail*(X,Y) *exemplify*(X,Y)	*elaborate*(X,Y)	SUP(Y,X)
8	*evalSem*(X,Y)	*evaluate*(X,Y)	DOM(X,Y)
9	*reason*(X,Y)	*explain*(X,Y)	GEN(Y,X)
10	*sequence*(X,Y)	*sequence*(X,Y)	SP(X,Y)
11	*backSem*(X,Y)	*background*(X,Y)	not SP(X,Y) SUP(Y,X)

O mapeamento funciona da seguinte forma: dada uma relação semântica entre duas proposições X e Y em uma base de conhecimento qualquer e as relações intencionais entre estas proposições, é possível produzir uma relação retórica envolvendo as duas proposições. De forma inversa, dada uma relação retórica, é possível desmembrá-la em uma relação semântica e algumas relações intencionais.

O mapeamento de Rino é muito interessante pelo fato de concretizar a distinção que se tem feito entre retórica e semântica: a retórica tem força argumentativa, enquanto a semântica não. Portanto, uma conclusão que pode ser obtida do mapeamento de Rino é a de que a retórica é a semântica acrescida de força argumentativa, que, neste caso, é dada pelas relações intencionais. De fato, em seus experimentos, Pardo constatou que são as relações intencionais que atribuem às proposições o caráter nuclear ou não durante a construção da estrutura retórica.

Apesar de ser um mapeamento interessante, Rino diverge das pesquisas anteriores nos seguintes pontos:

[5] Para mais informações sobre as definições das relações e exemplos de mapeamento em textos reais, vide as obras de referência.

[6] Assume-se, nesta coluna, que o primeiro argumento das relações retóricas é o núcleo, excetuando-se os casos de relações multinucleares (*list*, *contrast* e *sequence*).

1. a relação entre intenções e retórica no mapeamento de Rino não está em conformidade com a questão da nuclearidade refletir a dominância e vice-versa (por exemplo, casos 1 e 4 do mapeamento);
2. em alguns casos do mapeamento, há redundâncias no nível intencional (por exemplo, casos 2, 3 e 4 do mapeamento): nos casos em que existem as relações *supports* (SUP) e *generates* (GEN), não é necessário que se defina a relação de dominância (DOM), pois esta pode ser deduzida das primeiras.

2.4.2. Retórica e Semântica

O mapeamento entre retórica e semântica não foi tão explorado e formalizado quanto o mapeamento entre retórica e intenções. Alguns trabalhos, como Moser e Moore (1996), Moore e Pollack (1992), Hovy (1991, 1993) e Rino (1996a), defenderam que um texto deve ter as relações semânticas incorporadas à sua estrutura retórica, mas não explicitaram como as relações retóricas e semânticas se relacionam no discurso.

Pelo mapeamento de Rino (mostrado na subseção anterior), pode-se inferir que a retórica é a semântica acrescida de força argumentativa. Korelsky e Kittredge (1993) sugerem que as relações retóricas se estabelecem entre proposições relacionadas semanticamente. Eles ressaltam que, como acontece no mapeamento entre intenções e retórica, uma relação retórica pode se servir de várias relações semânticas no discurso, assim como uma relação semântica pode ser interpretada como várias relações retóricas. Por exemplo, uma relação retórica *evidence* pode ser observada entre proposições conectadas pelas relações semânticas *volitional cause* (trecho de texto 1), *non-volitional cause* (trecho de texto 2) e *elaboration* (trecho de texto 3), como mostram os trechos de texto abaixo retirados integralmente dos trabalhos de Korelsky e Kittredge (trechos 1 e 2) e de Mann e Thompson (1987) (trecho 3).

(1) *George Bush supports big business.*
He's sure to veto House Bill 1711.

(2) *Winters in Montreal are so cold.*
I need a fur coat.

(3) *George Bush definitely supports big business.*
He just voted House Bill 1711.

Para exemplificar o caso contrário, Korelsky e Kittredge usam o trecho de texto 4 abaixo, no qual há uma relação semântica *condition*, podendo-se reconhecer tanto a relação retórica *enablement* quanto a relação *motivation*.

(4) *Come home by 5:00.*
Then we can go to the hardware store before it closes.

Como se pode notar, as relações semânticas citadas nos exemplos acima são as relações retóricas *subject matter* da RST. Apesar de não serem iguais às relações semânticas "puras" de Jordan, as observações feitas anteriormente continuam válidas para estas. Moser e Moore, inclusive, reconhecem que as relações retóricas *subject*

matter não deveriam envolver nuclearidade, o que as tornariam, então, similares às relações semânticas de Jordan (1992).

Korelsky e Kittredge sugerem algoritmos para determinar a relação semântica a partir da retórica. Para o caso da relação retórica *evidence* e suas correspondentes semânticas, o algoritmo da Figura 9 é dado como exemplo.

Se a relação retórica *evidence* é observada entre duas proposições P1 e P2, em que P1 é o núcleo e P2 é o satélite, então

1) se há um agente consciente de tal forma que P1 e P2 fazem referência a suas ações, então a relação semântica *volitional cause* se estabelece entre as proposições
2) se não há um agente consciente, então a relação semântica *non-volitional cause* se estabelece entre as proposições
3) se P2 é uma proposição genérica, então a relação semântica *elaboration* se estabelece entre as proposições

Figura 9 – Algoritmo de Korelsky e Kittredge (1993) para mapeamento da relação retórica *evidence* em possíveis relações semânticas

Na mesma linha de Korelsky e Kittredge, Hovy (1991) sugere que as próprias definições das relações retóricas sejam enriquecidas com as informações semânticas. Entretanto, essa mudança na definição das relações retóricas causaria vários problemas (como apontado por Moore e Pollack): perda de modularidade das análises retórica e semântica no tratamento discursivo; proliferação das definições das relações, já que uma relação retórica pode ser observada com a ocorrência de várias relações semânticas; e, mais importante, as estruturas retórica e semântica podem não ser isomórficas, ou seja, podem possuir formatos estruturais diferentes. Os dois últimos problemas parecem ser ignorados pela maioria das pesquisas na área.

O principal problema, entretanto, parece ser o não isomorfismo entre as estruturas retórica e semântica. De fato, Dale (1993) sugere que a estrutura semântica de um texto não é uma árvore binária, mas um grafo. Isso ocorre pelo fato de uma relação semântica não ter as mesmas restrições de uma relação retórica, como conectar segmentos de texto adjacentes e estabelecer o que é núcleo ou não na relação. Teoricamente, mais de uma relação semântica pode se estabelecer entre quaisquer dois segmentos do texto (mesmo não adjacentes) e em qualquer direção (pois não se tem núcleo ou satélite para indicar a direção da relação). Para resolver esse problema, Moser e Moore sugerem que as relações semânticas sejam "parasitas" das relações retóricas, isto é, que elas se estabeleçam somente entre os segmentos das relações retóricas em questão (como fazem Korelsky e Kittredge), o que faria com que as estruturas retóricas e semânticas se tornassem isomórficas. A direção da relação, neste caso, seria indiferente, pois poderia ser facilmente acomodada na estrutura discursiva, uma vez que existe isomorfismo. Fazendo isso, Moser e Moore vão contra a afirmação de Jordan (1992) de ser "ingenuidade" tentar escolher uma só relação entre segmentos textuais. Deve-se ressaltar, entretanto, que os interesses de Moser e Moore e de Jordan são diferentes: Moser e Moore estavam preocupadas em formalizações que tornassem o problema computacionalmente tratável; Jordan, por sua vez, utilizava as relações semânticas como uma ferramenta descritiva de textos.

A próxima seção apresenta os pontos principais de alguns trabalhos que desenvolveram analisadores retóricos para a língua inglesa, visto que a retórica é o nível discursivo que tem recebido mais atenção nas pesquisas recentes.

3. Analisadores Retóricos

Recentemente, em PLN, vários trabalhos têm apresentado modelos formais e metodologias para o desenvolvimento de analisadores retóricos, destacando-se os trabalhos de Marcu (1997, 2000a), Carlson e Marcu (2001), Marcu e Echihabi (2002), Corston-Oliver (1998), Schilder (2002), Kurohashi e Nagao (1994) e Sumita et al. (1992). Os trabalhos listados mais relevantes são introduzidos na próxima subseção, com especial enfoque nos trabalhos de Marcu, por este ser um dos pesquisadores mais proeminentes na área atualmente. Ao final desta seção, é apresentada uma breve discussão sobre os marcadores discursivos, já que estes são utilizados por todos os trabalhos acima como principal base para a análise retórica.

3.1. Analisadores Retóricos

3.1.1. Marcu

Marcu (1997, 2000a) desenvolveu o primeiro *parser* retórico (conforme denominado por ele) para textos em inglês de domínio irrestrito. Baseado na RST, ao desenvolver o *parser*, Marcu identificou vários outros problemas para a automatização da análise retórica além dos problemas "clássicos" de ambigüidade e falta de formalização da RST, a saber:

- como determinar as unidades mínimas de significado dos textos de forma "consistente" para que a análise retórica seja passível de automatização;
- como "descobrir" as relações retóricas intra e intersentenciais de forma automática;
- uma vez descobertas as relações retóricas, como saber que segmentos textuais são núcleos e satélites das relações;
- como montar as estruturas retóricas válidas de um texto a partir das relações retóricas entre suas partes.

Cada uma das questões acima e as soluções propostas por Marcu são discutidas a seguir.

3.1.1.1. Determinação das Unidades Mínimas de Significado

A determinação das unidades mínimas de significado consiste, na realidade, no conhecido problema de segmentação textual[7]. Em seus trabalhos, Marcu propôs duas soluções para este problema: uma baseada em análise de corpus e outra baseada em técnicas de Aprendizado de Máquina.

Em (Marcu, 1997, 2000a), por meio de análise de um corpus de domínio irrestrito, Marcu produziu várias regras para determinação dos segmentos de um texto. As regras se basearam na ocorrência de sinais de pontuação do texto e de marcadores discursivos (os marcadores e suas funcionalidades serão discutidos

[7] Para mais detalhes sobre técnicas de segmentação textual, além das empregadas por Marcu, vide Pardo e Nunes (2002).

posteriormente), já que estes são um dos principais indicativos superficiais da estruturação textual. Desta maneira, a cada padrão de itens léxicos encontrados no texto, uma ação foi associada. As ações são responsáveis por informar ao *parser* retórico onde inserir as marcações de início e fim de segmento. A Tabela 6 mostra alguns exemplos de padrões lexicais e ações associadas definidos por Marcu. Na primeira linha da tabela, por exemplo, caso a palavra *Although* seja encontrada no início de uma sentença, o *parser* deverá inserir uma marca de segmento após a próxima vírgula que encontrar na sentença.

Tabela 6 – Padrões lexicais e ações para segmentação de Marcu (1997, 2000a)

Padrões Lexicais	**Ações**
Although no começo de uma sentença	Inserir marca de fim de segmento imediatamente após a próxima ocorrência de vírgula
because no começo de uma sentença	Inserir marca de início de segmento imediatamente antes do marcador discursivo
for example no fim de uma sentença	Não inserir marca de segmento alguma

Com esta técnica, Marcu conseguiu precisão (*precision*) de 90% e cobertura (*recall*) de 81%. Neste contexto, precisão indica quantos segmentos corretos foram detectados em relação a tudo que foi detectado e cobertura indica quantos segmentos corretos foram detectados em relação a tudo que deveria ter sido detectado.

Na outra abordagem, utilizando técnicas de Aprendizado de Máquina, mais especificamente, o classificador C4.5 (Quinlan, 1993), Marcu (2000a) lista as *features* (características) abaixo como sendo as principais para determinar se um item lexical do texto indica ou não a presença de uma marca de segmento:

- a classe gramatical do item lexical sob análise;
- as classes gramaticais dos dois itens que precedem e seguem o item lexical sob análise;
- se o item lexical sob análise é um marcador discursivo;
- se o item lexical é uma abreviatura;
- se há verbos nas proximidades do item lexical sob análise.

Com esta técnica, Marcu conseguiu um desempenho (*f-measure*, que é uma combinação das medidas de precisão e cobertura) de 97%.

Após coletar e revisar todo o conhecimento sobre segmentação textual gerado pelas técnicas acima, Carlson e Marcu (2001) produziram um manual para segmentação de textos em inglês. Por este manual, é possível determinar exatamente e, mais importante, de forma consistente, como reconhecer segmentos textuais.

Recentemente, Soricut e Marcu (2003) também utilizaram um modelo probabilístico para determinar as cláusulas mais prováveis de uma sentença. O modelo probabilístico foi treinado com as palavras consideradas núcleos (*heads*) dos constituintes (sujeito, objetos, predicativos, etc.) das estruturas sintáticas lexicalizadas das sentenças. Assim, dada uma nova sentença, o modelo a segmenta nos pontos mais prováveis utilizando seus núcleos, determinando, assim, suas cláusulas. Por exemplo, grosso modo, dada uma sentença S formada pelas palavras w_1, w_2 ... w_n, a probabilidade da palavra w_i (com $1 \leq i \leq n$) indicar um novo segmento é dada pela combinação da probabilidade (a) da própria palavra w_i, (b) da palavra que é núcleo do constituinte sintático a que w_i pertence (sujeito, objeto, predicativo, etc.) e (c) da palavra que é núcleo do constituinte sintático que domina o constituinte

sintático a que w_i pertence (por exemplo, o núcleo de um predicado domina o núcleo de um complemento nominal dentro dele) indicarem um novo segmento. Com esta técnica, Marcu atingiu um desempenho de 84%.

3.1.1.2. Determinação das Relações Retóricas

Para determinar as relações retóricas entre os segmentos de um texto, Marcu (1997, 2000a) faz uso dos marcadores discursivos presentes no texto. Os marcadores discursivos são elementos coesivos formados de uma ou mais palavras que explicitam o relacionamento que existe entre as partes de um texto (Koch, 1998; Kock e Travaglia, 1990). Por isso, nas pesquisas recentes, eles são um dos principais fatores para a automatização da análise retórica.

Para identificar os marcadores discursivos em um texto e diferenciá-los de marcadores sentenciais e pragmáticos, Marcu utiliza padrões lexicais, também obtidos por meio de análise de corpus, semelhantes aos padrões mostrados na Tabela 6. Apesar de marcadores sentenciais e pragmáticos possuírem formação semelhante aos marcadores discursivos, eles são distinguidos pelo fato de não refletirem a estrutura discursiva do texto. Os marcadores sentenciais são utilizados em uma sentença para ligar suas partes somente, sem função discursiva. Por exemplo, o "e" que forma o sujeito composto da sentença "João e Maria são irmãos." é um marcador sentencial. Marcadores pragmáticos, por sua vez, remetem o leitor a seu conhecimento de mundo. Por exemplo, na sentença "João foi preso de novo.", o marcador "de novo" leva o leitor a inferir que João já foi preso antes.

Pela análise de corpus que realizou, Marcu associou a cada marcador discursivo (com seu contexto de ocorrência) as relações retóricas possíveis de ocorrerem. Por exemplo, o marcador *although*, dependendo da posição em que ocorre na sentença, pode indicar a relação retórica *concession* ou *contrast* entre os segmentos onde o marcador é observado. Com isso, durante a análise automática de um texto, todas as relações retóricas possíveis entre segmentos conectados por marcadores são listadas.

Nos casos em que não há marcadores discursivos entre segmentos textuais, Marcu tenta inferir a relação retórica pela aplicação de algumas heurísticas simples. Por exemplo: se um segmento repete algumas palavras do segmento anterior e não há marcadores discursivos entre eles, então se estabelece uma relação *background*; caso contrário, estabelece-se uma relação *elaboration*, que é a relação mais comum e genérica no elenco de relações da RST. Com esta técnica, Marcu conseguiu precisão de 78% e cobertura de 47%. Neste caso, a precisão indica quantas relações retóricas foram corretamente identificadas em relação a tudo que se identificou e cobertura indica quantas relações retóricas foram corretamente identificadas em relação a tudo que deveria ser identificado.

Deixando de lado a abordagem baseada em análise de corpus, Marcu e Echihabi (2002) utilizaram um classificador *naive-Bayes* (Mitchell, 1997) para determinar as relações retóricas entre dois segmentos textuais. Como *features* para o aprendizado, eles utilizaram as próprias palavras dos segmentos. Com isso, eles esperavam capturar dois tipos de conhecimento: (i) que marcadores discursivos indicavam quais relações retóricas e (ii) conhecimento de mundo. Em relação ao conhecimento de mundo, considere o exemplo abaixo dado pelos autores, no qual há uma relação retórica *contrast* entre as sentenças:

John is good in Math and Science. Paul fails almost every class he takes.

Por este exemplo, o classificador *bayesiano* aprenderia a relação de oposição que há entre as palavras *good* e *fail*. Ao se deparar com um novo exemplo que contivesse estas palavras, o classificador conseguiria, então, inferir a relação *contrast*.

Apesar de ser uma abordagem promissora, Marcu e Echihabi a utilizaram para determinar um pequeno conjunto de 4 relações bem distintas, atingindo um desempenho de 49%. Para o conjunto completo de relações, levando-se em consideração que algumas relações possuem uma diferença de definição muito tênue, imagina-se que essa abordagem não seria informada o suficiente para diferenciá-las com precisão suficiente.

Recentemente, Soricut e Marcu (2003) utilizaram um modelo probabilístico para determinar somente as relações retóricas intra-sentenciais. Neste modelo, dada uma sentença e sua estrutura sintática lexicalizada, Soricut e Marcu mostraram que a estrutura retórica da sentença pode ser determinada com grande precisão a partir de informações léxico-sintáticas. Eles treinaram o modelo probabilístico com os núcleos (*heads*) das cláusulas das sentenças (na maioria dos casos, os verbos) classificados com as relações retóricas que se estabeleciam entre eles, para, então, determinar a estrutura retórica das cláusulas de novas sentenças com base em seus núcleos. Os autores conseguiram um desempenho médio de 75% com esta técnica.

3.1.1.3. Determinação de Núcleos e Satélites

Uma vez que os segmentos textuais e as relações retóricas entre eles foram identificados, Marcu (1997, 2000a) utiliza a ordem preferencial de realização de núcleos e satélites das relações retóricas para determinar que segmentos são núcleos e que segmentos são satélites. Para determinar a ordem entre o núcleo e o satélite de cada relação, Marcu recorreu a sua análise de corpus e associou aos marcadores discursivos compilados as possíveis ordenações entre os segmentos. Por exemplo, para o marcador *Although* que ocorre no começo de uma sentença, a cláusula a qual o marcador pertence é classificada como satélite e a cláusula seguinte é classificada como núcleo, ou seja, segue-se a ordenação satélite-núcleo. Esta técnica atingiu precisão de 85% e cobertura de 50%. Neste caso, precisão indica quantos núcleos e satélites foram identificados corretamente em relação a tudo que foi identificado e cobertura indica quantos núcleos e satélites foram identificados corretamente em relação a tudo que deveria ter sido identificado.

3.1.1.4. Determinação das Estruturas Retóricas Válidas

Por fim, Marcu (1997) abordou o problema que considerou um dos mais desafiadores, o de construir as estruturas retóricas "válidas" de um texto a partir das relações retóricas que se estabelecem entre seus segmentos. Segundo Marcu, a falta de formalização da RST não permitia que se automatizasse este passo. Marcu procedeu, então, a uma completa formalização da RST. Desta formalização, os pontos principais e inovadores que permitiram a automatização da análise retórica são:

- critério da composicionalidade: dadas duas estruturas retóricas RSTree1 e RSTree2

RSTree1 = rhet_rel(R1,S1,S2)
RSTree2 = rhet_rel(R2,S3,S4)

onde o predicado rhet_rel(R,Y,X) representa uma estrutura retórica cujo segmento Y é o satélite do segmento nuclear X na relação retórica R, é possível combinar RSTree1 e RSTree2 em uma estrutura maior RSTree3 por meio da relação retórica R3, se R3 se estabelece entre os núcleos das estruturas a serem combinadas, ou seja, se R3 se estabelece entre os segmentos S2 da RSTree1 e S4 da RSTree2.
Esta definição, quando necessária, é aplicada recursivamente até que se chegue aos nós-folha das estruturas a serem combinadas. Como exemplo, considere o texto da Figura 10, o qual foi retirado integralmente do trabalho de Marcu (2000a), segmentado e numerado para referência.

[No matter how much on wants to stay a nonsmoker,]$_1$ *[the truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life.]*$_2$ *[We know that 3.000 teens start smoking each day,]*$_3$ *[although it is a fact that 90% of them once thought that smoking was something that they'd never do.]*$_4$

Figura 10 – Texto 2 (Marcu, 2000a)

Suponha que as seguintes relações retóricas, correspondentes às estruturas retóricas elementares que conectam diretamente dois segmentos, tenham sido reconhecidas:

rhet_rel(*justify*,1,2)
rhet_rel(*concession*,4,3)
rhet_rel(*evidence*,3,2)
rhet_rel(*restatement*,4,1)

Uma possível estrutura retórica para o texto é a mostrada na Figura 11: tem-se a estrutura conectando os segmentos 1 e 2 pela relação *justify*; tem-se a estrutura conectando os segmentos 4 e 3 pela relação *concession*; para montar uma estrutura maior que abranja as subestruturas anteriores, é necessário encontrar uma relação retórica que conecte os núcleos das duas, que, neste caso, é a relação *evidence* que conecta os segmentos 2 (núcleo da primeira subestrutura) e 3 (núcleo da segunda subestrutura). Como a relação *restatement* estabelece-se entre segmentos que não são os núcleos das subestruturas, ela não pode ser usada no lugar de *evidence* para montar a estrutura final.

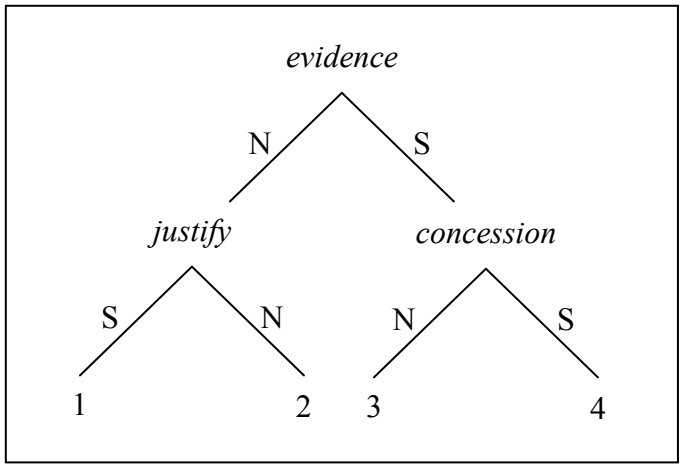

Figura 11 – Estrutura retórica para o Texto 2

- com base no critério da composicionalidade, dado o conjunto de relações retóricas que se estabelecem entre os segmentos textuais (como no exemplo anterior), é possível montar várias estruturas retóricas válidas para um mesmo texto. Para que fosse possível montar "todas" as estruturas retóricas válidas, e somente as válidas, Marcu (1997) desenvolveu vários algoritmos. O melhor deles, utilizando DCG (*Definite Clause Grammar*) (Pereira and Warren, 1980), produz todas as estruturas com tempo de execução linear. Neste algoritmo, dado o conjunto de relações retóricas entre os segmentos, é produzida uma gramática em DCG que produz todas as combinações possíveis entre as relações, montando, assim, todas as estruturas retóricas válidas.
  Para o exemplo anterior, pela aplicação do algoritmo de Marcu, a estrutura retórica da Figura 12 também poderia ser produzida e, como se pode verificar, pelo critério da composicionalidade, é uma estrutura válida. Portanto, para o conjunto de relações retóricas proposto para o Texto 2, as estruturas retóricas das Figuras 11 e 12 são "todas" as estruturas válidas possíveis de se produzir.

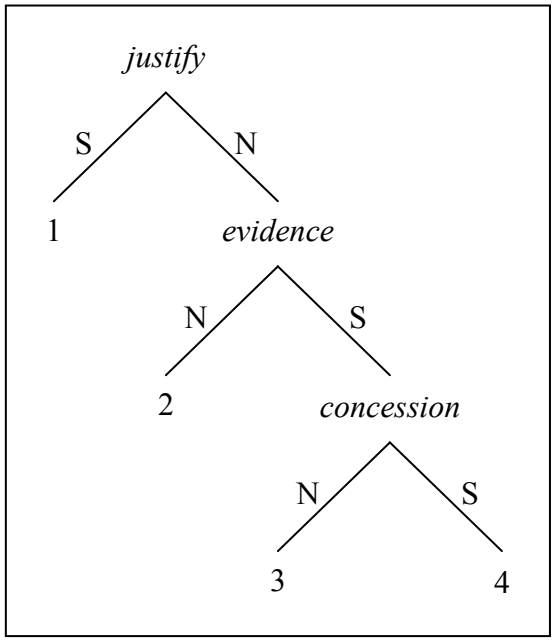

Figura 12 – Outra estrutura retórica para o Texto 2

3.1.2. Corston-Oliver

O analisador retórico produzido por Corston-Oliver (1998), chamado RASTA (*Rhetorical Structure Theory Analyzer*), foi desenvolvido a partir da análise de textos enciclopédicos. Ele se baseou na formalização proposta por Marcu (1997, 2000a).

Este trabalho merece destaque por abordar, além dos marcadores discursivos, as estruturas sintáticas (lexicalizadas) das sentenças de um texto para determinar as relações retóricas entre seus segmentos. Dentre as informações sintáticas disponíveis, Corston-Oliver levou em consideração, principalmente:

- se as cláusulas a serem relacionadas estavam subordinadas umas as outras ou não;
- se as cláusulas estavam na voz ativa ou passiva;
- se os núcleos dos constituintes sintáticos das cláusulas eram os mesmos ou não;
- as classes gramaticais das palavras das cláusulas.

Como exemplo, na Figura 13 são mostrados os critérios necessários listados por Corston-Oliver para que se estabeleça a relação retórica *cause* entre duas cláusulas A e B quaisquer de um texto (em inglês).

1. A cláusula A precede a cláusula B no texto
2. A cláusula A não é sintaticamente subordinada à cláusula B
3. A cláusula B não é sintaticamente subordinada à cláusula A
4. O sujeito da cláusula B é um pronome demonstrativo ou é modificado por um demonstrativo; ou as cláusulas A e B são coordenadas por um símbolo de dois pontos (;)
5. A cláusula B está na voz passiva e possui a palavra *cause*; ou a cláusula B contém a frase indicativa *result from* com o verbo estando possivelmente flexionado

Figura 13 – Critérios de Corston-Oliver para a relação retórica *cause*

É importante dizer que Corston-Oliver também trabalhou com o conceito de subespecificação, isto é, quando, por falta de informação disponível, não se consegue determinar a relação retórica que se estabelece entre dois segmentos textuais. Nestes casos, ele sinaliza na estrutura retórica que a relação existe, mas não a especifica.

3.1.3. Schilder

O trabalho de Schilder (2002) destaca-se pelo uso que faz de técnicas de Recuperação de Informação (RI) para auxiliar a análise retórica. Em uma primeira etapa de seu sistema de análise retórica, dado um texto, sua estrutura retórica parcial é produzida por meio da identificação dos marcadores discursivos presentes e das relações que estes indicam. A estrutura parcial é então completada pelo uso de técnicas de RI.

Uma estrutura parcial não contém todos os segmentos identificados em um texto. Isso ocorre pela falta de informação disponível (por exemplo, ausência de marcadores discursivos e ambigüidade retórica) para determinar onde os segmentos seriam anexados na estrutura retórica. Para decidir onde anexar os segmentos restantes na estrutura, Schilder utiliza informação sobre a topicalidade dos segmentos. Segmentos topicais são mais importantes para um texto e, portanto, devem ser anexados em posições mais importantes da estrutura retórica.

Para determinar a topicalidade dos segmentos restantes de um texto, Schilder os representa em vetores, seguindo a proposta de Salton (1971), e os compara com o vetor do título do texto sob análise. Os segmentos cujos vetores são mais próximos do vetor do título do texto recebem uma maior pontuação. Por fim, Schilder considera que os segmentos com maior pontuação são segmentos topicais, restringindo, assim, os locais aos quais estes segmentos serão anexados à estrutura retórica parcial.

É importante ressaltar que, como Corston-Oliver (1998), Schilder também permite subespecificação na estrutura retórica que produz para um texto. Desta forma, apesar de conseguir determinar a correta localização dos segmentos textuais na estrutura retórica, as relações podem não ser especificadas.

3.2. Marcadores Discursivos

Os marcadores discursivos[8] são essenciais para o bom desempenho da análise retórica automática, pois são os maiores indicadores superficiais das relações retóricas no texto. Por exemplo, ao se encontrar o marcador "entretanto", "contudo" ou "porém" conectando dois segmentos textuais, há grandes chances de haver uma relação retórica de oposição (*contrast*, *antithesis* ou *concession*) entre eles.

Na Lingüística e na Lingüística Computacional, há vários trabalhos sobre marcadores discursivos e sua função no discurso. Diz-se que eles determinam a estrutura do discurso e, ao mesmo tempo, são determinados por ela. Eles são pistas que o escritor do texto deixa para que o leitor consiga, com o mínimo esforço possível, entender o relacionamento entre as partes do texto e entender, portanto, o próprio sentido do texto (Koch e Travaglia, 1990; Koch, 1998).

Para o inglês, muitos trabalhos se destacaram no estudo dos mais diversos marcadores discursivos (por exemplo, Di Eugenio, 1992, 1993; Elhadad e McKeown, 1990; Hirschberg e Litman, 1987, 1993; Knott, 1995; Knott e Dale, 1996; Knott e Mellish, 1996; Grote et. al., 1997; Fraser, 1999; Oates, 1999; etc.). Para o português do Brasil, destacam-se os trabalhos de Koch (1998), Paizan (2001) e Dias da Silva e Oliveira (2002), os quais apresentam a função de vários marcadores discursivos, seus contextos de ocorrência e que relações retóricas eles indicam.

Além de indicar a relação retórica existente entre dois segmentos textuais, Grosz e Sidner (1986) afirmam que os marcadores discursivos também podem ser usados para indicar as relações intencionais entre dois segmentos. No exemplo dado pelas autoras, marcadores como "*firstly*", "*in the first place*", "*second*", "*then*", "*lastly*", entre outros, podem indicar uma relação intencional de *satisfaction-precedence* entre as intenções subjacentes aos segmentos que possuem tais marcadores.

Em uma argumentação diferente da tradicional, Korelsky e Kittredge (1993) afirmam que os marcadores discursivos indicam, na realidade, as relações semânticas entre os segmentos, e não as relações retóricas. Entretanto, usualmente, assume-se que os marcadores indicam, em primeira instância, a estruturação retórica de um texto.

---

[8] Segundo Hirschberg e Litman (1993) e Fraser (1999), os marcadores discursivos também podem ser chamados de conectivos discursivos, operadores discursivos, partículas discursivas, sinalizadores de discurso, conectivos fáticos, conectivos pragmáticos, expressões pragmáticas, formativos pragmáticos, marcadores pragmáticos, operadores pragmáticos, partículas pragmáticas, conjuntos semânticos, conectivos de sentenças, frases indicativas, palavras indicativas, etc.

A próxima seção apresenta alguns trabalhos que fazem uso de análise discursiva em aplicações de PLN, mais especificamente, pesquisas em Sumarização Automática de Textos (SA), pois esta área é uma das que mais tem se beneficiado das pesquisas em Análise de Discurso.

4. Análise Discursiva e Sumarização Automática de Textos

Um sistema capaz de produzir a estrutura discursiva de um texto pode ser de grande utilidade em PLN. De fato, várias pesquisas atuais têm apontado como esse conhecimento pode ser usado para melhorar os resultados do estado da arte em várias aplicações em PLN, desde Resolução de Anáforas a SA. Esta seção relatará os pontos principais de algumas pesquisas em SA que fizeram uso de análise discursiva, principalmente da análise retórica.

A retórica na SA tem recebido muito destaque devido a uma de suas características básicas, a de que uma relação retórica padrão possui um núcleo e um satélite e que o satélite das relações pode ser omitido sem grandes prejuízos para a satisfação do objetivo comunicativo de um texto (Mann and Thompson, 1987).

Nessa linha, Marcu (1997, 2000a), a partir da estrutura retórica extraída automaticamente de um texto-fonte (vide subseção 3.1.1), aplica um algoritmo para computar o grau de importância de cada segmento da estrutura retórica, considerando que os segmentos mais importantes (nucleares ou, como denominado por ele, "salientes") apresentam informações mais relevantes do texto, sendo adequados, portanto, para compor o sumário desse texto.

Ocálculo da importância dos segmentos é baseado na saliência e na profundidade em que cada segmento aparece na estrutura retórica. Quanto mais saliente e mais acima estiver na estrutura, maior a importância de um segmento.

A Figura 15 mostra a estrutura retórica[9] do texto *Using Computers in Manufacturing* (Jordan, 1980, p.225) (mostrado na Figura 14, com segmentos identificados e numerados para referência). Ao lado de cada relação dessa estrutura, é indicado o segmento mais saliente da sub-árvore correspondente. Desta forma, a ordem de importância dos segmentos para compor o sumário é dada por 2>1>3a>3b>4>6>5, onde 2 é mais importante que 1, que é mais importante que 3a, etc.

[9] É importante notar que os trabalhos relatados nesta seção utilizaram variações do conjunto de relações retóricas da RST.

1. Whether you regard computers as a blessing or a curse, the fact is that we are all becoming more and more affected by them.
2. The general level of understanding of the power and weaknesses of computers among manufacturing managers is dangerously low.
3a. In order to counteract the lack of knowledge, the Manufacturing Management Activity Group of the IprodE is organizing a two-day seminar on "Computers and manufacturing management"
3b. to be held at the Birmingham Metropole Hotel at the National Exhibition Centre from 21-22 March 1979.
4. The seminar has been specially designed by the IprodE for managers concerned with manufacturing processes and not for computer experts.
5. The idea is that delegates will be able to share the experiences of other computer users and learn of their successes and failures.
6. The seminar will consist of plenary sessions followed by syndicates where delegates will be arranged into small discussion groups.

Figura 14 – Texto *Using Computers on Manufacturing* (Jordan, 1980)

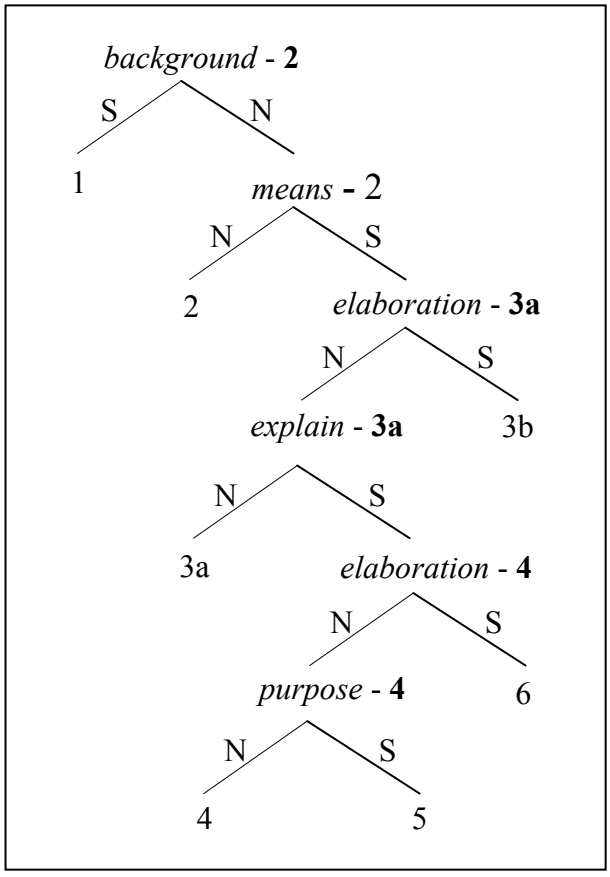

Figura 15 – Classificação de Marcu

Analisando-se passo a passo como se chegou a essa ordem de importância entre os segmentos, percorrendo a estrutura retórica de baixo para cima, tem-se:
1) entre o núcleo 4 e o satélite 5 da relação *purpose*, o núcleo 4 sobe para a relação *purpose*;

2) entre o núcleo 4 e o satélite 6 da relação *elaboration*, o núcleo 4 sobe para a relação *elaborate*;
3) entre o núcleo 3a e o satélite 4 da relação *explain*, o núcleo 3a sobe para a relação *explain*;
4) entre o núcleo 3a e o satélite 3b da relação *elaboration*, o núcleo 3a sobe para a relação *elaboration*;
5) entre o núcleo 2 e o satélite 3a da relação *means*, o núcleo 2 sobe para a relação *means*;
6) entre o núcleo 2 e o satélite 1 da relação *background*, o núcleo 2 sobe para a relação *background*.

Quantos desses segmentos serão utilizados para compor o sumário dependerá do tamanho desejado deste.

Em experimentos realizados, Marcu produziu resultados automáticos com 57% de cobertura (*recall*) e 51% de precisão (*precision*), produzindo uma *f-measure* de 53,8%. Neste caso, cobertura indica quantos segmentos do texto-fonte foram selecionados para compor o sumário em relação a todos os segmentos que deveriam ser selecionados e precisão indica quantos segmentos foram selecionados para compor o sumário em relação a todos os segmentos que foram selecionados.

O'Donnel (1997), em seu trabalho, também atribui valores de importância aos segmentos da estrutura retórica. Entretanto, O'Donnel não possui um analisador retórico automático, fornecendo manualmente a estrutura retórica a seu sumarizador.

O valor de cada segmento é calculado com base no valor do segmento mais nuclear mais próximo ao segmento em foco multiplicado pelo valor (peso) da relação retórica pela qual o segmento mais nuclear é imediatamente dominado. Os pesos das relações são pré-definidos de acordo com sua "taxa de relevância" em um processo de comunicação. Por exemplo, O'Donnel atribuiu o peso 0.40 para a relação *elaboration*, enquanto atribuiu o peso 0.70 para a relação *purpose*, já que esta deve apresentar informação mais importante para a satisfação do objetivo de um texto do que a relação *elaboration*.

O segmento mais nuclear na estrutura retórica recebe automaticamente o valor máximo (1.0); os demais valores são calculados a partir deste de forma recursiva, conforme se atinge os níveis mais profundos da estrutura. Desta forma, os segmentos mais acima na estrutura receberão valores maiores do que os mais abaixo. O'Donnel afirma que essa técnica pode produzir resultados ruins em alguns casos, já que, de acordo com seus experimentos, nem sempre a nuclearidade reflete a centralidade da informação, pois o autor do texto pode apresentar informações importantes em lugares do texto retoricamente irrelevantes.

A Figura 16 mostra a estrutura retórica do texto *Using Computers in Manufacturing* e os valores de cada segmento. Pesos fictícios foram associados às relações retóricas (*purpose*=0,70; *explain* e *means*=0,60; *background*=0,50; *elaboration*=0,40), resultando na seguinte ordem de importância: 2>3a>1>4>5>3b>6. Analisando-se passo a passo, de cima para baixo na estrutura retórica, tem-se:
1) o segmento 2, por ser o mais nuclear, recebe o valor máximo 1;
2) o valor do segmento 1 é igual ao valor do segmento 2 multiplicado pelo valor da relação *background*, ou seja, 1 x 0,50 = 0,50;
3) o valor do segmento 3a é igual ao valor do segmento 2 multiplicado pelo valor da relação *means*, ou seja, 1 x 0,60 = 0,60;

4) o valor do segmento 3b é igual ao valor do segmento 3a multiplicado pelo valor da relação *elaborate*, ou seja, 0,60 x 0,40 = 0,24;
5) o valor do segmento 4 é igual ao valor do segmento 3a multiplicado pelo valor da relação *explain*, ou seja, 0,60 x 0,60 = 0,36;
6) o valor do segmento 6 é igual ao valor do segmento 4 multiplicado pelo valor da relação *elaborate*, ou seja, 0,36 x 0,40 = 0,14;
7) o valor do segmento 5 é igual ao valor do segmento 4 multiplicado pelo valor da relação *purpose*, ou seja, 0,36 x 0,70 = 0,25.

Como no trabalho de Marcu, quantos desses segmentos serão utilizados para compor o sumário dependerá do tamanho desejado deste.

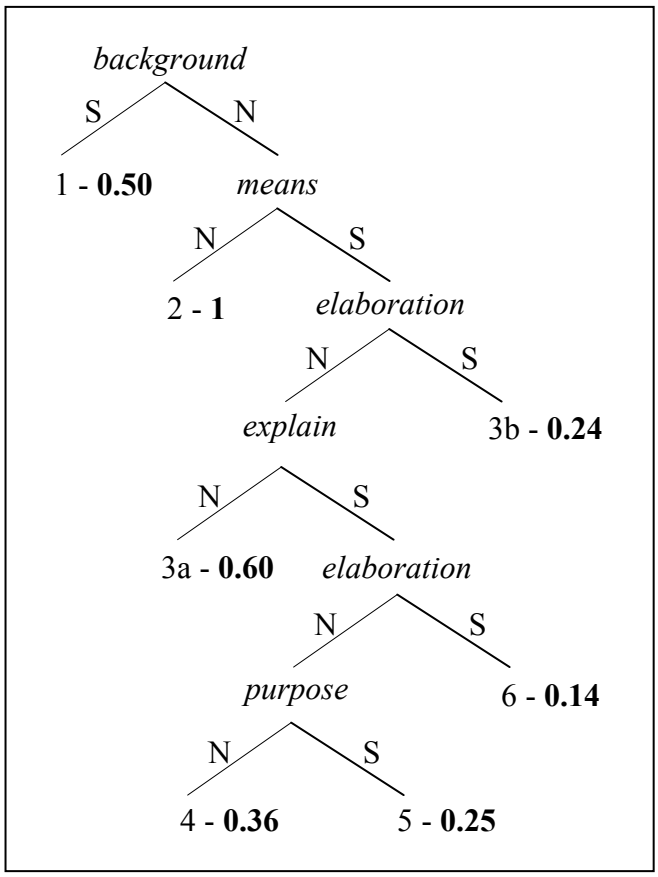

Figura 16 – Classificação de O'Donnel

Tanto Marcu como O'Donnel consideraram, de uma forma ou de outra, que a informação mais importante a compor um sumário é a informação mais nuclear em uma estrutura retórica.

Em uma outra linha, Pardo (2002), baseando-se no modelo de discurso de Rino (1996a) (já mostrado na subseção 2.4.1), implementou um gerador automático de textos chamado DMSumm (*Discourse Modeling SUMMarizer*) que, a partir da estrutura semântica e das relações intencionais de um texto-fonte, produz o(s) sumário(s) correspondente(s). A estrutura semântica e as relações intencionais, entretanto, devem ser especificadas manualmente.

Para fazer o mapeamento das relações semânticas e intencionais em relações retóricas, Pardo utilizou operadores de plano (Moore and Paris, 1993). Desta forma,

dependendo da ordem de aplicação dos operadores de plano, do objetivo comunicativo considerado e da proposição central do discurso (que deverá ser o segmento mais nuclear da estrutura retórica a ser produzida), vários sumários podem ser construídos. Assim, diferentemente de Marcu e O'Donnel, as estruturas retóricas produzidas pelo DMSumm já são consideradas as próprias estruturas retóricas dos sumários, não se excluindo satélite algum dessas estruturas. O DMSumm produziu sumários com 44% de precisão e 54% de cobertura.

Mais detalhes sobre análise discursiva aplicada à SA podem ser conseguidos em Pardo e Rino (2002), Ribeiro e Rino (2002) e Martins et al. (2001).

5. Conclusões

Este relatório apresentou uma revisão bibliográfica sobre as principais teorias discursivas existentes, alguns trabalhos que automatizaram a análise discursiva retórica e pesquisas em SA que fizeram uso de análise discursiva. De fato, a análise discursiva pode auxiliar bastante as pesquisas em PLN, como Resolução Anafórica, Tradução Automática, SA, etc. Entretanto, ainda não há para o português do Brasil ferramentas computacionais capazes de extrair de um texto seu conhecimento discursivo.

A revisão apresentada faz parte da investigação para o desenvolvimento de um analisador discursivo para o português do Brasil, chamado DiZer-PBr[10] (*DIscourse analyZER for BRazilian Portuguese*), que abrangerá os níveis retórico, semântico e intencional na análise discursiva. Ao estar funcional, este analisador poderá ser utilizado em muitas pesquisas sendo desenvolvidas no NILC.

Referências

Carlson, L. and Marcu, D. (2001). *Discourse Tagging Reference Manual*. ISI Technical Report ISI-TR-545.

Corston-Oliver, S. (1998). *Computing Representations of the Structure of Written Discourse*. PhD Thesis, University of California, Santa Barbara, CA, USA.

Dale, R. (1993). Rhetoric and Intentions in Discourse. In the *Proceedings of the Intentionality and Structure in Discourse Relations Workshop*, pp. 5-6. Ohio, USA.

Di Eugenio, B. (1992). Understanding natural language instructions: the case of purpose clauses. In the *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL'92)*, pp. 120-127. Newmark, DE.

Di Eugenio, B. (1993). *Understanding natural language instructions: a computational approach to purpose clauses*. Ph.D. thesis, University of Pennsylvania.

Dias da Silva, B.C.; Oliveira, M.F.; Moraes, H.R.; Hasegawa, R.; Amorim, D.; Paschoalino, C.; Nascimento, A.C. (2000). A Construção de um Thesaurus Eletrônico para o Português do Brasil. Nos *Anais do V Encontro para o*

[10] Mais detalhes sobre este projeto podem ser obtidos em www.nilc.icmc.usp.br/~thiago/DiZer.html

Processamento Computacional da Língua Portuguesa Escrita e Falada – PROPOR'2000. Atibaia – SP. Brasil.

Dias da Silva, B.C. e Oliveira, M.F. (2002). Inclusão de informação pragmático-discursiva na base lexical de um thesaurus eletrônico. *Estudos Lingüísticos*, Vol. 31.

Dik, S.C. (1997). *The Theory of Functional Grammar*. K. Hengeveld (ed.), Berlin: Mouton de Gruyter.

Elhadad, M, and McKewon, K. R. (1990). Generating connectives. In the *Proceedings of the International Conference on Computational Linguistics (COLING'90)*, Vol. 3, pp. 97-102. Helsinki, Finland.

Feltrim, V.D.; Aluísio, S.M.; Nunes, M.G.V. (2000). *Uma Revisão Bibliográfica sobre a Estruturação de Textos Científicos em Português*. Série de Relatórios do NILC. NILC-TR-00-11.

Feltrim, V.D.; Nunes, M.G.V.; Aluísio, S.M. (2001). *Um corpus de textos científicos em Português para a análise da Estrutura Esquemática*. Série de Relatórios do NILC. NILC-TR-01-4.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, Vol. 32, pp. 913-952.

Grosz, B. and Sidner, C. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, Vol. 12, N. 3.

Grote, B.; Lenke, N.; Stede, M. (1997). Ma(r)king concessions in English and German. *Discourse Processes*, pp. 87-117.

Halliday, M.A.K. (1985). *An Introduction to Functional Grammar*. Edward Arnold Press, London.

Hirschberg, J. and Litman, D. J. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, Vol. 19, N. 3, pp. 501-513.

Hoey, M.P. (1979). *Signalling in Discourse*. University of Birmingham, England.

Hoey, M.P. (1983a). *On the Surface of Discourse*. London: George Allen and Unwin.

Hoey, M.P. (1983b). The place of clause relational analysis in linguistic description. *English Language Research Journal*, N. 4.

Hoey, M.P. and Winter, E.O. (1986). Clause relations and the writer's communicative task. In B. Couture (ed.), *Functional Approaches to Writing*. London: Frances Pinter.

Hovy, E. (1991). Approaches to the planning of coherent text. In C. Paris, W. Swartout and W. Mann (eds.), *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pp. 83-102. Kluwer Academic Publishers, Boston.

Hovy, E. (1993). In Defense of Syntax: Informational, Intentional, and Rhetorical Structures in Discourse. In the *Proceedings of the Intentionality and Structure in Discourse Relations Workshop*, pp. 35-39. Ohio, USA.

Jordan, M.P. (1978). *The principal semantics of the nominals 'this' and 'that' in contemporary English writing*. PhD Thesis, The Hatfield Polytechnic and Birmingham University, England.

Jordan, M.P. (1980). Short Texts to Explain Problem-Solution Structures – and Vice Versa. *Instructional Science*, Vol. 9, pp. 221-252

Jordan, M.P. (1985a). Some clause relational associated nominals in technical English. *Technostyle*, Vol. 4, N. 1.

Jordan, M.P. (1985b). Some relations of surprise and expectation in English. In J. Hall (ed.), *The 11th LACUS Forum*. Columbia SC: Hornbeam Press.

Jordan, M.P. (1988). Some advances in clause relational theory. In J.D. Benson and W.S. Greaves (eds.), *Systemic Functional Approaches to Discourse*. Norwood NJ: Ablex.

Jordan, M.P. (1989). Relational propositions within the clause. In S. Embleton (ed.), *The 15th LACUS Forum*. Lake Bluffs IL: LACUS.

Jordan, M.P. (1992). An Integrated Three-Pronged Analysis of a Fund-Raising Letter. In W.C. Mann and S.A. Thompson (eds), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pp. 171-226.

Koch, I.V. (1998). *A Coesão Textual*. Editora Contexto.

Koch, I.V. e Travaglia, L.C. (1990). *A Coerência Textual*. Editora Contexto.

Knott, A. (1995). *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, University of Edinburgh, Scotland.

Knott, A. and Dale, R. (1996). Choosing a set of coherence relations for text generation: a data-driven approach. In M. Zock (ed), *Trends in Natural Language Generation: an Artificial Intelligence Perspective*, pp. 47-67. Heidelberg, Germany.

Knott, A. and Mellish, C. (1996). A feature-based account of the relations signaled by sentence and clause connectives. *Journal of Language and Speech*, Vol. 39, Ns. 2 and 3, pp. 143-183.

Korelsky, T. and Kittredge, R. (1993). Towards stratification of RST. In the *Proceedings of the Intentionality and Structure in Discourse Relations Workshop*, pp. 52-55. Ohio, USA.

Kuhn, D.; Abarca, E.; Nunes, M.G.V. (2000). *Corpus NILC - Situação em Maio/2000*. NILC-TR-00-7.

Kurohashi, S. and Nagao, M. (1994). Automatic detection of discourse structure by checking surface information in sentences. In the *Proceedings of the 15th International Conference on Computational Linguistics* (COLING'94), Vol. 2, pp. 1123-1127. Kyoto, Japan.

Maier, E. (1993). The Representation of Interdependencies between Communicative Goals and Rhetorical Relations in the Framework of Multimedia Document Generation. In the *Proceedings of the Intentionality and Structure in Discourse Relations Workshop*, pp. 70-73. Ohio, USA.

Maier, E. and Hovy, E. H. (1991). A Metafunctionally Motivated Taxonomy for Discourse Structure Relations. In the *Proceedings of the 3rd European Workshop on Language Generation*. Innsbruck, Austria.

Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD Thesis, Department of Computer Science, University of Toronto.

Marcu, D. (1999). A formal and computational synthesis of Grosz and Sidner's and Mann and Thompson's theories. In *Proceedings of the Workshop on Levels of Representation in Discourse*, pp. 101-108. Edinburgh, Scotland.

Marcu, D. (2000a). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. Cambridge, Massachusetts.

Marcu, D. (2000b). Extending a Formal and Computational Model of Rhetorical Structure Theory with Intentional Structures à la Grosz and Sidner. In the

Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000), Saarbrueken.

Marcu, D. and Echihabi, A. (2002). An Unsupervised Approach to Recognizing Discourse Relations. In the *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL'02), Philadelphia, PA.

Martins, C.B.; Pardo, T.A.S.; Espina, A.P.; Rino, L.H.M. (2001). *Introdução à Sumarização Automática*. Relatório Técnico RT-DC 002/2001, Departamento de Computação, Universidade Federal de São Carlos.

Maybury, M.T. (1992). Communicative Acts for Explanation Generation. *Int. Journal of Man-Machine Studies 37*, pp. 135-172.

Mitchell, T.M. (1997). *Machine Learning*. McGraw Hill, New York.

Moore, J.D. (1995). *Participating in Explanatory Dialogs: Interpreting and Responding to Questions in Context*. The MIT Press. Cambridge, Massachusetts.

Moore, J.D. and Paris, C. (1993). Plannig Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information. *Computational Linguistics*, Vol. 19, N. 4, pp. 651-694.

Moore, J. D. and Pollack, M. E. (1992). A problem for RST: the need for multi-level discourse analysis. *Computational Linguistics*, Vol. 18, N. 4, pp. 537-544.

Moser, M. and Moore, J. D. (1996). Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, Vol. 22, N. 3, pp. 409-419.

O'Donnell, M. (1997). Variable-Length On-Line Document Generation. In the *Proceedings of the 6th European Workshop on Natural Language Generation*, Gerhard-Mercator University, Duisburg, Germany.

Oates, S.L. (1999). *State of the Art Reporto n Discourse Markers and Relations*. Technical Report ITRI-99-08. Information Technology Research Institute. University of Brighton.

Paizan, D.C. (2001). *O uso da linguagem da Internet na produção de um módulo de ensino de leitura de inglês instrumental*. Dissertação de Mestrado. Faculdade de Ciências e Letras de Araraquara.

Pardo, T.A.S. (2002). *DMSumm: Um Gerador Automático de Sumários*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos – SP.

Pardo, T.A.S. e Nunes, M.G.V. (2002). *Segmentação Textual Automática: Uma Revisão Bibliográfica*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação - ICMC, Universidade de São Paulo, N. 185.

Pardo, T.A.S. and Rino, L.H.M. (2002). DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), *Advances in Natural Language Processing*, pp. 263-273 (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany.

Pereira, F.C.N. and Warren, D.H.D. (1980). Definite Clause Grammars for Language Analysis – A Survey of the Formalism and a Comparison with Augmented Transition Networks. *Artificial Intelligence*, Vol. 13, pp. 231-278.

Pinheiro, G.M. e Aluísio, S.M. (2003). *Corpus NILC: Descrição e Análise Crítica com Vistas ao Projeto Lacio-Web*. Série de Relatórios Técnicos do Instituto de Ciencias Matemáticas e de Computação – ICMC, Universidade de São Paulo, N. 190.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Ribeiro, G.F. and Rino, L.H.M. (2002). *A Sumarização Automática com base nas Estruturas RST*. Séries de Relatórios do NILC (DC-UFSCar). NILC-TR-02-05.

Rino, L.H.M. (1996a). *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. Tese de Doutorado. IFSC-Usp. São Carlos - SP.

Rino, L.H.M. (1996b). A sumarização automática de textos em português. In *Anais do II Encontro para o Processamento Computacional de Português Escrito e Falado*, pp. 109-119. Curitiba - PR.

Salton, G. (1971). *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall, New York.

Schilder, F. (2002). Robust discourse parsing via discourse markers, topicality and position. In J. Tait, B.K. Boguraev and C. Jacquemin (eds.), *Natural Language Engineering*, Vol. 8. Cambridge University Press.

Soricut, R. and Marcu, D. (2003). Sentence Level Discourse Parsing using Syntactic and Lexical Information. *To appear*.

Sumita, K.; Ono, K.; Chino, T.; Ukita, T.; Amano, S. (1992). A discourse structure analyzer for Japonese text. In the *Proceedings of the International Conference on Fifth Generation Computer Systems*, Vol. 2, pp. 1133-1140. Tokyo, Japan.

Winter, E.O. (1968). Some aspects of cohesion. In R.D. Huddleston, R.A. Hudson., E.O. Winter and A. Henrici (eds.), *Sentence and Clause in Scientific English*. University of London.

Winter, E.O. (1971). *Connection in science material: A proposition about the semantics of clause relations*. Centre for Information on English Language Teaching and Research.

Winter, E.O. (1974). *Replacement as a function of repetition*. PhD Thesis, University of London.

Winter, E.O. (1976). *Fundamentals of Information Structure*. Hatfield Polytechnic, Hertfodshire, England.

Winter, E.O. (1977). A Clause-Relational Approach to English Texts. A Study of Some Predictive Lexical Items in Written Discourse. *Structional Science*, Vol. 6, N. 1, pp. 1-92.

Winter, E.O. (1979). Replacement as a Fundamental Function of the Sentence in Context. In *Forum Linguitics*, Vol. 4, N. 2, pp. 95-133.

Winter, E.O. (1982). *Towards a Contextual Grammar of English*. London: George Allen and Unwin.