AuTema-Dis: uma arquitetura computacional para identificação da temática discursiva em textos em Língua Portuguesa

Ana Luísa Leal[1], Paulo Quaresma[2], Luís Rodrigues[2]

[1]Departamento de Português, Faculdade de Ciências e Humanidades, Universidade de Macau, Macau

[2]Departamento de Informática, Universidade de Évora, Évora, Portugal

analeal@umac.mo, pq@di.uevora.pt, lfrodrigues@gmail.com

Abstract. *This article describes the proposed methodology and the computational implementation of the Autema-Dis system. Its a computational architecture aiming to automatically analyze text and to present, after the processing phases, a representative structure of the textual macroproposition. Each processing phase is sequential and its performed without human intervention; textual structure classification, dependency trees organization, rhetorical relations identification, and the inference of the textual macroproposition.*

Resumo. *Este artigo tem como objetivo apresentar a metodologia e a implementação computacional do Autema-Dis. Trata-se de uma arquitectura computacional elaborada para realizar a análise automática de texto e apresentar, ao final do processamento, uma estrutura representativa da macroproposição textual. As etapas realizadas pelo sistema são sequenciais e sem a interferência humana; classificam as estruturas textuais, organizam árvores de dependência, atribuem automaticamente de algumas relações retóricas entre segmentos e elaboram uma macroproposição representativa do conteúdo do texto.*

1. Introdução

Este trabalho faz parte de um projecto de investigação desenvolvido no âmbito do doutoramento em Informática na Universidade de Évora, Portugal. Em revisão à literatura da área, observamos que alguns trabalhos desenvolvidos em Linguística Computacional direcionados à análise textual procuram explicar a realização e as relações de alguns elementos linguísticos e ocorrências linguísticas, devidamente marcados na estrutura textual. Tais estudos revelam-se, relacionados às questões de ordem morfológica, sintática e, conforme observamos poucas investigações avançam em direção à semântica. O texto é o concreto objeto de estudo, bem como os processos envolvidos e relacionados com a sua constituição e organização estrutural.

No sentido de avançar em direcção a um campo de investigação ainda a ser explorado na linguística computacional (LC), o qual lida com a total automatização e

Anais do III Workshop "A RST e os Estudos do Texto", páginas 43–58, Cuiabá, MT, Brasil, Outubro 24–26, 2011.

análise textual, sem a interferência ou manipulação humana das informações, elaboramos uma proposta de arquitetura para análise automática de texto/discurso[1], identificado neste estudo, conforme Pardo (2005). A pesquisa em questão direcionou-se, especificamente, à elaboração de uma arquitectura computacional que, a partir da sua implementação em sistema, efetuasse automaticamente análise de um texto em sua totalidade e gerasse uma estrutura sintética representativa da macroestrutura temática do discurso. Assim sendo, a partir da modelagem proposta construiu-se o protótipo AuTema-Dis, um analisador temático discursivo que articula informações de cunho formais do tipo morfossintáticas e informações relacionais, isto é, relações semânticas.

O texto é o objeto e a base de análise desta investigação, assim, para a composição da metodologia e realização automática da análise foi necessário considerar dois níveis: o nível das relações referente às questões sintáticas, intra e intersentencial e o nível conceitual relacionado às questões conceituais e significativas, suportado pelas relações retóricas[2], que se articulam e organizam a estrutura discursiva – o *discurso*.

2. A proposta

O projeto AuTema-Dis prevê uma intersecção de conhecimentos de áreas distintas, ou seja, Linguística e Informática. O objectivo do trabalho foi desenvolver uma base metodológica, cuja sistematização fosse capaz de:

- reconhecer a informação principal em um determinado discurso, considerando o resultado de uma análise sintática automática – *Palavras*[3] ;
- reorganizar as estruturas relevantes ao tema em árvores de dependência dos segmentos – DTS[4];
- atribuir automaticamente algumas relações retóricas entre os segmentos e subsegmentos organizados nas DTS's;
- produzir automaticamente uma estrutura sintética em língua natural, representativa da informação disposta na estrutura textual/discursiva, considerando os resultados das etapas anteriores.

2.1. Arquitetura AuTema-Dis

A arquitetura AuTema-Dis encontra-se representada através da figura 1.

1 Texto/Discurso: Conforme Pardo (2005) um texto possui uma estrutura subjacente altamente elaborada que relaciona todo o seu conteúdo, atribuindo-lhe coerência. A essa estrutura dá-se o nome de estrutura discursiva.

2 Relaçãos Retóricas: conforme a proposta de Mann and Thompson (1988).

3 Palavras: analisador sintático – *parser*, desenvolvido por E. Bick (2000), no âmbito do projeto VISL6, no Institute of Language and Communication da University of Southern Denmark.

4 DTS: conceito desenvolvido no âmbito deste estudo para designar árvore de dependência de segmentos.

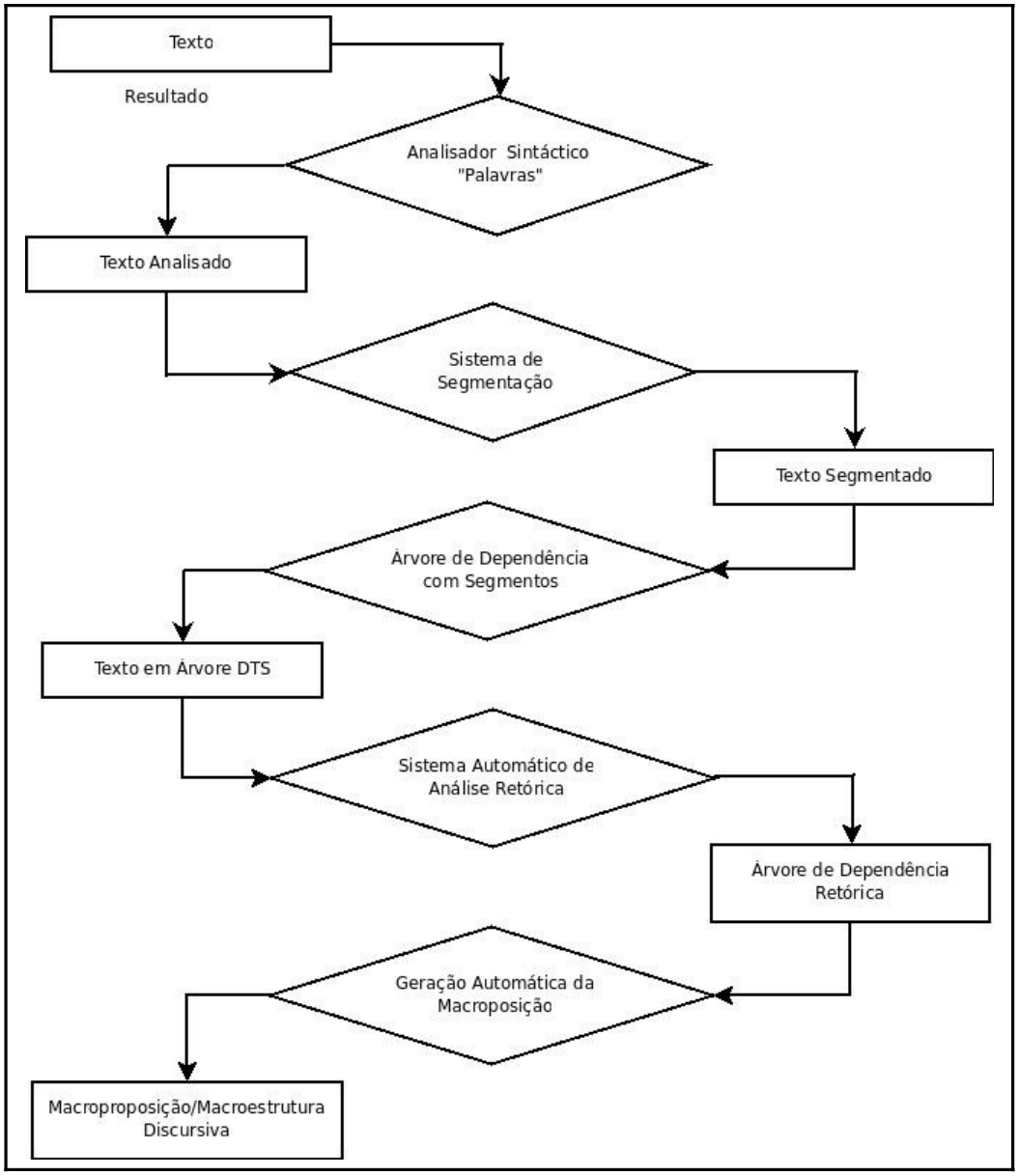

Figura 1: Arquitetura do sistema AuTema-Dis.

2.2. Base Metodológica

A proposta metodológica que apresentamos foi edificada com o objectivo principal de realizar análise textual, direcionando os seus resultados à elaboração automática da macroestrutura/macroproposição[5] discursiva. Para tal, desenvolvemos uma arquitetura do tipo modular, conforme figura 1, em que a execução dos diferentes módulos é capaz de realizar a análise textual, considerando-se os diferentes níveis linguísticos que constituem na íntegra a estrutura do texto.

5 Macroproposição: a definição do termo é empregada aqui conforme Dijk (1972). Macroproposições representam significado de parágrafos, seções de texto, grupo/conjunto de seções e eventualmente o papel do próprio texto. Elas são contrapartidas formais das noções intuitivas dos principais pontos e do resumo de um texto, representadas por macroestruturas.

No que diz respeito à constituição da metodologia, foi prevista a realização de quatro etapas de análise distintas e autônomas, mas relacionadas entre si. Cada módulo da arquitetura consiste em uma das etapas de análise e o resultado da execução de cada uma apresenta informações, para a execução da etapa seguinte até a conclusão de todo o processo.

A edificação dos quatro módulos básicos foi realizada a partir de observações empíricas a respeito da constituição do texto jornalístico. Elaborou-se um estudo piloto, em que foi feita uma análise manual em um conjunto constituído por dez textos, extraídos do corpus em formato digital do Jornal Público dos anos de 1994 e 1995, os quais fazem parte do conjunto aprendizado do corpus em Português Europeu PE. O objetivo desta análise manual foi identificar nos textos selecionados um padrão composicional, no que se refere à categorização morfo-sintático-semântica. Buscou-se características que fossem passíveis de comporem um conjunto de regras, as quais pudessem servir de base a um sistema automático para análise textual.

Assim, determinamos quatro módulos básicos para a realização da análise:

1. Módulo para identificação, classificação e segmentação dos constituintes textuais;
2. Módulo para organização arbórea – DTS's dos constituintes textuais;
3. Módulo para determinação das relações retóricas nas DTS's;
4. Módulo para identificação da macroproposição e produção da macroestrutura discursiva.

2.3. Etapas da Metodologia – AuTema-Dis

Módulo 1 – Identificação e Segmentação dos Constituintes Textuais

A primeira etapa da metodologia tem como objetivo específico identificar as estruturas que constituem um texto e segmentá-las de acordo a relação e importância com o tema. Para desenvolver o módulo de segmentação, realizamos uma análise manual, conforme mencionamos, em estudo piloto, assim, identificamos manualmente as regularidades estruturais no nível morfossintático. A seguir, recorremos ao *parser* Palavras para analisar automaticamente o mesmo conjunto de textos, verificamos os resultados gerados e comparamos com as análises manuais.

Considerando os resultados de ambas análises, buscou-se um padrão, uma regularidade na organização das estruturas dos textos e, a partir dessa classificação, elaboramos uma base heurística, constituída por um conjunto de regras de cunho morfossintático. As regras são formais, no sentido de serem passíveis de implementação em sistema computacional, fortemente orientadas por características morfológicas, sintáticas e, em algumas ocorrências, semânticas.

Palavras – Analisador Automático

Dos analisadores desenvolvidos para realizar automaticamente análise sintática, o que melhor apresentou resultados em comparação com a análise manual foi o *Palavras*, desenvolvido por Bick (2000).

As unidades textuais, definidas por Marcu (1997), Carlson e Marcu (2001) como unidades mínimas, ou EDU's, também corroborada por Pardo (2005) foram avaliadas para comporem a metodologia deste trabalho. Assumimos, conforme os autores, que as Edu's são as menores unidades de significação entre as quais são estabelecidas relações do tipo retóricas, entre outras possíveis.

Conforme pode ser evidenciado na figura 2, o resultado produzido pelo analisador Palavras apresenta as estruturas do texto devidamente etiquetadas com a codificação morfossintática e, em alguns casos, com alguma notação semântica. Além de delimitar o nível em que se encontra a estrutura analisada, a qual revela-se importante para categorização dos segmentos e subsegmentos, o *Palavra*s fornece uma codificação específica para cada elemento lexical, organiza e hierarquiza as estruturas conforme no nível da frase. A partir da marcação do nível em que se encontra cada uma das unidades, propusemos regras para classificar as estruturas do texto em *segmentos* e *subsegmentos,* conforme o nível de profundidade.

```
STA:fcl
=SUBJ:np
==>N:art('o' <artd> F S) A
==H:n('camara' F S) camara
=P:v-fin('nomear' PS 3S IND) nomeou
=,
=ADVL:adv('entretanto' <kc>) entretanto
=,
=PRED:np
==>N:art('um' <arti> M S) um
==H:n('grupo' <HH> M S) grupo
==N<:pp
===H:prp('de') de
===P<:n('trabalho' <am> <act-d> M S) trabalho
(restante conteúdo omitido)
```

Figura 2: Resultado do Palavras para um excerto do texto do Jornal Público-19950726-079.

Conjunto de regras para a segmentação

Conforme apresentamos, a partir dos resultados das análises manual e automática do *Palavras* nos textos, foi possível evidenciar regularidades na organização das estruturas e as relações estabelecidas entre elas, tais regularidades foram categorizadas em um conjunto de regras, conforme o número de ocorrências na totalidade dos corpora.

A base para a constituição das regras foi feita a partir do estudo piloto, utilizando a etiquetagem apresentada pela análise do *Palavras.* As regras foram devidamente implementadas em *prolog* e são utilizadas para: identificação dos constituintes textuais; classificação das unidades textuais em segmentos e subsegmentos; delimitação dos limites e fronteiras nas estruturas das quais fazem parte nos textos. Existem regras para determinar o que é um *segmento* ou *subsegmento,* conforme figuras 3 e 4.

Regra para Segmento	**Identificação da Regra**
UTT: acl	Enunciado sem verbo - títulos, manchetes (jornal) e cabeçalhos

EXC:fcl	Estrutura Exclamativa
QUE:fcl	Estrutura Interrogativa
NPHR:prop	Enunciado sem verbo - com estrutura nominal própria (nome próprio)
NPHR:np	Enunciado nominal
STA:fcl	Enunciado com oração finita

Figura 3: A tabela apresenta as regras para a identificação dos segmentos, bem como a sua definição terminológica.

Regras para Subsegmentos	**Identificação da Regra**
N<:fcl	Informação Acessória /Complementar
Advl:pp	Circunstância Genérica
Advl:advp	Circunstância Genérica
Advl:fcl	Circunstância Genérica
Pred:pp	Circunstância Genérica
Advl:cu	Circunstância Genérica
App:prop	Circunstância Apositiva Nome Próprio
Advl:acl	Avaliação
Sta:icl	Ação
Co:conj-c (’mas’)	Oposição / Antítese
Pred:np	Elaboração - Circunstância Genérica
App:np	Complementação Nominal Apositiva
Advl:adv - atemp	Quantificação Temporal
Advl:adv - aloc	Quantificação Locativa
Advl:np ou Advl:n	Circunstância de Tempo Decorrido

Figura 4: A tabela apresenta regras para identificação dos subsegmentos, bem como a sua classificação terminológica.

Módulo 2 – Organização Arbórea – DTS’s

O segundo módulo da metodologia prevê a organização dos constituintes textuais, identificados no módulo 1, em árvores tipo DTS’s. Trata-se de uma organização estrutural idealizada a partir da interação de características linguísticas (morfo-sintático-conceituais), devidamente etiquetadas direcionadas à representação automática do texto hierarquizado. No que se refere às características estruturais ponderadas para a organização dos constituintes textuais nas árvores, essas foram identificadas tendo em conta as informações advindas do módulo 1, isto é, identificação dos segmentos/subsegmentos e segmentação das estruturas.

O resultado do Palavras provê o nível de profundidade em que cada um dos constituintes se encontra no interior da estrutura da qual faz parte. A categorização dos níveis de profundidade é relevante para determinar a composição e a organização hierárquica das árvores DTS’s. A classificação dos níveis de profundidade apresentados pela análise realizada pelo Palavras fornece os dados que podem ser incorporados às regras de segmentação, o que pode ser evidenciado na figura 5, na sequência deste trabalho.

A marcação dos níveis em que se encontram os segmentos está associada às regras de segmentação e em conformidade com o nível de profundidade que os constituintes ocupam nas estruturas é possível identificar, em algumas situações, o tipo

de relação que se estabelece entre eles, seja ela, hipotaxis ou parataxis. Desta forma, os constituintes identificados automaticamente nas árvores DTS's são organizados a partir da interação entre as características sintático-estruturais independente de serem hipotáticas ou paratáticas, a organização arbórea possibilita-nos classificá-los de acordo com o papel que desempenham na estrutura, isto é, se o constituinte desempenha o papel de segmento ou de subsegmento.

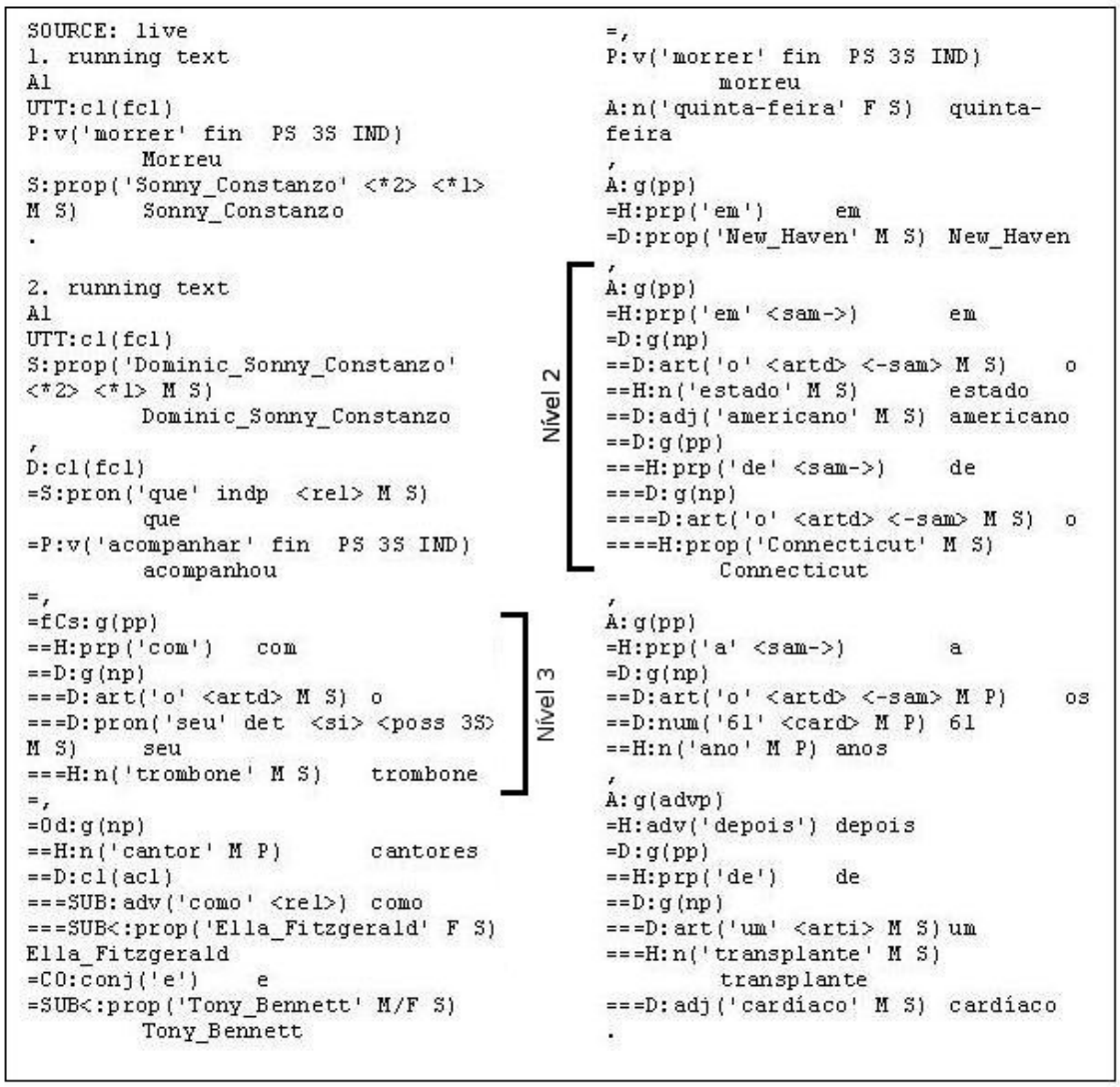

Figura 5: A figura apresenta a análise automática do Palavras com a marcação dos níveis de profundidade em que se encontram os constituintes na estrutura – texto publico-19950726 – 079

As DTS's organizam os segmentos principais como "nós" do nível 1 da árvore e os subsegmentos são identificados como "nós" de níveis 2 e 3, conforme figura 6. Os demais subsegmentos, ou seja, aqueles que se encontram em níveis de profundidade posterior ao 3º nível, isto é, subsegmentos do 4o e 5o níveis, não são organizados de maneira hierárquica em "nós", são mantidos em "nós" únicos e indissociáveis, conforme apresentado na figura 6.

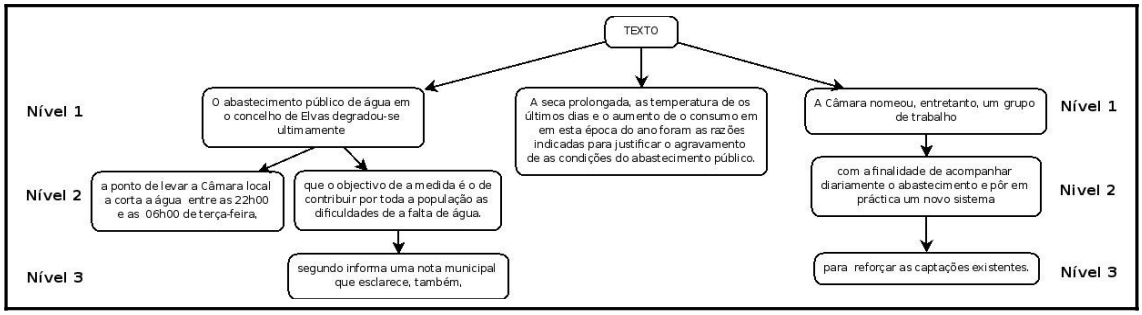

Figura 6: A figura representativa da organização hierárquica de um texto em uma DTS com especificação dos níveis – texto publico-19950726-079 .

As Regras de Segmentação Textual e Identificação dos Níveis dos Constituintes

Conforme apresentamos, o objetivo do módulo 2 da metodologia é organizar os segmentos e os subsegmentos em árvores do tipo DTS's, conforme a representatividade e o compromisso que cada uma das estruturas desempenha em relação ao tema do texto. Para realizar tal organização, a metodologia prevê a utilização das regras de segmentação, elaboradas para a identificação dos segmentos no módulo 1, e de informação complementar relacionada ao nível de profundidade em que se encontram cada um dos constituintes no texto.

Definimos na metodologia que as regras de segmentação recebem a caracterização de nível de profundidade, isto é, cada regra que identifica um segmento ou um subsegmento traz agregada uma informação que determina o lugar que cada segmento ou subsegmento pode ocupar na árvore DTS, conforme evidenciamos na figura 7. A priori, os constituintes identificados como *segmentos* ocupam os nós de 1o nível; já os constituintes identificados como *subsegmentos* ocupam os nós de 2o e 3o níveis. Nota-se que a análise realizada pelo Palavras determina níveis de profundidade além do 2º e 3º, entretanto, para fins metodológicos e de implementação do sistema, optou-se em considerar para segmentação e estruturação arbórea apenas estes dois níveis, ficando dos demais níveis associados aos nós de nível 2 e 3.

Níveis de Profundidade & Regra para Segmento
Todos segmentos encontram-se no nível 1 de profundidade na estrutura UTT : acl; EXC : fcl; QUE : fcl; NPHR : prop; NPHR : np; STA : fcl
Níveis de Profundidade & Regras para Subsegmentos
Todos segmentos encontram-se nos níveis 2 e 3 de profundidade na estrutura N<:fcl; Advl:pp; Advl:advp; Advl:fcl; Pred:pp; Advl:cu; App:prop; Advl:acl; Sta:icl Co:conj-c ('mas'); Pred:np; App:np; Advl:adv – atemp; Advl:adv – aloc; Advl:np ou Advl:n

Figura 7: Regras de segmentação e os níveis de profundidade que os segmentos e os subsegmentos podem ocupar em uma estrutura.

Módulo 3 – Identificação das Relações Retóricas em DTS´s

O módulo 3 apresenta a metodologia para a identificação automática das relações retóricas entre segmentos e subsegmentos a partir da configuração em DTS's. Para a

realização do processo de análise manual e posterior automatização, utilizou-se dois conjuntos de relações retóricas, o conjunto proposto por Mann e Thompson (1988) e o conjunto proposto por Marcu (2001). Apesar de contarmos com dois grupos de relações, as relações retóricas e estruturais propostas por ambos autores não foram utilizadas na íntegra em nosso estudo, optamos pela utilização de apenas algumas das relações de Mann e Thompson e de Marcu em função das relações identificadas nos textos dos corpora. Das duas propostas analisadas, foram utilizadas as seguintes relações retóricas:

- Mann and Thompson – Relações Retóricas: circunstância; Avaliação; Antítese, Elaboração.
- Daniel Marcu – Relações Estruturais: Same-unit; Parentética.

Além das relações selecionadas dos dois grupos mencionados, foi necessário identificar novas relações, desenvolvidas exclusivamente para suprir particularidades evidenciadas nos textos dos corpora, sendo algumas delas exclusivamente sintáticas e outras discursivas. As novas relações foram definidas em conformidade com a teoria das relações retóricas que prevê na sua origem a possibilidade de livre ampliação, de acordo com as particularidades de cada texto, com os critérios e objetivos estabelecidos pelo analista e com a finalidade da análise. Neste sentido, além das seis relações retóricas utilizadas, provenientes do rol das relações de Mann e Thompson e Marcu, agregou-se mais cinco relações, são elas: Apositiva de Nome Próprio; Quantificação Temporal, Quantificação Locativa, Ação, Circunstância de Tempo Decorrido.

A constituição deste módulo 3 conta com as informações advindas dos módulos 1 e 2. O módulo 1 identifica e segmenta os constituintes do texto; o módulo 2 recebe e processa as informações do módulo 1, classifica e organiza os constituintes em árvores de dependência. O módulo 3 utiliza a organização arbórea, realizada no módulo 2 e atribui algumas relações retóricas entre os segmentos, nós de 1o nível, e subsegmentos, nós de 2o e 3o níveis, em conformidade com o tipo de estrutura etiquetada pelo *Palavras* e o nível em que se encontra o segmento, conforme figura 8.

Relações Retóricas	Regra para Segmento
A definir	UTT : acl EXC : fcl QUE : fcl NPHR : prop NPHR : np STA : fcl
Relações Retóricas	Regra para Subsegmento
Same-Unit	N<:fcl
Circunstância Genérica	Adv1:pp Adv1:advp Adv1:fcl Pred:pp Adv1:cu
Circunstância Apositiva Nome Próprio	pp:prop
Avaliação	Adv1:acl
Ação	Sta:icl
Oposição/Antítese	Co:conj-c ('mas')
Circunstância Apositiva Nome Próprio	Pred:np
Complementação Nominal Apositiva	App:np
Quantificação Temporal	Adv1:adv - atemp
Quantificação Locativa	Adv1:adv - aloc
Circunstância de Tempo Decorrido	Adv1:np ou Adv1:n

Figura 8: Relações retóricas indexadas às regras para a identificação dos segmentos e subsegmentos.

As relações retóricas elegidas perfazem um total de 11 relações, sendo estas relações apenas para as ligações entre os segmentos e subsegmentos, e subsegmentos – subsegmentos. No momento, a metodologia ainda não está desenvolvida para contemplar as relações entre dois ou mais segmentos e alguns casos de relações entre segmentos e subsegmentos, o que justifica o número reduzido de relações implementadas no sistema. A figura 9 apresenta uma DST com as relações retóricas entre os segmentos e os subsegmentos.

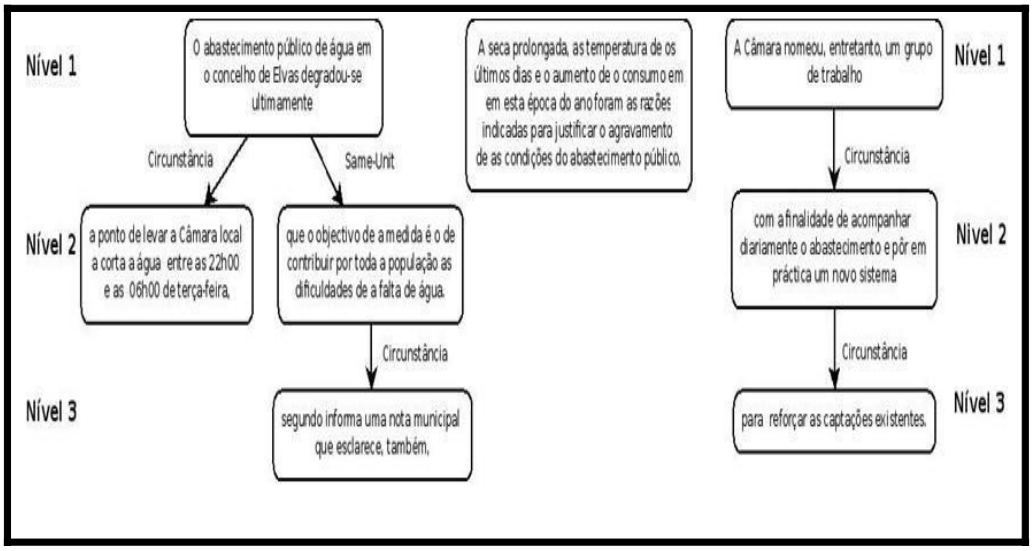

Figura 9: Árvore DST com as relações retóricas entre os constituintes hierarquizados.

Módulo 4 – Representação Estrutural da Macroproposição Textual

O módulo 4 manipula os segmentos e subsegmentos devidamente identificados e dispostos em árvores, em que cada um dos constituintes desempenha um papel específico em relação ao tema do texto.

A identificação dos segmentos e subsegmentos que participam da composição da macroestrutura/macroproposição está relacionado com os níveis de profundidade. Observou-se, pelas análises realizadas nos corpora que, quanto mais interno estiver um subsegmento em uma estrutura, menor será o seu comprometimento com o tema global, isto é, quanto mais afastado estiver o subsegmento da estrutura que ocupa a posição do nó principal, ou nó de primeiro nível, maior será a probabilidade deste constituinte ser considerado descartável em relação à composição do tema. A seleção dos segmentos a participarem da macroproposição está condicionado à aplicação das macrorregras, conforme propõe Dijk (1992), Alguns subsegmentos tornam-se candidatos a não participarem da composição da macroproposição/macroestrutura, sendo eliminados pela aplicação da macrorregra de apagamento ou supressão.

Considerando os resultados da análise realizada nos corpora e em conformidade com a possibilidade de implementação deste módulo em sistema computacional, optamos pela utilização de apenas uma das três macrorregras, isto é, utilizamos a macrorregra apagamento ou supressão. A opção pela aplicação da macrorregra de apagamento deve-se às restrições de implementação deste módulo em sistema

computacional. As outras macrorregras são importantes, no entanto, no nível em que se encontra esse estudo e devido à complexidade que exige a implementação das outras duas macrorregras, optamos por trabalhar apenas com a que fosse imediatamente passível de implementação.

Em um sentido pragmático, o módulo 4 prevê a seleção dos constituintes que ocupam os nós de primeiro nível e alguns do nível 2, de acordo com as regras de eleição dos segmentos e subsegmentos. Selecionados os constituintes, organiza-se a macroestrutura/macroproposição do texto analisado.

Figura 10: A figura apresenta a macroestrutura/macroproposição de um texto processado pelo AuTema-Dis.

3. Avaliação da Metodologia em Sistema Computacional

A avaliação de cada um dos módulos previstos na metodologia é realizada de duas formas, isto é, manual e automática, e os resultados podem ser evidenciados, conforme demonstramos, em tabelas com os resultados e dados estatísticos. Outrossim, ressaltamos que o processo de avaliação realizado foi executado de forma especial, permitindo-nos dar cobertura a dois processos distintos, isto é:

- elaboração da metodologia proposta para análise discursiva automática;
- constituição e execução do sistema computacional, AuTema-Dis, estruturado a partir da metodologia proposta.

A execução computacional é a forma pragmática empregada para avaliarmos e validarmos a metodologia desenvolvida, especificamente, para análise automática discursiva e para avaliarmos a execução do sistema usamos as seguintes medidas: precisão – *precision*, cobertura – *recall* e *F-measure*. Especificamente ao caso da

avaliação da macroestrutura/macroproposição, não foi possível aplicar a medida de precisão, por se tratar de uma produção de caráter subjetivo. Assim, a avaliação da execução do sistema na realização do módulo 4, que concerne à apresentação automática da estrutura temática do texto, foi realizada em termos de coerência/incoerência e completude/incompletude.

3.1. Módulos Avaliação e Resultados

Identificação e Classificação dos Constituintes

A avaliação da metodologia proposta para o módulo 1 é realizada considerando-se: as regras elaboradas para identificação e segmentação dos constituintes textuais, implementadas em sistema computacional e a sua posterior execução e validação dos resultados obtidos. Os resultados podem ser conferidos nas figuras 11 e 12.

Segmentos - Totalidade dos Corpora			
	Precisão	Cobertura	F-Measure
Macromédia	0.76	0.75	0.75
Micromédia	0.77	0.75	0.76

Figura 11: Identificação dos Segmentos.

Subsegmentos - Totalidade dos Corpora			
	Precisão	Cobertura	F-Measure
Macromédia	0.65	0.63	0.63
Micromédia	0.73	0.67	0.70

Figura 12: Identificação dos Subsegmentos.

Os dados descritos nas figuras 11 e 12 representam os resultados obtidos com a execução do sistema no módulo 1, no que diz respeito a identificação dos *segmentos* e *subsegmentos*. Os dados representam a totalidade dos textos dos corpora, ou seja, 40 textos, 20 em Português Europeu e 20 em Português Brasileiro.

Organização Automática das DTS's – árvores de dependência dos segmentos

O módulo 2 apresenta a metodologia desenvolvida para realizar computacionalmente a classificação e a organização dos constituintes textuais em árvores do tipo DTS's, previamente identificados no módulo 1.

A partir da implementação computacional da metodologia proposta no módulo 2, foi possível executar automaticamente a identificação e a organização dos segmentos e subsegmentos nas DTS's sem a interferência humana. Especificamente, na sequência do processamento, o módulo 2 recebe os resultados da etapa concluída no módulo 1, a qual identifica os constituintes textuais sem classificá-los, transformando esses resultados, até então não categorizados, em um novo conjunto de regras, as quais são utilizadas na geração automática da estrutura arbórea.

Na figura 13, é possível evidenciar os resultados da performance do sistema Autema em organizar as estruturas nas DTS´s em comparação com a organização manual. Além disso, medimos a precisão do sistema em identificar

adequadamente o nível de profundidade das estruturas. Observamos que a avaliação neste nível foi realizada considerando-se todos os textos dos corpora, inclusive do aprendizado.

Regras e Representação DTS na totalidade dos corpora - 50 textos			Correção	
Regras	Nº de Ocorrências Manual	Nº Ocorrências Sistema	nº Profundidade	%
N<:fcl	38	40	29	73%
Advl:pp	239	205	148	72%
Advl:advp	15	16	15	94%
Advl:fcl	30	34	26	76%
Pred:pp	19	16	13	81%
Advl:cu	7	7	7	100%
App:prop	16	14	12	86%
Advl:acl	8	8	3	38%
Sta:icl	10	13	8	62%
Co:conj-c ('mas')	14	12	11	92%
Pred:np;	13	7	0	0%
App:np	8	10	8	80%
Advl:adv - atemp	52	47	25	53%
Advl:adv - aloc	1	3	0	0%
Advl:np ou Advl:n	6	8	3	38%
Média	31.7	29.3	20.5	70%

Figura 13: Representação automática das DTS.

Identificação Automática das Relações Retóricas em DTS's

Na edificação do módulo 3, foi previsto a atribuição automática de algumas relações retóricas entre segmentos e subsegmentos em textos em Língua Portuguesa. Devido às questões de ordem semântica, esta etapa foi avaliada de maneira diferenciada em relação as demais etapas implementadas computacionalmente. Optou-se por realizar dois tipos de avaliações específicas, são elas:

- uma avaliação holística – em que se verifica qualitativamente a execução do sistema AuTema-Dis, no processo relativo à atribuição das relações retóricas entre os segmentos e subsegmentos;
- uma avaliação pontual – em que se avalia quantitativamente a precisão do sistema AuTema-Dis em atribuir corretamente as relações retóricas entre os segmentos e subsegmentos.

Os resultados obtidos podem ser evidenciados abaixo na figura 14 .

A figura 14 apresenta quais relações foram atribuídas automaticamente, quantas estavam corretas e quantas realmente são necessárias comparando-se com os resultados da atribuição manual.

A partir dos resultados evidenciados através das medidas de avaliação, verificou-se que o AuTema-Dis está capacitado a realizar a tarefa de atribuir relações retóricas entre os segmentos, apesar das suas limitações. Todavia, salientamos, que os resultados obtidos nesta etapa do processamento são satisfatórios no âmbito desta proposta.

Avaliação da Automatização das Relações Retóricas - 50 textos				
Relações	Regra	N de Rel. Auto.	N de Rel. Corretas	N de Rel. Existentes.
Same-Unit	N<:fcl	40	29	38
Circunstância Genérica	Adv1:pp Adv1:advp Adv1:fcl Pred:pp Adv1:cu	200 15 33 16 7	148 15 26 13 7	239 15 30 19 7
Circunstância Apositiva Nome Próprio	App:prop	14	12	16
Avaliação	Adv1:acl	8	3	8
Ação	Sta:icl	11	8	10
Oposição Antítese	Co:conj-c ('mas')	12	11	14
Elaboração Circunstância Genérica	Pred:np	6	0	13
Complementação Nominal Apositiva	App:np	8	8	8
Quantificação Temporal	Adv1:adv - atemp	47	25	52
Quantificação Locativa	Adv1:adv - aloc	3	0	1
Circunstância de Tempo Decorrido	Adv1:np ou Adv1:n	6	3	6

Figura 14: Correção dos resultados na atribuição das relações retóricas

Identificação Automática da Macroestrutura/Macroproposição

Para avaliarmos a produção e a apresentação da macroestrutura/macroproposição gerada automaticamente pelo sistema AuTema-Dis, são efetuados dois processos:

1. avaliação das regras para a organização macroestrutural.
2. avaliação do sistema constituído por estas regras.

Inicialmente, verificamos manualmente se as regras desenvolvidas para identificação, seleção e organização da macroestrutura/macroproposição estão adequadas a realizarem as tarefas estabelecidas. Em um momento posterior, procedemos a sua implementação no sistema computacional AuTema-Dis. Em relação à avaliação do resultado da produção da macroestrutura/macroproposição foi realizada uma análise contrastiva entre a identificação da estrutura representativa do tema proposto pelo analista e a identificação apresentada como resultado da execução do sistema AuTem-Dis, conforme pode ser observado nas figuras 15 e 16.

Avaliação da Macroestrutura/Macroproposição Automática Conjunto Avaliação - Corpus Folha de São Paulo - 1994/1995				
Textos	Correção	Nº de Erros	Erro Palavras	Erro Regras
Nº FSP950101-011	90%	1	0	1
Nº FSP950101-032	80%	1	1	0
Nº FSP950101-054	100%	0	0	0
Nº FSP950101-084	50%	5	5	0
Nº FSP950111-014	100%	0	0	0
Nº FSP950111-026	100%	2	2	0
Nº FSP950111-034	100%	0	0	0
Nº FSP950111-036	100%	0	0	0
Nº FSP950117-048	90%	2	1	1
Nº FSP950117-074	100%	0	0	0
Nº FSP940101-132	80%	2	2	0
Nº FSP940101-124	90%	2	2	0
Nº FSP940101-107	100%	0	0	0
Nº FSP940101-102	100%	0	0	0
Nº FSP940101-095	100%	0	0	0
Nº FSP940101-092	80%	3	1	2
Nº FSP940101-085	100%	2	2	0
Nº FSP940101-079	90%	1	1	0
Nº FSP940101-074	100%	0	0	0
Nº FSP940101-066	100%	1	1	0
Média	92.5%	1.9	0.9	0.2

Figura 15: Identificação Automática da Macroestrutura/Macroproposição nos textos em Português Brasileiro.

Avaliação da Macroestrutura/Macroproposição Automática Conjunto Avaliação - Corpus Jornal Público - 1994/1995				
Textos	Correção	Nº de Erros	Erro Palavras	Erro Regras
Nº 19940504-070	90%	1	0	1
Nº 19940505-024	80%	1	1	1
Nº 19940505-071	80%	2	1	1
Nº 19941911-083	70%	3	2	1
Nº 19941012-011	100%	0	0	0
Nº 19941025-045	70%	2	1	1
Nº 19950416-032	60%	2	2	0
Nº 19950795-167	70%	3	3	0
Nº 19950912-022	60%	2	2	0
Nº 19950924-121	50%	3	2	1
Nº 19950422-141	60%	1	0	1
Nº 19950423-011	50%	4	3	1
Nº 19950629-083	100%	0	0	0
Nº 19950629-119	70%	3	2	1
Nº 19951011-139	90%	1	0	1
Nº 19951011-150	100%	0	0	0
Nº 19951114-163	90%	1	0	1
Nº 19951114-169	80%	2	0	2
Nº 19951220-045	100%	1	1	0
Nº 19951229-044	100%	0	0	0
Média	78.5%	1.6	1	0.65

Figura 16: Identificação Automática da Macroestrutura/Macroproposição nos textos em Português Europeu.

4. Considerações Finais

O trabalho desenvolvido contribui para a compreensão e entendimento de como se pode processar automaticamente a identificação estrutural/formal e a organização conceitual de textos escritos em Português Brasileiro e Português Europeu. Em lato sensu, apresentamos abaixo as contribuições que se destacam na proposta em questão.

4.1. A Metodologia

A metodologia utilizada no desenvolvimento desta pesquisa é constituída por etapas diferenciadas de análise textual, as quais representam hierarquicamente a estruturação formal e a estruturação conceitual de textos em Português:

1. segmentação automática dos textos em PB e PE em segmentos e subsegmentos, em conformidade com as características formais e conceituais dos corpora;
2. identificação e atribuição automática de algumas relações retóricas RST entre os segmentos e subsegmentos dos textos dos corpora;
3. identificação e produção automática de uma estrutura representativa da macroproposição dos textos dos corpora, analisados pelo sistema.

4.2. Aplicabilidade do sistema teórico-computacional

A aplicabilidade dos resultados obtidos na investigação pode ser constitutiva em pesquisas que envolvam:

1. sistemas computacionais capazes de recuperar informação contida em textos de forma automática ou em sistemas de pergunta-resposta.

2. sistemas computacionais que realizam automaticamente a sumarização em textos em linguagem natural;

4.3. Metodologia AuTema-Dis: limitações

A metodologia AuTema–Dis apresenta algumas limitações, nomeadamente:

1. tratamento adequado ao elemento verbal, considerando-se a construção de uma ontologia robusta para os verbos.
2. identificação de um conjunto muito restrito de relações retóricas e relações estruturais, ponderando-se a possibilidade de desenvolver regras mais específicas para uma identificação mais conceitual.
3. produção pouco refinada para a macroproposição/macroestrutura.
4. resultado da análise automática do *Palavras* – o analisador em algumas ocasiões apresenta falhas na análise e geração de resultados.
5. reconhecimento dos pronomes anafóricos e indexação nominal. A metodologia AuTema- Dis não prevê a resolução dos anafóricos.
6. tratamento dos marcadores discursivos a partir das relações retóricas instituídas por esses elementos linguísticos.

A proposta metodológica para análise automática discursiva e a sua implementação em sistema computacional apresentada no âmbito deste trabalho, representou um desafio, devido à interdisciplinaridade envolvida no processo. Verificou-se a necessidade e a dificuldade em compor a metodologia que realizasse exaustivamente a análise textual completa, sem a interferência humana e que, além disso, fosse passível à implementação em sistema computacional, o qual encontra-se disponível à comunidade em http://analu.xdi.uevora.pt/interfaceweb/.

5. Referências Bibliográficas

Bick, E. (2000) The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press.

Carlson, L. and Marcu, D. (2001) Discourse tagging reference manual. Technical Report ISITR545, ISI Technical Report.

Dijk, T. A. V. (1972). Some Aspects of Text Grammars. The Hague:Mouton.

Dijk, T. A. V. (1992) Cognição, discurso e interação. São Paulo: Contexto.

Mann, W. and Thompson, S. (1988) Rhetorical structure theory: toward a functional theory of text organization. 3(8):243–281.

Marcu, D. (1997) The Rhetorical Parsing, Summarization and Generation of Natural Language Text Organization. PhD thesis.

Marcu, D. (2000) Extending a formal and computational model of rhetorical structure theory with intentional structures à la grosz and sidner. In The 18th International Conference on Computational Linguistics (COLING2000).

Leal, A. L. Quaresma P, and R. Chishman. (2006) From syntactical analysis to textual segmentation. In Vieira et al., PROPOR, Springer LNAI 3960. pages 252–255

Pardo, T. (2005) Métodos para Análise Discursiva Automática. PhD thesis.