

UNIVERSIDADE DO VALE DO RIO DOS SINOS

CENTRO DE CIÊNCIAS DA COMUNICAÇÃO

**USO DE INFORMAÇÃO DE CORREFERÊNCIA E  
ANÁFORA PARA VERIFICAÇÃO DA COESÃO E COERÊNCIA  
TEXTUAL NA SUMARIZAÇÃO AUTOMÁTICA**

AUTOR: JORGE CÉSAR BARBOZA COELHO

**MONOGRAFIA DE CONCLUSÃO DO CURSO DE LETRAS**

Orientadora: Renata Vieira

São Leopoldo, junho de 2007

## RESUMO

Face ao desenvolvimento das chamadas Tecnologias de Informação e Comunicação - TICs, a pesquisa e a construção de sistemas de Processamento da Língua Natural - PLN ganham uma posição de destaque. Área essa que impõe diversos desafios, pois a habilidade humana de comunicação é sustentada por um sistema extremamente complexo. Dessa forma, a qualidade do resultado produzido por muitos desses sistemas ainda é bastante comprometida.

É nesse contexto que se insere este trabalho. Entre diversos sistemas e aplicações do PLN, este trabalho se ocupa do estudo da sumarização automática e, em especial, da qualidade do resultado produzido por tais sistemas. O aspecto qualidade aqui, muito discutido na literatura da área de Estudos do Discurso com a denominação de “coerência textual”, restringiu-se principalmente a questões de coesão referencial. Esse enfoque deve-se ao fato de que os estudos de coesão referencial são muito próximos aos estudos de sistemas de resolução de correferência, e o princípio que baseou esta pesquisa é o de que informações sobre as cadeias de correferência originais de um texto (que podem em parte ser obtidos automaticamente por esses sistemas) podem guiar o processo de sumarização automática com vistas à produção de sumários com melhor qualidade.

Este trabalho baseou-se em um corpus constituído de 50 textos do corpus PLN-BR e seus sumários, bem como na anotação lingüística nos níveis morfossintático, semântico e de discurso (relações anafóricas e de correferência). A própria construção do corpus para esse estudo foi parte integrante deste trabalho.

Palavras-chave: correferência e anáfora, sumarização, coesão e coerência textual, processamento da língua natural.

## **ABSTRACT**

With the development of Information and Communication Technologies - ICTs, the research and the construction of Natural Language Processing - NLP systems are becoming more prominent. This area imposes diverse challenges, since the human communication ability is supported by an extremely complex system. Therefore, often the results produced by these systems are far from perfect.

In this context, this work is presented. Among many systems and NLP applications, this work focuses on the study of automatic summarization, in special, on the quality of the results produced by these systems. The quality aspect here, much discussed in the literature of the area of the Discourse Studies with the denomination of “textual coherence”, is mainly restricted to the questions of referential cohesion. This approach was chosen because the studies of referential cohesion are close to the studies of coreference resolution systems. We believe that the information about the coreference chains in texts can guide the automatic summarization process, aiming at the generation of summaries with higher quality.

This work is based on a corpus consisting of 50 texts extracted from the corpus PLN-BR and their summaries, containing also the linguistic annotation regarding lexical, syntactic, semantic and of the discourse (anaphoric and coreference relations) levels. The building the corpus for this study was also a part of this work.

Word-key: coreference and anaphora, summarization, textual cohesion and coherence, natural language processing.

*“Dedico este trabalho a minha  
adorada mãe, minha sublime fonte de  
felicidade, inspiração, motivação e força.”*

## AGRADECIMENTOS

*Agradeço a todos que me ajudaram, colegas, professoras e amigos. Seria difícil mencionar todos aqui. Entretanto, quebrando o protocolo, preciso agradecer em especial a brilhante professora, pesquisadora e orientadora Renata Vieira. Também, não posso deixar de reconhecer toda a ajuda dos meus queridos amigos do Laboratório de Engenharia da Linguagem - UNISINOS e do NILC/UFSCar, todos amigos que tenho orgulho de trabalhar. Entre eles preciso iluminar minha terna amiga de todas as horas, referência como pessoa e profissional, Sandra Collovini de Abreu.*

## SUMÁRIO

INTRODUÇÃO .....	7
1 FUNDAMENTAÇÃO TEÓRICA .....	15
1.1 Conceção de texto .....	16
1.2 Coerência textual .....	18
1.2.1 Fatores de coerência .....	21
1.3 Coesão referencial .....	34
1.4 Sumarização Automática.....	48
2 CORPUS <i>SUMM-IT</i> .....	58
2.1 Seleção do corpus.....	59
2.2 Anotação do corpus .....	60
2.2.1 Anotação XCES.....	62
2.2.2 Anotação automática dos programas PALAVRAS e Palavras Xtractor .....	64
2.2.3 Anotação de correferência e anáforas .....	67
2.3 Sumários .....	76
2.3.1 Elaboração dos textos tarjados e sumários manuais .....	77
2.3.2 Geração dos extratos automáticos .....	82
3 SUMARIZAÇÃO, CORREFERÊNCIA E COESÃO REFERENCIAL.....	86
3.1 Estudo dos sumários automáticos sob a ótica das cadeias de correferência .....	88
CONSIDERAÇÕES FINAIS .....	97
REFERÊNCIAS BIBLIOGRÁFICAS.....	102
ANEXOS .....	111
Anexo 1 - Relatório das atividades realizadas pelo autor desta monografia .....	112
Anexo 2 - Instruções para anotação de relações anafóricas e referência dêitica.....	115
Anexo 3 - Exemplo de relatório de cadeias de correferência .....	125
Anexo 4 - Anotação RST .....	127
Anexo 5 - Orientações gerais para equipe de sumarizadores.....	130

## INTRODUÇÃO

A humanidade está num processo muito dinâmico de mudanças em todas as dimensões de sua existência. Um fato novo que parece estar na raiz dessas mudanças é a digitalização crescente da informação e a evolução dramática da tecnologia para lidar com a informação digitalizada. Alguns acham que se iniciou uma grande revolução, parecida talvez com a Revolução Industrial, mas outros objetam o uso do termo revolução. Edgar Morin (1999) prevê que durante o processo em curso todas as instituições da atual existência serão renegociadas diante da presente revolução da tecnologia da informação, ou melhor, diante da sociedade da informação. Revolução ou não, poucos duvidam do fato de que a evolução tecnológica está transformando de modo muito rápido os aspectos econômicos, sociais e culturais da humanidade. Essas quebras de paradigmas estão relacionadas também a um outro fato intimamente ligado à tecnologia da informação: o surgimento de uma comunidade digital, marcadamente, polimórfica, plurilíngüe e multicultural. Ela, com efeito, tem promovido uma democratização do acesso ao conhecimento e potencializado o avanço, principalmente, da ciência e da economia.

Entretanto, a sociedade da informação também pode ser vislumbrada como a sociedade da demasia, da dispersão e da transitoriedade da informação, uma vez que a geração do conhecimento é marcada pela proliferação desenfreada, desarticulada e multifacetada de conteúdos. A maior disponibilidade dos conteúdos não tem implicado, necessariamente, comunicação de melhor qualidade. O que se percebe é a entropia do sistema de informação global aumentando progressivamente, a ponto de motores de busca e caixas de mensagens se congestionarem de contra-informações e pseudo-informações, interferindo negativamente nas potencialidades comunicativas das redes de computadores.

Ao entender que a elaboração de protocolos e políticas de produção e difusão de conteúdos é improvável na atual conjuntura de globalização, se instala, em caráter de urgência, o desenvolvimento de estratégias para tratar a expressiva dispersão da informação do lado da recepção. No que toca o ciberespaço – Internet e intranets – isso envolve, inevitavelmente, a elaboração de uma representação computacional, não somente da forma, mas também do conteúdo dos arquivos digitais.

Na busca desse modelo de representação, outro aspecto da sociedade da informação emerge: a natureza extraordinariamente multidisciplinar e multicultural dos processos em curso. As conseqüências dessa afirmação são difíceis de aceitar, até porque a sua aceitação implica “dores de cabeça monumentais”, já que é consideravelmente difícil acostumar-se a uma disciplina e aprofundar-se nela a ponto de dominá-la. Mais complicado é fazer isso em várias disciplinas, e ser capaz de entender ainda a interação entre elas. Continuamente a ênfase parece encontrar-se na interação ampla, ativa e significativa de duas ou mais disciplinas, isto é, os estudos e conhecimentos necessários para a compreensão do dia-a-dia e para o exercício



profissional são cada vez menos verticais, de menor especialização, para serem cada vez mais horizontais, de maior abrangência temática.

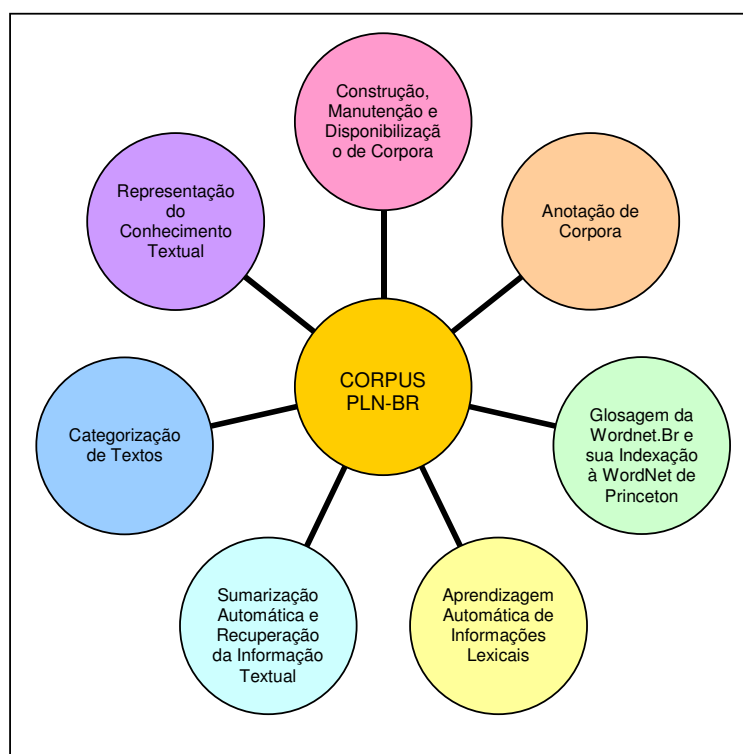
Diante desse panorama, vem se consolidando um programa de investigação científica que, genericamente, é denominado de “Ciências da Informação”, que corresponde, efetivamente, a um empreendimento multidisciplinar, com desdobramentos nas Ciências da Computação, na Lingüística, na Biblioteconomia e na Filosofia, cujas fronteiras estão sendo freqüentemente cruzadas e cuja interpenetração se torna cada vez mais evidente. Exemplo disso pode ser observado no âmbito do Processamento das Línguas Naturais - PLN, que tem por objetivo a representação e o tratamento da informação veiculada nos documentos verbais. Entretanto, nessa sinergia, surge um aspecto problemático: enquanto avança rapidamente o desenvolvimento de tecnologias para o inglês, japonês, francês e alemão; para o português, o avanço é lento – mesmo com o número expressivo de falantes. Isso é preocupante visto que, conforme Jeffrey Sachs (2002) e Edgar Morin (1999), “o mundo não é mais dividido por ideologia, mas por tecnologia”. Eles argumentam, com autoridade, sobre o poder divisivo da tecnologia e como a percepção da tecnologia e a capacidade de inovação de cada país determina o seu futuro. Observa-se que cada vez mais a habilidade de processar determinada língua natural instala-se como um fator expressivo de desenvolvimento.

A fim de enfrentar esses problemas, reuniram-se sete diferentes equipes de pesquisadores e desenvolvedores (USP/São Carlos, UFSCar, UNESP/Araraquara, PUC/RS, PUC/RJ, UNISINOS e Mackenzie) para constituir o projeto de pesquisa PLN-BR - *Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil*<sup>1</sup>.

---

<sup>1</sup> Disponível em: <http://plnbr.tripod.com/>

Numa tentativa de integração dos esforços que já vêm sendo despendidos, isoladamente, por eles, em diferentes áreas, todas elas envolvendo o tratamento computacional do português, o projeto PLN-BR visa à convergência dos problemas que, em última análise, dizem respeito à heterogeneidade e à diversidade dos recursos existentes para pesquisa do processamento de bases textuais e à criação de um espaço comunitário de ação e investigação, a partir do qual possam ser construídos recursos a serem compartilhados. Conforme mostra a Tabela 1 e a Figura 1, apesar das diferentes perspectivas de trabalho, todas as equipes têm em comum o corpus PLN-BR.



**Figura 1 – Projeto PLN-BR**

Desse grupo, duas equipes de pesquisa, o Núcleo Interinstitucional de Lingüística Computacional - NILC/UFSCar e o Laboratório de Engenharia da Linguagem da UNISINOS, estreitaram os laços de interação ao formar o projeto ProCaCoSa - *Processamento de Cadeias*

de *Correferência para a Sumarização Automática de Textos em Português*<sup>2</sup>. Esse projeto visa rastrear e diagnosticar alguns problemas desencadeados pela desconsideração dos fenômenos de natureza correferencial (correferência e anáfora), ou melhor, pelo desatendimento dos problemas decorrentes do não tratamento das cadeias de correferência durante a seleção e estruturação do conteúdo de sumários gerados automaticamente. Aqui, se considera *cadeia de correferência* uma série de eventos lingüísticos em que é possível observar relações de identidade, associação e dependência referencial.

PROJETO	RESPONSÁVEL	INSTITUIÇÃO
Construção, Manutenção e Disponibilização de Corpora	Sandra Maria Aluísio	NILC-USP/S. Carlos
Anotação de Corpora	Renata Vieira	UNISINOS
Glosagem da Wordnet.Br e sua Indexação à WordNet de Princeton	Bento Carlos Dias-da-Silva	UNESP/Araraquara
Aprendizagem Automática de Informações Lexicais	Violeta de San Tiago Dantas Barbosa Quental	PUC/RJ
Sumarização Automática e Recuperação da Informação Textual	Lúcia Helena Machado Rino	UFSCar
Categorização de Textos	Vera Lúcia Strube de Lima	PUC/RS
Representação do Conhecimento Textual	Ronaldo Martins	MACKENZIE

**Tabela 1 – As sete equipes de pesquisa do projeto PLN-BR**

É justamente na junção desses dois tópicos – resolução anafórica e sumarização automática – que se fundamenta este trabalho. Este é subordinado aos projetos ProCaCoSa e PLN-BR, subprojetos do projeto CROWS – *Construção de ontologias para Web Semântica*<sup>3</sup>, coordenado pela orientadora deste trabalho de conclusão. Este estudo, portanto, compartilha com eles motivações e metodologias. Antes de prosseguir, cabe destacar que o PLN é uma subárea da Inteligência Artificial – IA, sendo que esta é por lado uma ciência, que procura

<sup>2</sup> Disponível em: [http://www.inf.unisinos.br/~renata/laboratorio/procacosa\\_index.htm](http://www.inf.unisinos.br/~renata/laboratorio/procacosa_index.htm)

<sup>3</sup> Disponível em: [http://www.inf.unisinos.br/~renata/laboratorio/crows\\_index.htm](http://www.inf.unisinos.br/~renata/laboratorio/crows_index.htm)

estudar e compreender o fenômeno da inteligência, e por outro um ramo da engenharia, na medida em que procura construir instrumentos para apoiar a inteligência humana. IA é inteligência como modelagem computacional, que tenta emular o pensamento dos peritos e os fenômenos cognitivos humanos. A construção de máquinas inteligentes pressupõe a existência de estruturas simbólicas (representação), a capacidade de elas poderem raciocinar (procura) e a existência de conhecimentos (matéria prima).

Nessa tríplice, este trabalho se mobiliza na construção de um corpus denominado *Summ-it* (matéria prima) tratado com diferentes níveis de anotações (representação). Além disso, este texto visa reportar as tarefas realizadas pelo presente autor na busca (procura) de empregar informações de correferência e anáfora no contexto de sumarização automática. Todavia, antes, cabe registrar que este é fruto de experiências ligadas ao Laboratório de Engenharia da Linguagem. Desde 2003, houve a participação do autor, como bolsista, do projeto TeXto - *Acesso a informações em bases textuais* e do projeto DIRPI - *Desenvolvimento e Integração de Recursos para Pesquisa de Informação*. Atualmente, há o vínculo como colaborador nas equipes dos projetos CROWS - *Construção de ontologias para a Web Semântica*, ProCaCoSa - *Processamento de Cadeias de Correferência para a Sumarização Automática de Textos em Português* e PLN-BR - *Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil*.

Assim, em continuidade aos trabalhos realizados, este texto se instala como registro da construção do corpus *Summ-it* (Capítulo 2) e das análises previstas pelo projeto ProCaCoSa, precisamente, um exame dos sumários produzidos por um sumarizador automático, o

GistSumm, sob a ótica das cadeias de correferência, a fim de detectar problemas de coerência decorrentes da ausência de resolução de anafórica – Capítulo 3.

Visto que a construção do *Summ-it* envolveu, e ainda envolve, muitos especialistas, são elencadas no Anexo 1, em ordem cronológica, as tarefas que efetivamente foram realizadas pelo autor deste trabalho relacionando-as com as demais tarefas efetuadas pelo grupo de pesquisa. Além desse, é oferecido aqui outros anexos: as instruções de anotação de correferência, anáforas e referências dêiticas (Anexo 2), um relatório (exemplo) da anotação das cadeias de correferência (Anexo 3), uma descrição breve da anotação RST (Anexo 4) e as orientações ao sumarizadores profissionais (Anexo5).

O corpus *Summ-it* que é constituído por 50 textos do corpus PLN-BR busca servir de base a estudos avançados para diferentes tarefas do PLN, notadamente, à Resolução de Anáforas<sup>4</sup> e à Sumarização Automática. Corpora similares são disponibilizados pelo *Linguistic Data Consortium* - LDC<sup>5</sup>, tais como os corpora MUC-6 e ACE utilizados para o inglês.

O caráter inovador deste trabalho se encontra na fusão tecnológica no sentido da otimização das pesquisas em informação: um único corpus (o *Summ-it*) foi anotado com informações bibliográficas, lexicais, sintáticas, sobre as relações de correferência e anafóricas e sobre relações retóricas. Esses níveis de anotação foram preparados dentro dos padrões internacionais, que tem como palavras de ordem a *reusabilidade* (característica de um corpus ser usável em mais de um projeto de pesquisa e por mais de um grupo de pesquisadores) e a

---

<sup>4</sup> Subárea do PLN, a Resolução Anafórica, ou Resolução de Anáforas, compreende o desenvolvimento de aplicações computacionais que mapeiem automaticamente relações anafóricas, de correferência e casos de pronomes zero (elipses).

*extensibilidade* (isto é, a capacidade de corpora serem melhorados em várias direções, por exemplo, com a provisão de um nível a mais de análise lingüística ). Esse processo de anotação do corpus *Summ-it* é reportado no segundo capítulo deste trabalho. Nesse capítulo, também é relatado outro recurso desenvolvido para compor esse corpus: um conjunto de sumários. Para cada um dos 50 textos do *Summ-it*, foi gerado, por um sumarizador automático, um extrato e, por sumarizadores profissionais, foi elaborado um resumo informativo e um texto com marcações.

Entretanto, ao entender a delicadeza das informações lingüísticas envolvidas nas atividades aqui descritas, primeiramente, são apresentados, no próximo capítulo, os fundamentos que permeiam este trabalho.

---

<sup>5</sup> Disponível em: [http:// www ldc.upenn.edu/](http://www ldc.upenn.edu/)

# 1 FUNDAMENTAÇÃO TEÓRICA

De início, antes de montar um recorte teórico para reflexões tanto sobre questões lingüísticas quanto computacionais, pede-se, com reverência, licença aos inúmeros estudiosos das referidas comunidades científicas e seus respectivos aparatos teóricos, muitos, aliás, já consagrados. Sabe-se que, a partir de agora, o terreno é denso e acidentado, cheio de altos e baixos, e, nesse percurso, vários são os caminhos, perspectivas, e, neste trabalho, a escolha de um em detrimento dos demais não significa de forma alguma que esse se instala em superioridade àqueles. Acredita-se que as escolhas teóricas do presente trabalho tende muito mais aos aspectos computacionais adotados aqui. Além disso, conforme já mencionado, este trabalho busca atender uma série de atividades em andamento no âmbito das pesquisas a que ele está subordinado.

O presente estudo organiza a fundamentação teórica da seguinte forma: na Seção 1.1, é apresentada a concepção de texto (sentido e linguagem<sup>6</sup>) que permeia este trabalho; nas duas seções posteriores, concentra-se nas relações entre texto e coerência (Seção 1.2) e a idéia de coesão referencial (Seção 1.3); e, na última seção, são discutidos os tópicos relativos ao PLN, ou melhor, os referentes à Sumarização Automática.

---

<sup>6</sup> Os termos língua e linguagem são utilizados alternadamente ao longo deste trabalho sem uma distinção específica.

## 1.1 Concepção de texto

A preparação de uma representação do conteúdo de textos como atividade comum à redação de sumários está relacionada à concepção de texto e de sentido. O próprio conceito de texto depende das concepções que se tenha de língua e de sujeito.

Na concepção de língua como representação do pensamento e de sujeito como senhor absoluto de suas ações e de seu dizer, o texto é visto como um produto – lógico – do pensamento (representação mental) do produtor, nada mais cabendo ao leitor senão “captar” essa representação mental, juntamente com as intenções (psicológicas) do produtor, exercendo, pois, um papel essencialmente passivo.

Na concepção de língua como código – portanto, como mero instrumento de comunicação – e de sujeito como (pré)determinado pelo sistema, o texto é visto como simples produto da decodificação de um emissor a ser decodificado pelo leitor/ouvinte, bastando a este, para tanto o conhecimento do código, já que o texto, uma vez codificado, é totalmente explícito. Também nesta concepção o papel do “decodificador” é essencialmente passivo.

Já na concepção interacional (dialógica) da língua (KOCH, 2002; REITER e DALE, 2000), na qual os sujeitos são vistos como construtores, o texto passa a ser considerado o próprio lugar da interação e os interlocutores, como sujeitos ativos que – dialogicamente – nele se constroem e são construídos. Dessa forma há lugar, no texto, para toda uma gama de implícitos, dos mais variados tipos, somente detectáveis quando se tem, como pano de fundo, o contexto sociocognitivo dos participantes da interação.

Adotando essa última concepção, a compreensão passa a ser uma atividade interativa altamente complexa de produção de sentidos, que se realiza, evidentemente, com base nos



elementos lingüísticos presentes na superfície do texto e na sua organização, mas que requer a mobilização de um vasto conjunto de saberes e a sua reconstrução no interior do evento comunicativo.

A configuração veiculadora de sentidos se dá ao modo como os elementos presentes na superfície textual, aliados a todos os elementos do contexto sociocognitivo, são mobilizados na interlocução. O sentido é um dado a ser deduzido no código; modelo hermenêutico. Ele é um construto a ser engendrado no processo interativo, criado pelo sujeito-leitor, de acordo com as suas circunstâncias e os seus propósitos, sua bagagem, seus pontos de vistas etc. O sentido de um texto é, portanto, construído na interação texto e sujeitos.

Contudo, na atual conjuntura digital (realidade virtual, Internet, Inteligência Artificial etc.), cabe propor uma manutenção na noção de sujeito: na interação texto, há um novo participante, os programas computacionais. Hoje, não é novidade um sistema que, de modo autônomo, toma decisões e realiza tarefas a partir do processamento de textos. Aliás, atualmente, se observa uma verdadeira “corrida” tecnológica na busca de novos recursos que adquiram conhecimento por meio de bases textuais presentes na Internet – exemplo disso, é a famosa empresa Google<sup>7</sup>.

Todavia, o avanço dessas tecnologias está instalado sobre conceitos delicados, tais como, informatividade, coerência e coesão, noções intuitivas para os falantes da língua materna, mas de difícil incorporação e manipulação em sistemas computacionais. Esses conceitos de medidas nebulosas são o foco da atenção nas próximas duas seções.

---

<sup>7</sup> A empresa Google investe pesado em: PLN, IA, *Data Mining*, *Machine Learning*, *Information Retrieval* etc. – mais detalhes em: <http://labs.google.com/>

## 1.2 Coerência textual

Já há algum tempo, uma questão é amplamente debatida pela Lingüística Textual e pela Sumarização Automática: as relações entre texto e coerência. Pode-se dizer que tal discussão advém do momento em que se nota que os sentidos do texto não estão nele em si, mas estão sujeitos a fatores de diversas ordens: lingüísticos, cognitivos, socioculturais, interacionais etc.

Dentre vários trabalhos produzidos, este texto evoca três célebres aparatos teóricos: o de Robert A. Beaugrande e Wolfgang U. Dressler, o de Michael Halliday e Ruqaiya Hasan e o de Michel Charolles – que serviram de base aos estudos elaborados no país por Koch, Fávero, Travaglia entre outros.

Pode-se dizer que esses estudos sobre o texto não estão centrados em uma gramática, mas sim, na busca do que chamaram de “critérios de textualidade”. Segundo Kock e Travaglia (1989, p.21), “a textualidade ou a textura é aquilo que faz de uma seqüência lingüística um texto e não um amontoado de letras. A seqüência é percebida como texto quando aquele que a recebe é capaz de percebê-la como uma unidade significativa global”. Logo, se pode falar em textos coerentes, com sentido global, e textos incoerentes, com falta de sentido global. Beaugrande e Dressler (1981, p.24) dizem que um “texto incoerente é aquele em que o leitor ou ouvinte não consegue descobrir qualquer continuidade de sentido, seja pela discrepância entre os conhecimentos ativados, seja pela inadequação entre conhecimentos e o seu universo cognitivo”.

Entretanto, Charolles (1989) argumenta que não existem textos incoerentes, uma vez que não há regras de “boa formação de textos” que se apliquem a todas as situações possíveis. Conforme o teórico, tudo é muito relativo, dependendo muito dos interlocutores do texto e da situação em que esses estão circunscritos. Para iluminar essa questão de coerência textual, a partir da década de oitenta, Charolles (1989) instaura dois princípios: o de *cooperação* e de *interpretabilidade*. Em síntese, o primeiro estabelece que o leitor, no momento inicial com o texto, está potencialmente inclinado a tomá-lo como coerente. O segundo princípio, de interpretabilidade, coloca que, a priori, todos os textos seriam “aceitáveis”, desde que imersos no contexto adequado. Em linhas gerais, pode-se dizer que um texto pode ser incoerente em uma situação comunicativa, mas coerente em outra situação. Exemplos disso são os textos com algum aspecto lúdico, como os poemas. Como Charolles diz: “o texto será incoerente se seu produtor não souber adequá-lo à situação, levando em conta intenção comunicativa, objetivos, destinatário, regras socioculturais, outros elementos da situação, uso dos recursos lingüísticos etc. Caso contrário, será coerente” (p.43).

Esses dois princípios foram somados ao modelo teórico de Charolles (1978) que conta, ainda, com quatro metarregras de coerência: *repetição*, *progressão*, *não-contradição* e *relação*. Conforme a metarregra de repetição, um texto coerente deve apresentar, em seu desenvolvimento, “elementos de recorrência estrita”. Segundo a metarregra de progressão, um texto coerente deve oferecer, em seu desenvolvimento, uma “contribuição semântica constantemente renovada”. Essas duas regras, em linhas gerais, apontam para a necessidade de haver, em textos coerentes, retomadas de elementos já enunciados e, ao mesmo tempo, acréscimo de informação. Conforme o teórico, é essa costura de informações novas e dadas que permitem edificar textualmente a coerência. Geralmente, as retomadas se dão por

mecanismos de coesão referencial e, na progressão, desempenham uma função importante nos mecanismos de coesão seqüencial (KOCH, 1989). Em outras palavras, a coerência manifesta-se parcialmente no texto por mecanismos coesivos.

Outra metarregra é a de não-contradição. Segundo ela, para o texto ser coerente, “é preciso que no seu desenvolvimento não se introduza nenhum elemento semântico que contradiga um conteúdo posto ou pressuposto por uma ocorrência anterior, ou deduzível desta por inferência” (CHAROLLES, 1978, p.21). Já pela metarregra de relação o texto será coerente caso “os fatos que se denotam no mundo representado estejam relacionados”.

As metarregras de Charolles não dão conta, sozinhas, das questões referentes à coerência textual, fato reconhecido pelo próprio teórico. Por isso, a partir de agora, os aspectos relativos à coerência levantados até aqui são incrementados e formalizados em *fatores*. Aliás, é importante destacar que os sete critérios de textualidade apontados por Beaugrande e Dressler (1981), alguns centrados no texto e outros centrados no usuário, são tomados, aqui, como fatores de coerência, com alguns incrementos. Para os fins deste trabalho, abstrai-se a divisão entre propriedades centradas no texto e no usuário, assim como a noção de coerência adotada pelos teóricos – a ser entendida, não como uma propriedade entre outras, mas como resultado associado a múltiplos fatores, efeito de uma construção em situação de comunicação. Sublinhado isso, se passa à formalização de alguns fatores de coerência.

### 1.2.1 Fatores de coerência

O primeiro fator indubitavelmente é *o conhecimento lingüístico*. Porém, conforme dito anteriormente, é ilusão acreditar que a compreensão do significado de uma mensagem ocorre exclusivamente com base nas palavras e na sintaxe. Vários teóricos chamam a atenção para a relação do lingüístico com o cognitivo e com o pragmático.

Numa dimensão pragmática, pode-se apontar Grice (1975), que estabelece princípios conversacionais – aplicáveis a questões de coerência textual. Em sintonia com os demais trabalhos reportados, seu modelo estabelece um *Princípio da Cooperação*, “faça sua contribuição conversacional tal como é requerida no momento em que ocorre pelo propósito ou direção do intercâmbio em que está engajado”, que rege a comunicação humana e do qual derivam quatro máximas: i) *Máxima da Quantidade*, “Faça que sua contribuição seja tão informativa quanto for requerido para o propósito corrente da conversação; não a faça mais informativa do que o requerido”; ii) *Máxima da Qualidade*, “Não diga o que acredita ser falso; não diga senão aquilo para o que você possa fornecer evidência adequada”; iii) *Máxima da Relação*, “Seja relevante pertinente; e iv) *Máxima do Modo*, “Seja claro”.

No que tange ao conceitual-cognitivo, Beaugrande e Dressier dizem que há relações entre o nível gramatical e o conceitual do texto, sendo que “a cadeia gramatical somente se estende por pequenas partes do texto, enquanto a cadeia conceitual abrange o texto todo” (1981, p. 31). Beaugrande (1980) aponta e exemplifica algumas correlações entre os níveis gramatical e conceitual. Por exemplo, uma estrutura que relaciona, no nível gramatical, um sujeito com um verbo, no nível conceitual, considera-se o sujeito como um agente se o verbo

for de ação ou um objeto se o verbo for de estado, como nas sentenças: *O dinossauro se alimentava de animais maiores que ele* e *O dente foi encontrado nas últimas escavações*<sup>8</sup>.

Prince (1981) e Yule (1981) chamam a atenção para a relação das formas lingüísticas com a estrutura informacional-cognitiva, mais uma função do conhecimento lingüístico no estabelecimento da coerência. Operações complexas, como as inferências, são possíveis graças a essas relações. Aliás, as *inferências* ganharam, por alguns autores (por exemplo, KOCH 1989; FÁVERO 2000) o *status* de fator para questões de coerência.

Classicamente, a inferência é vista como uma relação entre duas idéias do discurso. Beaugrande e Dressier (1981) dizem que inferência é a operação que consiste em fornecer conceitos e relações plausíveis para preencher lacunas em “um mundo textual” – modelo de mundo representado em cada texto. Segundo eles, uma inferência visa freqüentemente resolver um problema de continuidade de sentido. Conforme Brown e Yule (1983), inferências são conexões que as pessoas fazem quando tentam alcançar uma interpretação do que lêem.

Charolles (1983) afirma que a atividade de interpretação é regida pelo princípio da coerência, que orienta aquele que interpreta o texto a estabelecer relações que não estão expressas na superfície do texto: elas são as inferências que podem ser ou não lingüisticamente fundadas. É interessante verificar que teóricos, como Charolles, separam as inferências lingüisticamente fundadas das não lingüisticamente fundadas. Nessa tarefa, alguns

---

<sup>8</sup> Trechos extraídos do texto CIENCIA\_2005\_6825 corpus PLN-BR.

sobrecarregam o léxico de seus modelos com vistas de reter o máximo de inferências dentro do domínio lingüístico. Contudo, é possível perceber que as determinações lingüísticas têm cedido a outras determinações, como as psicológicas.

Beaugrande e Dressier (1981, p.102) apresentam objeções ao uso das inferências na explicação do processo de compreensão de textos ou como parte do modelo que representaria esse processo por duas razões: primeiro, porque “as inferências admitidas neste processo seriam escolhidas arbitrariamente” e, segundo, porque “as inferências admitidas são poucas, uma vez que os usuários podem fazer muitas outras”. Na maioria das vezes, é possível realizar muitas inferências a partir das instruções lingüísticas. Há textos que visam proporcionar muitos veios de inferências, tais como textos humorísticos, publicitários e literários.

Em grande parte, as inferências dependem da noção de *conhecimento de mundo*. Esse conceito pode ser ilustrado como um tipo de enciclopédia do mundo e da cultura arquivado na memória. A memória humana pode ser compreendida como a capacidade dos seres humanos adquirir, conservar e evocar informações através de dispositivos neurobiológicos e da interação social. Cummings e Bensonf (2005, p.85) explicam que a memória é “um conjunto de procedimentos que permite manipular e compreender o mundo, levando em conta o contexto atual e as experiências individuais, recriando esse mundo por meio de ações da imaginação”.

Existem vários tipos de memórias que se relacionam para formar “a memória” que usamos no dia-a-dia. Os principais sistemas de memória reconhecidos pela neurologia atual são a *memória sensorial*, a *memória operacional* e a *memória de longa duração* (ALLEGRI, HARRIS e DRAKE, 2006; BRUN et al., 2006; HODGES, SALMON e BUTTERS, 2006).

Essa última divide-se ainda em *memória declarativa* (subdividida em *memória episódica* e *memória semântica*) e *memória procedimental*. Essa é a capacidade de reter e processar informações que não podem ser verbalizadas, como tocar um instrumento musical ou dirigir um veículo. Já a memória declarativa é a capacidade de verbalizar uma experiência anterior. Esta pode ser subdividida em memória episódica e memória semântica. A primeira é os registros de experiências de vida pessoais e eventos. A informação em memória episódica é associada com um lugar e/ou tempo particular. Na segunda, a memória semântica, as informações não estão associadas com um tempo particular ou lugar. Essa inclui principalmente o conhecimento sobre palavras, idiomas, e símbolos; seus significados, as relações entre eles; e regras para usá-los e manipulá-los.

Todo esse conhecimento não está organizado de modo caótico, ele geralmente se organiza em blocos. Esse conhecimento em blocos pode ser dividido: *conhecimento enciclopédico* (*background knowledge*), que representa tudo o que se conhece e que está arquivado na memória de longa duração; e *conhecimento ativado* (*foreground knowledge*), que é trazido à memória operacional.

A esses se acrescentam os modelos cognitivos globais, blocos de conhecimentos referentes aos conceitos intensamente utilizados na interação humana. Esses são composições cognitivas de conhecimento consolidadas pelo uso e encubadas na memória permanente.

Entre os modelos cognitivos globais, os *frames*, *esquemas*, *planos* e *scripts* vêm sendo adotados pela lingüística no processamento cognitivo dos textos. Ao lado deles, aparecem os cenários e modelos mentais. Alguns desses modelos foram propostos pelos estudos de Inteligência Artificial (*frames*, *scripts*), outros pela psicologia da cognição (cenários,



esquemas, modelos mentais). É importante observar que há flutuação terminológica, de modo que um mesmo conceito pode ser empregado com nomes diferentes. Há, também, estudos que utilizam uma única denominação para todos os tipos de modelos cognitivos (teoria dos *frames*, teoria dos esquemas etc.). Por isso, se adota a proposta de Beaugrande e Dressier (1981) com incrementos:

*frames* – conjunto de informações conservadas na memória sob uma espécie de “rótulo”, sem que haja uma ordenação entre elas, por exemplo: jantar (noite, talheres, sopa, vinho etc.), praia (mar, sol, calor, surf, bronzado, filtro solar etc.);

*cenários* – quadro visual (por vezes, enriquecido por informações audíveis e olfativas) impressa na memória que registra um lugar em que decorreu algo, por exemplo, imagem de um teatro (poltronas, palco, cortinas etc.) ou de uma sala de aula (lousa, cadeias, mesas etc.);

*esquemas* – conjunto de informações guardadas na memória em seqüência temporal ou causal, por exemplo: um dia de mãe (observa a criança, amamenta-a, troca a fralda, dá banho etc.);

*planos* – grupo de informações armazenadas, em seqüência, na memória sobre como agir para atingir certo objetivo, por exemplo, instalação de um televisor (abrir a embalagem, ler o manual, conferir os acessórios, ligar a alimentação elétrica etc.);

*scripts* – conjunto de informações registradas, em seqüência, na memória sobre modos de agir altamente estereotipados em uma cultura, até mesmo em termos de linguagem, por exemplo, cerimônias religiosas, praxes jurídicas, apresentações formais.

Aos itens elencados anteriormente, é importante acrescentar as macroestruturas ou superestruturas. A noção de *macroestrutura* foi proposta por Van Dijk (1981) para a interpretação coerente de um texto. Ela pode ser entendida como uma estrutura profunda semântica do texto, que é representada por uma macroproposição alcançada por meio de macrorregras que abstraem o conteúdo proposicional de seqüências textuais e organizam o conteúdo em termos de hierarquização. A macroestrutura é definida no nível da representação semântica global do texto. Ela pode ser comparada a uma espécie de sistema psicológico que norteia o planejamento, execução, compreensão, armazenamento e reprodução do texto. Pode-se dizer que determinar a macroestrutura de um texto é estabelecer sua coerência, pelo menos em termos semânticos. O enunciado que expressa a macroestrutura é chamado também de *macroproposição textual*.

A partir desses cinco modelos cognitivos globais, entende-se que o “mundo textual” nunca será uma cópia fiel do mundo real, uma vez que o produtor do texto recria o mundo sob um ponto de vista – carregado de crenças, desejos e intenções. Entretanto, a instauração da coerência de um texto exige certa correspondência entre os conhecimentos nele ativados e o conhecimento de mundo do leitor. Esse conjunto de informações incomuns (produtor – texto – leitor) que inclui mais do que o domínio de regras gramaticais é denominado *conhecimento compartilhado* (*common ground*). A postulação básica dessa noção é que, entre quaisquer pessoas, há sempre uma porção de conhecimentos partilhados que podem ser utilizados como pano de fundo da comunicação verbal.

Intimamente relacionado às noções de conhecimento de mundo e conhecimento compartilhado está a concepção de *intertextualidade*. Segundo Beaugrande e Dressier (1981),

a intertextualidade diz respeito aos fatores que tornam uma compreensão plena de um texto dependente de um ou mais textos previamente existentes. Essas maneiras podem incluir fatores relativos a conteúdo, aspectos formais e aspectos ligados a tipos textuais (descritivo, injuntivo, narrativo, dissertativo etc.).

Os fatores ligados a conteúdo são bastante evidentes e se ligam a questões de conhecimento de mundo. Para ilustrar, pode-se tomar artigos jornalísticos que cobrem um mesmo acontecimento, durante dias. Cada artigo pressupõe que os leitores conheçam as matérias sobre o mesmo assunto publicado anteriormente.

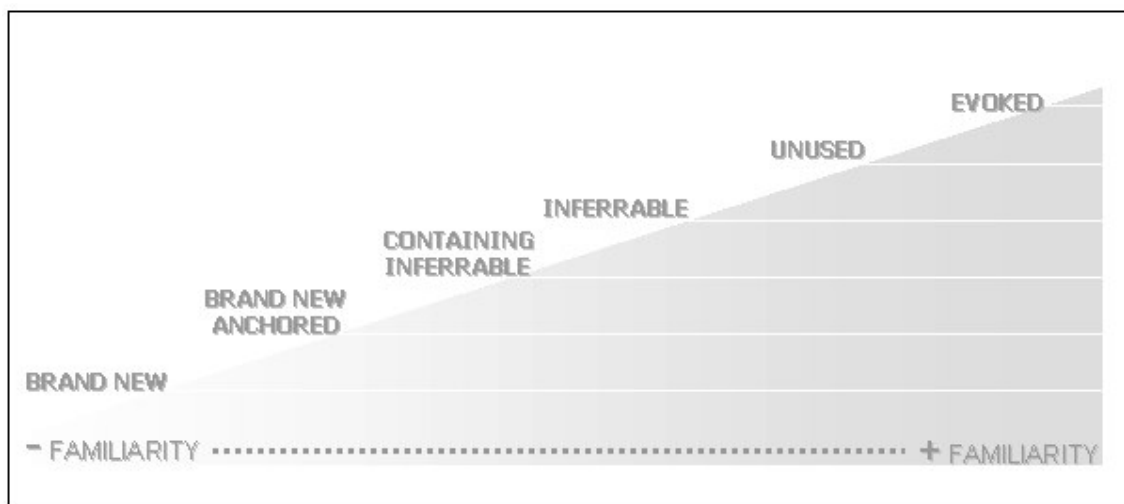
A intertextualidade é posta em relação aos aspectos formais e aos aspectos ligados aos tipos textuais na medida em que um texto somente pode ser bem interpretado, visto como coerente, se for esse compreendido como um exemplar dentre uma família – grupo de textos que apresentam propriedades estruturais e lexicais em comuns. No que se refere à estrutura de cada espécie de texto, é fundamental retomar a noção de *superestrutura*, que Van Dijk define como estruturas globais características de certos discursos, esquemas discursivos provenientes de “um aprendizado intuitivo ou sistematicamente dirigido, conhecimentos convencionais que envolvem além de uma seqüência esquemática, características de linguagem, de recursos retóricos ou estilísticos etc; as superestruturas são culturalmente dependentes” (1992, p.28).

Pode-se entender a partir disso que as superestruturas estabelecem uma correspondência com intertextualidade tipológica, uma vez que são internalizadas nas práticas sociais. Sua natureza é, parcialmente, diferente dos modelos cognitivos globais já mencionados, já que os usuários têm dificuldade em explicitá-las, sem um aprendizado sistemático, diferentemente do que ocorre com os *frames*, esquemas, planos e *scripts* –

geralmente um indivíduo não tem grandes dificuldades de falar sobre “o que dever ser feito para comprar um refrigerante?”.

Outro fator também relacionado ao conhecimento de mundo e o conhecimento compartilhado, relevante ao toca questões de coerência textual, é a *informatividade*. Essa noção tem a ver com a previsibilidade, redundância, das informações apresentadas pelo texto (BEAUGRANDE e DRESSIER, 1981). A informatividade é inversamente proporcional aos pontos antevistos, antecipados, pelo leitor, ou seja, quanto maior forem as coisas antevistas pelo leitor (estruturas lingüísticas, estratégias retóricas, fatos, objetos, hipóteses etc.), menor será a informatividade do texto. O grau de informatividade pode influenciar na coerência. Por exemplo, se, as informações de um texto forem extremamente inesperadas, novas, provavelmente esse texto possa ser considerado incoerente, ou pelo menos hermético.

Tradicionalmente, se considera *nova* a informação que o texto apresenta como não sendo recuperável nos trechos precedentes e *dada* aquela recuperada de alguma forma nas passagens anteriores (HALLIDAY e HASAN, 1976). A noção de *dado* e *novo* tem apresentado flutuações terminológicas. Prince (1981, 1992) discute os modelos binários e elabora uma nova proposta. Ela, ao entender o texto como um conjunto de instruções sobre como edificar um discurso particular, apresentando entidades, atributos e relacionamentos entre entidades, propõe uma escala de *familiaridade assumida* (*assumed familiarity*), conforme pode ser observado na Figura 2.



**Figura 2 – Escala de familiaridade assumida (*assumed familiarity*) proposta por Prince**

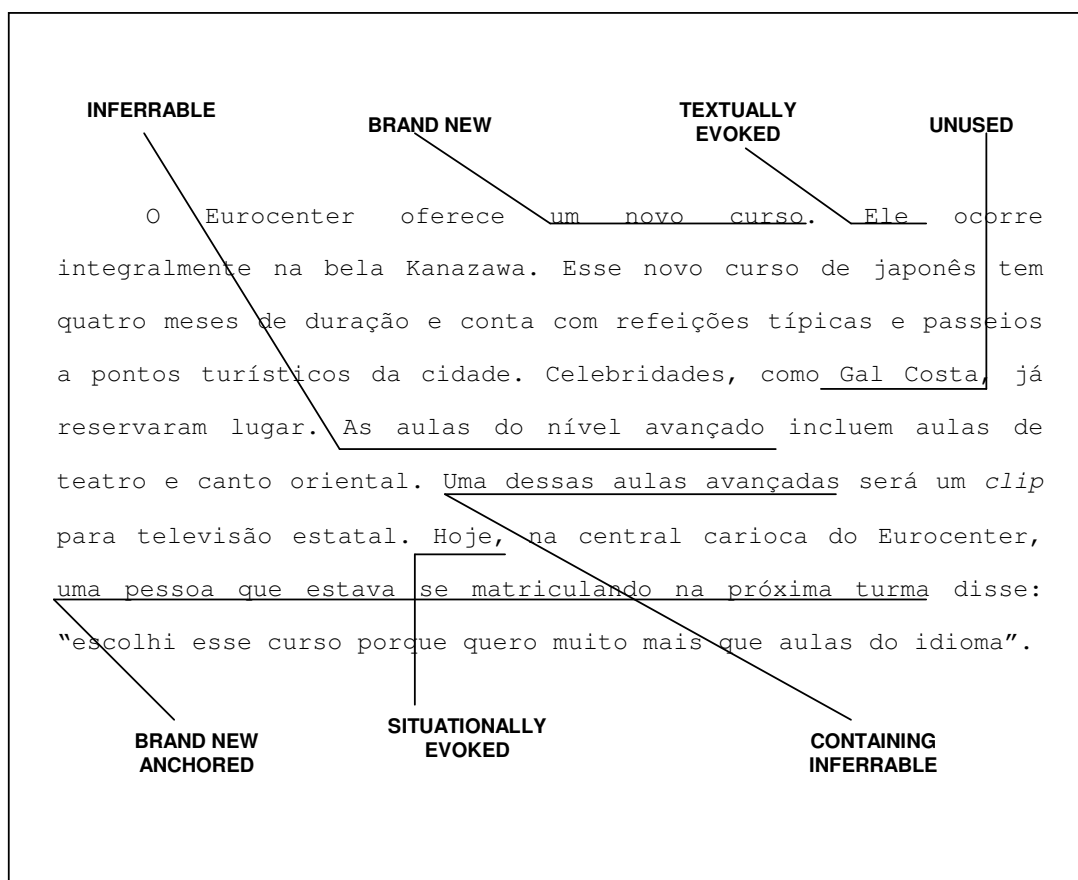
Seu trabalho oferece uma taxonomia com tipos de entidades. Nova é a entidade que está sendo introduzida no discurso pela primeira vez e, segundo Prince, pode ser de dois tipos: *brand new*, quando o produtor precisa “criá-la” a partir do texto, ou *unused*, quando se supõe que ela já é familiar ao leitor – exemplos na Figura 3.

Uma entidade *brand new*, ainda, pode ser *brand new anchored* ou exclusivamente *brand new*. Ao contrário dessa, as entidades *brand new anchored* “apóiam-se” semanticamente em alguma outra entidade por meio de um sintagma nominal contido no sintagma nominal que introduz a entidade – exemplos na Figura 3.

Além dessas, há as entidades do tipo *inferrable* que podem ser do tipo: *containing inferrable* ou exclusivamente *inferrable*. Essa é o tipo de entidade que pode ser inferida pelo leitor por meio de outras entidades já evocadas. Uma *containing inferrable* é aquela em que a entidade a partir da qual a inferência é feita é representada por um item lexical contido dentro

do sintagma nominal que introduz a entidade classificada como *containing inferrable* – exemplos na Figura 3.

Uma informação dada num modelo binário é classificada, por Prince, como *evoked*. Essa pode ser de dois tipos: *textually evoked*, aquela que pode ser recuperada no texto, e *situationally evoked*, entidade que é recuperada na situacionalidade – isto é, os casos de *dêixis*<sup>9</sup> – exemplos na Figura 3.



**Figura 3 – Exemplificação da taxonomia de Prince**<sup>10</sup>

<sup>9</sup> Segundo Prince (1981, 1992), a *dêixis* é o fenômeno da linguagem humana que consiste em fazer um enunciado referir-se a uma situação que pode ser quanto à enunciação, quanto ao momento da enunciação ou quanto ao lugar onde ocorre a enunciação, estado ou processo.

<sup>10</sup> O trecho foi extraído do texto 23 do corpus 2 do projeto TeXto.

Essa escala de familiaridade funciona dentro do seguinte princípio: o produtor evita introduzir novas entidades, quando as velhas são suficientes para o propósito comunicativo. Prince acredita que isso leve a um fenômeno possivelmente válido para o discurso: a tendência de empregar sintagmas nominais que sejam tão altos na escala de familiaridade quanto possível. Sua proposta tem por base a relação entre conhecimento de mundo e conhecimento compartilhado.

Com efeito, a última classe mencionada do modelo Prince, *situationally evoked*, chama a atenção para mais um fator importante no que tange a coerência textual: a *situacionalidade*. Bastos (1985) afirma que a coerência se engendra ao se inserir o texto numa determinada situação comunicativa. Desse modo, é possível dizer que caso a condição de situacionalidade não ocorra, o texto pode se apresentar como incoerente. Como dito anteriormente, foi a não definição de uma situação comunicativa adequada ou possível que, na maioria das vezes, levou teóricos a considerarem certos textos como incoerentes.

Soma-se ao fator da situacionalidade, a *focalização* – outro fator importante nos estudos de coerência textual. Ela pode ser entendida como os pontos de vista pelos quais as entidades evocadas no texto passam a ser vistas. Segundo Grosz (1981), os interlocutores focalizam sua atenção em pequena parte do que sabem e acreditam, e a enfatizam. Assim, algumas entidades além de assumirem uma posição central no discurso também são selecionadas e percebidas através de perspectivas que afetam tanto a fala do produtor quanto a interpretação do leitor. Por exemplo, num texto, certa perspectiva pode evocar um diálogo, como uma conversa ou uma discussão – quem sabe até mesmo como uma briga. Em conformidade com Grosz, pode-se dizer que, entre a língua e a focalização, há um caminho de

duas vias: o que é dito influencia a focalização e vice-versa. As pistas que o produtor “fornece” ao leitor sobre o que está focalizando podem ser lingüísticas e não-lingüísticas (por exemplo, gráfico, tabelas, fotografias etc.). O teórico defende que a focalização não só torna a comunicação mais eficiente, como, na verdade, a torna possível. Evidentemente, a focalização afeta a capacidade e a possibilidade do leitor estabelecer a coerência de um texto.

Estreitamente, ligados à focalização, estão outros dois fatores importantes para coerência comentados anteriormente, mas não formalizados: a *intencionalidade* e a *aceitabilidade*. Numa ponta está a intencionalidade, tratando os aspectos que dizem respeito à intenção do produtor. Na outra ponta está a aceitabilidade, representando a atitude do leitor de aceitar o conjunto de sentenças como um texto coesivo e coerente, que passa a ele algum proveito ou relevância. Quando Brown e Yule (1983), Charolles (1987a) e Grosz (1981) argumentam sobre o princípio de cooperação por parte dos interlocutores – um busca tecer um texto que faça sentido e o outro busca ver a produção como algo que foi concebido para fazer sentido, agindo em função disso – eles estão, de certa forma, chamando a atenção para a intencionalidade e a aceitabilidade.

Esses teóricos afirmam que a intencionalidade compreende todas as maneiras como os interlocutores usam os textos para alcançar seus objetivos comunicativos, enquanto a aceitabilidade inclui a aceitação como disposição ativa de participar de um discurso e compartilhar um propósito comunicativo. Em sentido amplo, essas duas noções relacionam-se com o que vem sendo chamado pela lingüística como argumentatividade, ou melhor, subjacente aos aspectos cognitivos do uso lingüístico, existe uma atividade básica: a *argumentação*. Koch (1984, p.12) diz que a atividade de interpretação fundamenta-se,



exatamente, na convicção de que quem produz um texto tem “determinadas intenções, consistindo a intelecção na captação dessas intenções, o que leva a prever, por conseguinte, uma pluralidade de interpretações”.

Então é possível compreender a intencionalidade e aceitabilidade como “faces” do princípio de cooperação, princípio que se fundamenta na idéia de que o texto faz sentido, é coerente e se faz de tudo para buscar esse sentido.

Intimamente associado à intencionalidade e à aceitabilidade está o que Giora (1985) define como a *consistência* e a *relevância*. Ela defende que esses dois fatores são básicos para que um texto possa ser considerado coerente. Os fatores da consistência e da relevância estão relacionados com as metarregras de não-contradição e de relação (CHAROLLES, 1978) ou com as Máximas da Qualidade e da Relação (GRICE, 1975), uma vez que se pode dizer que um texto será coerente quando seus enunciados tratam, em linhas gerais, de um mesmo tópico discursivo sem serem contraditórios. Em harmonia com esses teóricos, a estudiosa caracteriza a consistência como a propriedade de cada enunciado, no mundo textual, poder ser considerado não-contraditório em relação aos demais. Quanto à relevância Giora diz que “um conjunto de enunciados será relevante para um tópico discursivo se eles forem interpretáveis como predicando algo sobre um mesmo tema” (p.28). Desse modo, se pode dizer que a relevância não é uma propriedade entre pares de enunciados, mas entre os enunciados e um tópico discursivo, ou melhor, um supertópico discursivo. Portanto, para que vários segmentos textuais com diferentes tópicos discursivos possam preencher o requisito de relevância, eles devem estar ligados por um supertópico discursivo subjacente em termos de *aboutness* (ser sobre algo).

Não foi possível discutir todos os fatores de coerência que contribuem para a construção do sentido global do texto. Também não foi esse o objetivo deste trabalho. No entanto, foi possível delinear alguns fatores e uma noção de coerência textual que serve de base para observação do processo de sumarização (manual e automático). Como o foco deste trabalho é voltado para a relação entre os processos de correferência e a sumarização automática, na próxima seção, é abordada, em particular, a coesão referencial.

### **1.3 Coesão referencial**

De acordo com a discussão apresentada na seção anterior, é possível perceber que o estabelecimento da coerência depende dos elementos lingüísticos (seu conhecimento em uso), assim como, evidentemente, da sua organização no contexto lingüístico. Essa tessitura tradicionalmente denominada de coesão textual, embora não suficiente, é condição básica para um texto. Apontam para isso, as metarregras de repetição e de progressão de Charolles (1978) e a proposta de textualidade de Halliday e Hasan (1976).

Reconhecendo a variedade de processos tratados pela coesão textual – a anáfora (pronominal, nominal, possessiva etc.), a dêixis, o uso dos artigos, as marcas de temporalidade, a elipse, as modalidades, a subordinação e a coordenação, a ordem de palavras, o componente lexical, a recuperação pressuposicional, a tematização (tema-remática, tópico-comentário) etc. – é focalizado, aqui, a coesão referencial. Essa se detém, especialmente, nas estratégias de designação, examinando as relações de identidade e/ou de dependência entre as entidades evocadas (relações de sentido). Mais especificamente, é focalizado o estudo da coesão referencial com vistas na resolução automática de correferência e anáforas. Por isso se adota como base os trabalhos anteriormente realizados pelas equipes do Laboratório de

Engenharia da Linguagem. Esses têm por objetivo a recuperação das cadeias de correferência como parte de um objetivo maior de Extração (automática) de Informação de bases textuais. Apesar desse objetivo específico, os fenômenos lingüísticos a serem compreendidos nesses trabalhos estão fortemente ligados à noção de coesão referencial, por isso a preocupação com o tema. Essas pesquisas têm em sua gênese o trabalho de Vieira (1998), que contempla os processos de correferência para o caso específico das descrições definidas<sup>11</sup> da língua inglesa a partir de um minucioso estudo comparativo das propostas de Russell (1919), Clark (1977), Hawkins (1978), Sidner (1979), Prince (1981, 1992), Löbner (1985), Fraurud (1990) e Strand (1996). Posteriormente, esse trabalho foi atualizado pelas investigações realizadas nos projetos de pesquisa ANACORT - *Anotação automática de co-referência textual*<sup>12</sup>, COMMON-REFs - *Um modelo computacional unificado para o tratamento de referências*<sup>13</sup> e TeXto - *Acesso a informações em bases textuais*<sup>14</sup>, que consideram a língua portuguesa. Outra base relevante é a serie de estudos do projeto VENEX - Projeto colaborativo entre a *Universita' di Venezia* e a *Essex University*<sup>15</sup>, que envolve questões relacionadas à língua italiana.

Esses estudos estabelecem que o discurso pode ser representado como um sistema contendo entidades (*discourse entities*) e relações entre essas entidades. As entidades podem figurar como seres, lugares, fatos, idéias etc. Notadamente, as instruções lingüísticas que disparam e atualizam essas entidades do discurso são as expressões referenciais realizadas por sintagmas nominais. A organização dessas expressões está intimamente ligada aos fatores

---

<sup>11</sup> Considera-se uma descrição definida aquele sintagma nominal iniciado por determinante artigo definido.

<sup>12</sup> Disponível em: [http://www.inf.unisinos.br/~renata/laboratorio/anacort\\_index.htm](http://www.inf.unisinos.br/~renata/laboratorio/anacort_index.htm)

<sup>13</sup> Disponível em: [http://www.inf.unisinos.br/~renata/laboratorio/common\\_index.htm](http://www.inf.unisinos.br/~renata/laboratorio/common_index.htm)

<sup>14</sup> Disponível em: [http://www.inf.unisinos.br/~renata/laboratorio/texto\\_index.htm](http://www.inf.unisinos.br/~renata/laboratorio/texto_index.htm)

<sup>15</sup> Disponível em: <http://cswww.essex.ac.uk/staff/poesio/>

anteriormente abordados – intencionalidade e aceitabilidade, consistência e relevância, intertextualidade, focalização, informatividade, conhecimento lingüístico, compartilhado e de mundo – pois o texto, além de ser influenciado por aspectos cognitivos, objetiva ser coesivo, uma vez que se realiza “com base no *já dito*, no que *será dito* e no que é *sugerido*” (KOCH, 2002, p.85). Nessas operações, é possível observar que certos termos e/ou expressões apresentam uma relação de identidade com outros termos e/ou expressões do próprio texto, constituindo, assim, cadeias de correferência. Em (1), pode-se notar que as expressões *um novo curso*, *ele*, *esse novo curso de japonês* e *esse curso* manifestam uma relação de identidade, ou melhor, elas engendram uma mesma entidade, por isso, é dito que essas são expressões correferentes, logo, compõem uma cadeia de correferência.

( 1 ) O Eurocenter oferece um novo curso. Ele ocorre integralmente na bela Kanazawa. Esse novo curso de japonês tem quatro meses de duração e conta com refeições típicas e passeios a pontos turísticos da cidade. Celebrities, como Gal Costa, já reservaram lugar. As aulas do nível avançado incluem aulas de teatro e canto oriental. Uma dessas aulas avançadas será um *clip* para televisão estatal. Hoje, na central carioca do Eurocenter, uma pessoa que estava se matriculando na próxima turma disse: “escolhi esse curso porque quero muito mais que aulas do idioma”.<sup>16</sup>

Uma entidade engendrada por uma cadeia de correferência ao ser atualizada também é mantida em foco. Por isso, esse processo não deve ser tomado como algo estático e estável, ao contrário, são procedimentos complexos intimamente ligados aos fatores da seção anterior. Para ilustrar, em (1), é possível perceber que a entidade promovida pela expressão *um novo curso* é incrementada pela expressão correferente *esse novo curso de japonês*,

---

<sup>16</sup> O primeiro parágrafo do texto 23 do corpus TeXto.

especificamente, pelo modificador *japonês*. Esse tipo de atualização, por se dar entre expressões correferentes que apresentam nomes núcleos idênticos, é chamado de correferência direta. São incluídas nessa noção também as ocorrências de repetição exatas, tais como *O Eurocenter* e *o Eurocenter*. Relacionamentos diretos, além de serem a forma mais comum de focalização (saliência) também são uma das maneiras mais seguras de evitar ambigüidades relativas à coesão referencial.

Geralmente, se detecta uma situação de ambigüidade quando mais de um candidato se apresenta para assumir uma relação de correferência, sendo que cada um deles evoca uma entidade diferente. Em (2), nota-se um exemplo de ambigüidade. Para a expressão *o famoso especialista em fotocinética*, apresentam-se como candidatos (para correferência), as expressões *Olacio Dietzsch*, *Eduardo Macchione*, *Kiyomi Koide* e *José Hirata*. Em linhas gerais, isso ocorre, porque essas cinco expressões atendem tanto a critérios relativos à flexão genérica e numérica, quanto aos critérios relativos a questões de saliência. Um critério que pode desfazer essa ambigüidade é o conhecimento de mundo – *Kiyomi Koide* é um reconhecido pesquisador em fotocinética.

Embora a noção de saliência não resolva esse caso, ela é um instrumento útil em situações de ambigüidade, uma vez que permite uma classificação, de graduação crescente ou decrescente, das entidades do discurso segundo uma escala de grandeza. Uma entidade saliente é aquela em foco, atualizada, retomada, por outras expressões. Em (1), a entidade “o novo curso de japonês do Eurocenter” é mais saliente do que “a bela cidade de Kanazawa”. Para desfazer uma ambigüidade é possível, também, verificar qual expressão está mais próxima ou qual a função sintática dos candidatos. Em adição, emprega-se a idéia de papéis

temáticos (agente, instrumento etc.). Esses aspectos também estão relacionados à saliência. As situações de ambigüidades estão, consideravelmente, ligadas aos fatores de informatividade, de focalização, de conhecimento compartilhado e de conhecimento de mundo, por isso, são recursos relativos a esses fatores, os mais eficientes.

(2) Os físicos gostam de dizer que a biologia estaria perdida sem eles. Um grupo do Instituto de Física da USP está ajudando a inflar ainda mais o ego da categoria, ao construir o primeiro aparelho no país que usa laser para analisar moléculas biológicas.

O instrumento tem um nome indecifrável: espectrômetro de massa de ionização por dessorção a laser com auxílio de matriz (Maldi, na sigla em inglês). Mas sua função pode ser definida de um jeito bem simples: uma balança para pesar proteínas.

O aparelho de Maldi foi projetado e construído pelo grupo do LIP (Laboratório de Instrumentação e Partículas), do qual fazem parte os pesquisadores Olacio Dietzsch, Eduardo Macchione, Kiyomi Koide e José Hirata.

O famoso especialista em fotocinética adiantou que o instrumento está sendo testado na análise de polímeros (longas cadeias de uma mesma molécula) e, dentro de seis meses, deverá ser usado rotineiramente com proteínas.<sup>17</sup>

O exemplo anterior, que ilustra uma situação de ambigüidade, também exemplifica outra forma de atualização, de retomada: a correferência indireta. Essa compreende aqueles casos em que as expressões correferentes apresentam nomes núcleos diferentes. Outro exemplo de relacionamento indireto é: *o primeiro aparelho no país que usa laser para analisar moléculas biológicas* e *o instrumento* – em (2). Uma relação de correferência indireta pode ser motivada por várias questões: estéticas, temáticas, argumentativas etc. Por exemplo, no que toca aspectos temáticos, pode-se dizer que uma retomada indireta pode promover uma

---

<sup>17</sup> Os quatro primeiros parágrafos do texto CIÊNCIA\_2002\_22034 do corpus *Summ-it*.

dimensão nova de desenvolvimento informativo, tal como, em (3), na relação *um novo dinossauro – o animal* e *um novo dinossauro – o fóssil*. A primeira abriu espaço para asserções sobre o ser vivo (seus hábitos), enquanto a segunda reporta informações sobre a criatura não viva (seu estado de conservação e posição na hierarquia paleontológica).

( 3 ) Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de um novo dinossauro no Brasil. O animal era carnívoro e habitava o nordeste brasileiro há 110 milhões de anos, no período Cretáceo. A estrutura óssea desse raptor permitia a ele ser ágil e veloz, provavelmente se alimentando de presas um pouco menores que ele.

O fóssil, batizado de *Santanaraptor placidus raptor*, é único, pois foi encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele. Com os tecidos preservados, os cientistas esperam poder saber mais sobre o modo de vida e a evolução dos répteis. O *Santanaraptor* pode ocupar uma posição no grupo *Tyrannoraptora*, o mesmo do *Tyrannosaurus rex*.<sup>18</sup>

A correferência indireta é um processo complexo e fértil da linguagem natural. Nele é possível identificar liames que orbitam entre as relações de sinonímia, hiponímia e hiperonímia. Em sintonia com as concepções de texto e sentido assumidas aqui, sinonímia é entendida como a relação de sentido entre dois vocábulos que têm significação muito próxima. Não se entende sinonímia como uma relação de equivalência, pois essa, rigorosamente, é reflexiva, simétrica e transitiva. São compreendidas como sinônimas todas as operações lexicais de ordem paradigmáticas que não modificam a entidade (que restituem à sua identidade). Por exemplo, o vocábulo *casa* está numa relação de sinonímia com *moradia*, assim como, *aparelho* e *instrumento* – em (2). À medida que a relação de sentido entre dois

---

<sup>18</sup> Os dois primeiros parágrafos do texto CIÊNCIA\_2000\_17088 do corpus *Summ-it*.

vocábulos se distancia, pode ocorrer uma relação de hiponímia, ou seja, aquela existente entre uma palavra de sentido mais específico e outra de sentido mais abrangente, que tem com a primeira propriedades em comuns. Por exemplo, *chocolate* (específico) está numa relação de hiponímia com *doce* (genérico), assim como, *um novo dinossauro* e *o animal* – em (3). A relação inversa à hiponímia é denominada de hiperonímia, isto é, aquela estabelecida entre um vocábulo de sentido mais genérico e outro de sentido mais específico. Por exemplo, *doce* (genérico) está numa relação de hiperonímia com *chocolate* (específico), assim como, *um novo dinossauro* e *esse raptor* ou *o animal* e *esse raptor* – em (3).

A correferência indireta aponta para outra noção importante neste estudo: a anáfora. Ela, em certa medida, está relacionada ao conceito de correferência, já que, em sentido amplo, a anáfora se define como toda retomada de uma expressão em um texto, mantendo-se a relação de identidade – como em *esse novo curso de japonês* e *esse curso* ou em *a bela Kanazawa* e *a cidade* – em (1). Ao contrário da correferência, em relacionamentos anafóricos, há uma idéia de direção, ou melhor, de movimento para trás (retomadas, remissões), por isso, a noção de antecedente é funcional no fenômeno anafórico. Ela estabelece, no relacionamento binário, qual termo retoma (expressão anafórica) e qual é retomado (expressão antecedente). Em anáforas, de forma análoga à correferência, também são utilizadas as designações de relação direta e indireta, ou seja, anáfora direta (*esse novo curso de japonês* – *esse curso*) e anáfora indireta (*a bela Kanazawa* – *a cidade*).

Entretanto, em sentido restrito, o conceito de anáfora está embutida a idéia de dependência. Exemplos disso são as relações: *um novo curso* – *ele* (em 1), *os físicos* – *eles* (em 2) e *um novo dinossauro* – *ele* (em 3). Aqui se observa, explicitamente, que os dois



pronomes pessoais não possuem interpretação própria, essa é recebida de uma outra expressão nominal antecedente – *um novo curso*, *os físicos* e *um novo dinossauro* respectivamente. Nota-se que esse tipo de operação pode ocorrer tanto com entidades de caráter genérico (*os físicos* e *eles*) quanto com entidades de caráter específico (*um novo curso* e *ele*). De forma análoga, porém menos explícita, é possível verificar, também, a idéia de dependência em versões reduzidas das instruções lingüísticas que retomam a entidade já introduzida de modo mais completo (em termos lexicais). Exemplo disso são: *esse novo curso de japonês* (versão estendida) e *esse curso* (versão reduzida) ou *a bela Kanazawa* (versão estendida) e *a cidade* (versão reduzida) – em (1). Em (3), também é possível examinar outros exemplos: *um novo dinossauro* e *o animal* ou *o fóssil*.

Essa noção de dependência pode ser ampliada, caso se entenda que uma entidade é resultado de sua cadeia de correferência, que se alimenta das informações disparadas pelas sentenças em que seus nodos de interpretação estão presentes. Por exemplo, na sentença *o Eurocenter oferece um novo curso* (em 4), a entidade “curso” pode ser sintetizada em *um novo curso do Eurocenter* caso se inclua o seu relacionamento com a entidade “Eurocenter”. Se for efetuada a mesma operação, adicionando as informações da sentença seguinte, esse resultado pode ser expresso em *um novo curso do Eurocenter na bela Kanazawa*. Desse modo, pode-se dizer que a instrução lingüística *esse curso* não está somente se alimentando das informações presentes nas expressões *um novo curso*, *ele* e *esse novo curso de japonês*, mas também está se abastecendo das informações presentes nas sentenças em que essas expressões correferentes estão instaladas.

Nessa perspectiva, a idéia de dependência pode chegar às raias da não correferencialidade. Em outras palavras, dentro do arcabouço da anáfora, um termo/expressão pode consultar (se “apoiar” semanticamente em) outro termo/expressão anterior para engendrar uma nova entidade do discurso. A expressão que apresenta esse tipo de dependência interpretativa é chamada expressão anafórica associativa. Um exemplo de anáfora associativa é a expressão *as aulas do nível avançado* – em (1). Ela, apesar de não ser correferente a nenhuma expressão anterior, apresenta parte do seu significado ancorado na expressão *esse novo curso de japonês*, tanto que a entidade evocada por ela pode ser alcançada pela “soma” dos elementos lexicais da expressão anafórica associativa (*As aulas do nível avançado*) e da sua âncora semântica (*esse novo curso de japonês*): as aulas do nível avançado do novo curso de japonês. Logo, uma anáfora pode ser um fenômeno de natureza inferencial.

Com base nos trabalhos apresentados no início desta seção, este trabalho concentrar-se em cinco classes de relacionamentos anafóricos associativos: *element-of*, *subset-of*, *part-of*, *entity-attribute* e *possessor-thing*. O primeiro, *element-of*, configura-se quando a entidade evocada é um elemento de um par ou grupo previamente introduzido, tal como ocorre entre *as aulas do nível avançado* (entidade grupo) e *uma dessas aulas avançadas* (entidade elemento do grupo) – em (1) – ou entre *as turbinas das asas* (entidade par) e *a turbina da esquerda* (entidade elemento do par) – em (5). Quando o elemento ocorre primeiro que o par/grupo, configura-se a relação inversa, denominada *element-of-inv*. Um exemplo de relacionamento inverso está presente em (5): relação entre *a turbina traseira do F-996* (entidade elemento do grupo) e *as turbinas* (entidade grupo).

( 5 ) Ao contrário do que muita gente pensa, a corrida armamentista não foi colocada de lado. Algumas nações investem em armas biológicas outras em armas nucleares. Porém, todas elas não abrem mão de uma poderosa força aérea. Ontem, a marinha dos USA apresentou o caça mais letal já projetado, o F-996.

Devido ao peso, o F-996 alcança mais de 500 metros por segundo com carga total de combate. No bico e nas asas há canhões de alta precisão e nos vinte metros de envergadura carregam 32 mísseis.

À noite, ele é praticamente invisível graças a um sistema contenção de som e luminosidade presente na turbina traseira do F-996. Mesmo os radares mais modernos não o detectam.

As turbinas das asas permitem um pouso na vertical. Além disso, as turbinas conseguem reduzir pela metade a temperatura fazendo com que ele escape de mísseis com sensores térmicos. Ao lado da turbina da esquerda, há um laser que pode tanto guiar seus mísseis por quilômetros como desviar mísseis, tipo Tomarock, de alvos aliados.<sup>19</sup>

O segundo relacionamento associativo, *subset-of*, abarca as expressões anafóricas que estão ancoradas semanticamente numa relação entre conjuntos. Às vezes, ela pode ser um subconjunto ou ser o conjunto, por isso, nesse liame não há relação inversa. Um exemplo de relação anafórica associativa *subset-of* é verificado entre *camundongos* [*com essa alteração genérica*] e *os machos* ou entre *camundongos* e *as fêmeas*; ou entre *energia* e *a lipídica* – em (6).

( 6 ) Um tratamento para a obesidade que faz você comer mais, perder peso e reduzir a taxa de gordura do corpo é o que sugere um estudo britânico publicado hoje na revista científica *Nature*.

---

<sup>19</sup> Os três primeiros parágrafos do texto 19 do corpus TeXto.

Por enquanto é só sugestão: o tratamento foi testado em camundongos. Mas os resultados levaram os cientistas a chamar o próprio estudo de “abordagem promissora” contra a obesidade. No centro do método está um gene humano descoberto recentemente, o UCP-3, cujos mecanismos de ação ainda não são totalmente conhecidos. Sabe-se que está envolvido no processamento de energia pelas células e que um gene da mesma família, o UCP-1, está ligado à queima de gordura.

O gene UCP-3 foi inserido em camundongos e manipulado para produzir, em excesso, a proteína determinada por ele. Os camundongos com essa alteração genética comeram até 54% mais que os camundongos normais. Apesar disso, pesavam até 23% a menos que seus companheiros. A porcentagem de tecido adiposo (gordura) sobre o volume total do corpo dos bichos também diminuiu, nos machos, em 44% e nas fêmeas, em 57%. Sua atividade física não diferiu significativamente em relação à dos camundongos normais. Isso quer dizer que os camundongos transgênicos reduziram a gordura de seu corpo, em relação à massa muscular, sem fazer ginástica. Os animais magros consumiram mais energia até para respirar. Em vez de armazená-la, transformaram-na em calor, principalmente a lipídica.<sup>20</sup>

Quando a expressão anafórica evoca uma entidade que é parte de outra já introduzida no discurso ocorre o terceiro relacionamento associativo, *part-of*. Esse vínculo pode ser observado na relação entre *o F-996* e *o bico* ou entre *o F-966* e *as asas* – em (5). Quando a parte ocorre antes do todo, surge a relação inversa, chamada *part-of-inv*.

A quarta relação, *entity-attribute*, compreende aqueles casos registrados entre entidades em que uma é um atributo da outra numa situação de anáfora associativa (dependência). Em outras palavras, a expressão anafórica evoca uma entidade que é um

---

<sup>20</sup> Os três primeiros parágrafos do texto CIÊNCIA\_2000\_17113 do corpus Summit..

atributo de outra já presente no discurso. Em (5), identificam-se duas relações *entity-attribute*: *o F-996* e *o peso*; *o F-996* e *a metade da temperatura*.

Tema de estudos em várias línguas, as anáforas associativas *possessor-thing* são uma herança do caso genitivo da língua latina. Esse tipo de relacionamento se estabelece quando a entidade antecedente possui a entidade evocada pela expressão anafórica. Por exemplo, nota-se uma relação *possessor-thing* entre *o ministro dos transportes* e *o apartamento* – em (7).

(7) Ontem, o ministro dos transportes foi convocado pela polícia para prestar esclarecimentos. Semana passada, foram encontradas, no apartamento, várias mercadorias contrabandeadas.<sup>21</sup>

Nem sempre é fácil distinguir um relacionamento *part-of* de um *entity-attribute*. Os testes lingüísticos baseados em construções lingüísticas gerais, tais como *x de y* ou *x tem y* não são de muita ajuda, pois, se pode dizer que “casas têm janelas/portas/cômodos” e que “casas têm uma altura/largura/preço”. Um critério razoavelmente útil é o *status* ontológico: as partes (*part-of*) tendem a ser objetos concretos, tais como portas, asas, rodas, motores, ossos, órgãos etc., e os atributos (*entity-attribute*) se detêm a objetos abstratos, tais como altura, peso, velocidade etc. Os testes lingüísticos que usam verbos mais específicos podem ser proveitosos: ao contrário das asas e do bico de um avião, pode-se dizer que o peso e a temperatura são um atributo dos objetos. Inversamente, se pode afirmar que as asas e o bico são partes dos aviões, mas não que o comprimento é uma parte do jato. Embora esses dois relacionamentos anafóricos possam apresentar, em certa medida, uma idéia de posse, as relações *possessor-thing* não são aqueles entre um objeto e um de seus atributos ou suas

---

<sup>21</sup> O primeiro parágrafo do texto 04 do corpus TeXto.

partes. O vínculo anafórico associativo *possessor-thing* é um complemento das relações *part-of* e *entity-attribute*, ou seja, ele dá conta das relações de posse em sentido específico. Evidentemente, as anáforas associativas não se encerram em cinco fenômenos. Por exemplo, geralmente, os epítetos<sup>22</sup> estão envolvidos em processos anafóricos associativos.

Em síntese, uma expressão anafórica pode retomar, atualizar, uma entidade anterior numa relação de identidade (anáfora direta ou indireta) ou pode engendrar uma nova entidade do discurso cuja interpretação está ancorada semanticamente em outros termos/expressões anteriores (anáfora associativa). Nos limites entre os conceitos de anáfora correferente e anáfora associativa se encontra a noção de anáfora encapsuladora<sup>23</sup>. Essa ocorre quando a expressão anafórica retoma trechos de texto precedente maior que um sintagma (geralmente, sentenças ou parágrafos, em alguns casos até mesmo todo o texto precedente), sintetizando-os em uma nova entidade do discurso. Em (6), observa-se que o pronome *isso* sintetiza a sentença *os camundongos com essa alteração genética comeram até 54% mais que os camundongos normais*, convertendo-a em uma entidade do discurso. Processo similar pode ser observado, em (8), com a expressão *a proposta*. Esse exemplo se distingue do anterior pelo volume de texto, de informação, sintetizado. Como nos exemplos, estudos realizados (por exemplo, VIEIRA, GASPERIN e SALMONALT, 2005; COELHO, COLLOVINI e VIEIRA, 2005) mostram que freqüentemente as anáforas encapsuladoras são disparadas por expressões formadas por expressões compostas por um pronome demonstrativo e um núcleo nome (tais como, *essa proposta*, *essa opinião*, *esse plano*, *esse parecer* etc.) ou por um artigo definido e um núcleo nome (tais como, *a suposição*, *a atitude*, *a constatação*, *o debate*, *a idéia* etc.) – geralmente o núcleo é um nome abstrato. Além disso, na maioria das vezes, esse

---

<sup>22</sup> Qualquer adjetivo ou expressão com valor de adjunto atributivo não ligada ao substantivo por um verbo.

processo influencia e é influenciado por outros. Por exemplo, em (8), pode-se dizer que a entidade evocada pela expressão *a proposta* é um tópico discursivo central, uma vez que compõem uma cadeia de correferência (*a proposta*, *a idéia* e *esse ponto de vista*) e/ou vice-versa.

(8) O presidente da Comissão Nacional de Ética em Pesquisa, William Saad Hossne, disse ontem, na 52ª Reunião Anual da Sociedade Brasileira para o Progresso da Ciência, que a comunidade científica internacional pode alterar a Declaração de Helsinque, sobre ética em pesquisas com humanos.

Em estudos no Terceiro Mundo, os cientistas se desobrigariam de fornecer aos doentes o melhor tratamento médico conhecido. A proposta dá aos pesquisados o direito de receber terapia dada pelo governo de seu país \_que pode ser nenhuma.

A idéia surgiu após estudos em Ruanda e na Tailândia em que cientistas deram a grávidas com HIV um regime de AZT mais breve do que o recomendado.

Contra esse ponto de vista, Hossne defende a norma atual: em pesquisa de tratamentos, os doentes devem receber ao menos o remédio mais eficiente já descoberto para sua doença. Hossne citou o estudo de Tuskegee (EUA), em que negros com sífilis não foram tratados por 40 anos para que a evolução da doença fosse estudada.<sup>24</sup>

O aparato teórico empregado aqui contempla também outra propriedade da linguagem natural, a dêixis. Ela, resumidamente, consiste em fazer uma expressão evocar um aspecto da situação comunicativa. São casos dêiticos as expressões: *hoje*, em (1), e *ontem*, em (5). Conforme pode ser percebido, essas expressões se apresentam de maneira vaga, inespecífica, sendo claras, consistentes, ao se recuperem as datas dos textos respectivamente. Uma

---

<sup>23</sup> Denominação inspirada em Koch (2002).

<sup>24</sup> Os quatro primeiros parágrafos do texto CIÊNCIA\_2000\_17101 do corpus Summit.

expressão dêitica pode evocar os interlocutores da enunciação, o discurso em si, o momento da enunciação, o lugar ou uma posição do local onde ocorre a enunciação.

De posse do exposto, pode-se dizer que, excetuando a dêixis, os demais processos (e outros) reúnem-se para compor cadeias de correferência e circunstanciar relações de anáforas, lembrando que toda composição é norteada pelos fatores intencionalidade e aceitabilidade, consistência e relevância, focalização, informatividade, sendo que, na base dos procedimentos, está o conhecimento lingüístico, o conhecimento compartilhado e o conhecimento de mundo. Embora existam outros fenômenos referentes à coesão referencial (por exemplo, a catáfora e a elipse), eles não são tratados neste trabalho devido aos fins computacionais deste. Reitera-se que este estudo visa apresentar subsídios para a Sumarização Automática com base nas informações sobre correferência e anáforas. Para compreender melhor isso, na sequência, se aborda o tópico Sumarização Automática.

## 1.4 Sumarização Automática

Os primeiros experimentos de automatização do processo de sumarização textual remontam aos anos cinquenta. Assim como ocorreu em outras áreas, em Sumarização Automática, à medida que os experimentos se multiplicavam, convenções terminológicas e metodológicas se consolidavam. Justamente, sobre esse tópico que esta seção se detém.

De início, são elencados alguns termos técnicos da área importantes para compreensão do presente trabalho. Para efeito de síntese, fixa-se a atenção nos termos consensuais, evitando expor flutuações terminológicas. O primeiro e mais habitual deles é a designação *texto-fonte*, que se refere ao texto que foi ou será sumarizado. O termo *sumarização*



compreende tanto as etapas de planejamento ou análise quanto os processos de edição do sumário, existindo uma dicotomia entre sumarização *automática* e *manual*. Em adição, há sumarização *multidocumentos*, cujo objeto de estudo é a geração automática de sumários de um assunto por meio da exploração de vários textos-fonte que discorrem sobre a matéria pertinente a esse determinado assunto. Por extensão de sentido, há, também, as expressões *sumarizador automático* (sistema de sumarização automática) e *sumarizador extrativo* (sistema de sumarização automática que geram, exclusivamente, extratos – um tipo específico de sumário).

Em Sumarização Automática, *sumário* e *resumo* são denominações sinônimas que definem, basicamente, o resultado final do processo de sumarização – automática ou manual. Sumários podem ser classificados em: indicativos, informativos e críticos (*evaluative*). *Sumários indicativos* apresentam apenas os tópicos essenciais do texto-fonte, sem incluir detalhes de resultados, argumentos e conclusões (HUTCHINS, 1987). Ao contrário, *sumários informativos* contêm, além dos tópicos essenciais, informações principais de resultados, argumentos e conclusões. Por exemplo, se um texto-fonte for organizado em função de dados, métodos, hipóteses e/ou conclusões, um sumário informativo, diferentemente de um indicativo, deverá conter as informações principais de cada um desses. Por isso, no campo de Sumarização Automática, os sumários informativos são considerados possíveis substitutos dos respectivos textos-fonte, na medida que informam com menor detalhamento. *Sumários críticos* também apresentam tópicos essenciais do texto-fonte, contudo, é patente neles algum juízo de valor em relação ao conteúdo e/ou à organização do texto-fonte. Ainda, cabe destacar algumas peculiaridades referentes às denominações sumários, índices e extratos. *Índices* são sumários indicativos que, obrigatoriamente, se apresentam em forma de lista (de tópicos ou

itens), freqüentemente, compostos por sintagmas nominais – tais como *vida marinha*, *a proposta de Rino* ou *comidas*. O termo *extrato* define os sumários resultantes da simples justaposição de sentenças; sendo reservada a designação *extratos ideais* para aqueles realizados por especialistas – geralmente, empregados na avaliação de aplicações computacionais.

A convenção terminológica da área também conta com os conceitos: *taxa de compressão*, *granularidade* e *saliência*. Essa última é equivalente à noção de *significância* ou *relevância* amplamente explorada em PLN. Ela norteia os critérios de seleção e associação de objetos textuais para a edição de sumários, podendo ser entendida como uma medida de relevância relativa dos itens textuais (RATH, RESNICK e SAVVAGE, 1961; BOGURAEV e KENNEDY, 1997; PARDO e RINO, 2001). Por exemplo, as instruções lingüísticas com alta *saliência* devem estar no foco da atenção do discurso e/ou apresentam, em relação ao conteúdo do texto-fonte, significativa informatividade. Logo, de alguma forma, devem fazer parte do sumário. A *granularidade* refere-se à segmentação do texto-fonte, sendo essa relativamente arbitrária em Sumarização Automática. As mais comuns são a granularidade individual (palavras) e a granularidade sentencial (sentença). Também relativamente arbitrária é a *taxa de compressão*, percentagem que responde pelo volume de redução do texto-fonte, sendo que ela está ligada à granularidade. Supondo-se que, para geração de um sumário, seja definida uma granularidade sentencial, a taxa de compressão será medida em número de sentenças que deverão ser excluídas do sumário final. A fórmula geral de taxa de compressão é apresentada na Figura 4.

$$\text{Taxa de compressão} = 1 - \left( \frac{\text{tamanho do extrato}}{\text{tamanho do texto-fonte}} \right)$$

**Figura 4 – Fórmula geral de taxa de compressão**

Por exemplo, para produzir um sumário que corresponda a 30% do tamanho do texto-fonte, deve-se ajustar a taxa de compressão para 70%; para gerar um sumário que corresponda a 10% do texto-fonte, deve-se parametrizar a taxa de compressão para 90%. Mesmo arbitrária, a comunidade da área sugere, para fins de pesquisa, valores de granularidade e de compressão de acordo com a abordagem adotada.

Em PLN, notadamente, em Sumarização Automática, essas abordagens podem ser divididas naquelas que empregam *conhecimento profundo* das que utilizam *conhecimento superficial*. Sob esse rótulo se encontram informações concernentes relativamente ao cotexto (contexto lingüístico), já sob aquele se encontram informações referentes ao contexto lingüístico e extralingüístico. Tradicionalmente, a abordagem que aplica o conhecimento profundo é denominada de fundamental enquanto a que faz uso de conhecimento superficial é chamada de abordagem empírica.

Conforme dito, uma abordagem empírica se detém, relativamente, à superfície do texto, usando técnicas baseadas em corpora e/ou estatísticas. Alguns desses métodos matemáticos são tradicionais na área. Um deles é o método de palavras-chave (LUHN, 1958). Ele parte do pressuposto de que as idéias centrais de um texto podem ser expressas por algumas palavras-chave. Black e Johnson (1988) apontam o princípio da repetição para

concepção dessa técnica, à medida que as idéias centrais se desenvolvem no texto, os termos-chave aparecem com maior frequência, a fim de assegurar a organização dos tópicos. Portanto, uma solução para elaborar um sumário pode estar presente em uma distribuição estatística a partir da frequência de termos. Geralmente, para o cálculo de frequência, são consideradas as categorias gramaticais abertas em detrimento das fechadas. Por exemplo, Earl (1970) ajusta seu sistema para levar em conta, exclusivamente, os substantivos mais frequentes de um texto-fonte a fim de elencar palavras-chave. Elas são responsáveis por mapear o conteúdo textual que será preservado para composição do sumário.

Há uma variação importante dessa técnica. Ela se concentra, unicamente, nas unidades lexicais do título e subtítulo para formar a lista de palavras-chave que nortearão o processo de sumarização. Em casos de textos extensos, tais como monografias e romances também inclui-se o resumo. Edmundson (1969), ao armar esse raciocínio, afirma que os itens lexicais presentes em um título têm maior probabilidade de serem representativos do tópico principal do texto-fonte. Mani e Maybury (1999) corroboram Edmundson ao argumentarem que o título tem o poder de avançar comunicativamente em termos de expectativas por meio de processos cognitivos. Aliás, a escolha do título pode decidir a orientação da leitura. Segundo esses teóricos, o título representa a base para a primeira seleção entre as possibilidades de expectativas. Ele constitui, em determinados gêneros textuais, uma paráfrase reduzida do texto-fonte.

Aos métodos de palavras-chaves, acrescentam-se os de localização. Baxendale (1958), com base em estudo de corpora, estabelece uma correspondência entre a posição de uma sentença e sua importância para interpretação do texto. Seus estudos mostram que a primeira e

a última sentença podem representar idéias principais de um parágrafo. Por exemplo, em 85% dos casos de uma amostra de 200 parágrafos, a primeira sentença apresenta a idéia central a ser desenvolvida pelo respectivo parágrafo. Em 7% dos casos, a idéia central está localizada na última sentença. Mesmo com o resultado pouco significativo das últimas sentenças, Baxendale orienta selecioná-las, pois elas exercem, nas amostragens pesquisadas, uma função de elo entre os parágrafos, aliás, função detectada, também, nas primeiras sentenças. Segundo o teórico, com essas medidas, há mais probabilidade da coesão textual estar assegurada na edição do sumário.

Em adição às técnicas anteriores, Paice (1981) expôs duas propostas: o método de frase auto-indicativa e o método de palavras sinalizadoras (*cue phrases*) ou marcadores lingüísticos. Essa técnica utiliza um dicionário formado a partir de vocábulos relevantes de um gênero textual específico. À medida que determinada sentença apresenta um termo presente no dicionário, ela tem seu peso incrementado, sendo que as sentenças com maior peso são selecionadas para conceber o sumário. O método de frase auto-indicativa funciona de modo análogo, pois, basicamente, seu diferencial está no repositório que, ao invés de termos, contém orações. Em artigos científicos, dois exemplos de frases auto-indicativas são: *o objetivo deste artigo é investigar* e *neste artigo, é descrito um método para*.

Ao passo que as limitações de hardware e software eram superadas e os repositórios lingüísticos de grande porte surgiam, os métodos estatísticos ganhavam força. Destacam-se, aqui, os corpora de grande porte, utilizados em *Data Mining*. Eles influenciaram o aparecimento de uma área significativa para o campo de Recuperação de Informação e Sumarização Automática: *Text Mining* (HEARST, 1999). Essa sinergia possibilitou que a

distribuição de pesos e os cálculos de frequência evoluíssem consideravelmente, como mostram, por exemplo, os estudos de Rau e Brandow (1993) e Larocca Neto et al. (2000).

Ainda que tais métodos tradicionais apresentem resultados interessantes para determinados gêneros textuais (produzindo sumários úteis para objetivos bem definidos), há uma grande possibilidade de esses processos gerarem sumários incoerentes ou desconexos. Isso ocorre porque, freqüentemente, essas técnicas não contam com recursos de rescrita dos segmentos selecionados, sendo que, na maioria das vezes, os sumários resultam da justaposição de sentenças extraídas do seu cotexto original, ou seja, são extratos. Rino (1996) adverte sobre limitações dos métodos superficiais, avisando, por exemplo, que os títulos podem representar um engodo em relação ao assunto e que uma informação relevante pode estar em uma posição intermediária do parágrafo – não necessariamente na primeira ou na última sentença de cada parágrafo.

Em vista dessas dificuldades e de outras, certos grupos de pesquisa investem em conhecimento profundo para modelagem de sistemas de sumarização. Como na abordagem fundamental empregam-se teorias lingüísticas e modelos formais, pode haver chances consideráveis de avanços para área. Nesse contexto de trabalho, geralmente, aos textos, são associadas, representações computacionais, tais como, aquelas que descrevem relações semânticas e retóricas ou identificam relações de intencionalidade. Com esse alto grau de representação, por via de regra, utilizam-se, no processo de planejamento de sumarização, refinados artifícios de manipulação simbólica e motores de inferência apoiados em raciocínio lógico. Mesmo para cálculos básicos, são empregadas medidas complexas (CLOCKSIN e MELLISH, 1981).

Rino (1996) explica que, ao contrário dos métodos estatísticos, a abordagem fundamental conta, na maioria das vezes, com um processo de realização lingüística do sumário. Esse processo pode se efetivar de várias maneiras (REITER e DALE, 2000; SCOTT e SOUZA, 1990). Por exemplo, é possível aproveitar aplicações computacionais já disponíveis como FUF (ELHADAD, 1991), que se vale de *f-structures* (KAPLAN e BRESNAN, 1982) como representação de entrada de dados, ou Nigel (MANN e MATTHIESSEN, 1985), que lança mão de “um modelo sistêmico de representação discursiva” fundamentado nas teorias de Halliday (1985). Também se pode empregar, no processo de realização lingüística, *case frames* (funções matemáticas das relações retóricas e semânticas) ou *templates* e *canned text* (estruturas rígidas especificadas a partir das propriedades lexicais e semânticas do respectivo gênero textual).

Entretanto, o mérito da abordagem fundamental está, especialmente, no uso de teorias formais, tais como, *Rhetorical Structure Theory* - RST (MANN e THOMPSON, 1987) ou *Veins Theory* - VT (CRISTEA, IDE e ROMARY, 1998). Algumas frentes combinam teorias nas arquiteturas, tais como a proposta por Rino (1996), implementada por Pardo (2002a), que utiliza a RST e a *Grosz and Sidner Discourse Theory* - GSDT (GROSZ e SIDNER, 1986), sendo que essa se preocupa com relações de intenção e o foco de atenção enquanto aquela trata as relações retóricas.

Todavia, quaisquer que sejam as abordagens, lugar de destaque ocupa a avaliação. Entre as perspectivas, destaca-se a avaliação do desempenho dos sumarizadores automáticos e da funcionalidade e/ou qualidade dos sumários gerados automaticamente, isto é, o quanto um sumário atende as necessidades do usuário. Em geral, nessa perspectiva de avaliação dos

sistemas computacionais, usam-se *gold standards* (KUPIEC, PETERSEN e CHEN, 1995) – padrões de referência definidos por especialistas em sumarização textual (PARDO, 2002a; PARDO, NUNES, RINO, 2004)<sup>25</sup>.

Em síntese, pode-se dizer que, em Sumarização Automática, os métodos apoiados em conhecimento profundo oferecem vantagens significativas para o processamento computacional, permitindo tratar aspectos ignorados pelas técnicas estatísticas – apesar dessas serem menos dependentes de esforços e recursos manuais.

Entretanto, em acordo com Rino, seja em nível superficial ou profundo, “cada qual apresenta tanto méritos quanto dificuldades” (1994, p.2). Por isso, as últimas tendências adotam metodologias híbridas, buscando arquitetar e a implementar sistemas voltados para a funcionalidade e para o usuário nunca esquecendo de aprimorar os métodos de avaliação.

A partir do exposto neste capítulo, é possível perceber a complexidade cognitiva e computacional da tarefa de interpretação do texto. Na busca de fornecer subsídios para essa intrigante tarefa, investiu-se, aqui, na construção de um corpus que promova a sinergia de informações: o corpus *Summ-it*. Ele é constituído por 50 textos do caderno Ciências do jornal Folha de São Paulo que foram tratados com diferentes níveis de anotação: bibliográfica, lexical, sintática, RST e de correferência e anáfora. O *Summ-it* também conta com sumários, ou seja, para cada um de seus textos (que nesse contexto são os textos-fonte) há um sumário

---

<sup>25</sup> Há casos em que os *gold standards* também são usados no treinamento de sistemas.



manual, um texto tarjado e um extrato automático. O corpus *Summ-it* com os meios que se pode dispor é o tópico do próximo capítulo.

## 2 CORPUS *SUMM-IT*

Um corpus lingüístico é uma amostra autêntica de uma língua ou variedade lingüística. Trabalhos lingüísticos baseados em corpus remontam o século XIX, muito antes do uso dos computadores. Sem a praticidade e a precisão dos computadores, os trabalhos baseados em corpora extensos estavam significativamente sujeitos a erros e a imprecisões além de serem uma tarefa bastante árdua (SARDINHA, 2000).

Para a língua inglesa, vários corpora foram construídos, desde o pioneiro corpus *Brown*, lançado em 1964 com 1 milhão de ocorrências. Em termos de megacorpora balanceados, tanto o *British National Corpus* (BNC), para a variante britânica, quanto o *American National Corpus* (ANC), para a americana, contribuem para o desenvolvimento de ferramentas de PLN e para a descrição da língua e construção de recursos, tais como dicionários e gramáticas. Além disso, esses corpora impulsionam o desenvolvimento de formatos padrões de anotação e codificação, além de arquiteturas para dados e para ferramentas de manipulação de corpus. São esses padrões internacionais que ajudam a criar corpora que sejam intensivamente usados, reusáveis e extensíveis.

Segundo Ide e Brew (2000), a *reusabilidade* (característica de um corpus ser usável em mais de um projeto de pesquisa e por mais de um grupo de pesquisadores) e a *extensibilidade*

(isto é, a capacidade de corpus serem melhorados em várias direções, por exemplo, com a provisão de um nível a mais de análise lingüística) são aspectos a serem considerados em projetos de corpora lingüísticos, já que, na atual conjuntura, os trabalhos de pesquisas não devem estar encerrados em si, mas em diálogo com a comunidade global.

Com vistas nessas orientações, neste capítulo, é descrito o corpus *Summ-it* juntamente com seus diferentes níveis de anotação (morfológica, sintática e sobre relações de correferência, RST etc.) – Seção 2.2 – e com seus sumários, desenvolvido especialmente para pesquisas relativas à Sumarização Automática – Seção 2.3. Entretanto, antes de reportar o processo de anotação e elaboração dos sumários, é relatada, na seção seguinte, a seleção dos textos originais (textos-fonte).

## 2.1 Seleção do corpus

O corpus *Summ-it* está ligado ao projeto de pesquisa CROWS, em relação aos cuidados nos usos de padrões de anotação de acordo com a *Web Semântica* e, por sua vez, a seus subprojetos: ProCaCoSa, em relação à anotação de correferência, e PLN-BR, em relação ao cuidados em produzir recursos computacionais.

Os textos do *Summ-it* foram coletados do corpus PLN-BR. Conforme dito na parte introdutória deste trabalho, o projeto PLN-BR tem por objetivo geral a construção de um espaço interinstitucional de interação e intercâmbio de práticas de análise e investigação lingüístico-computacional acerca da representação e da recuperação de informação de natureza semântica e discursiva veiculada por enunciados produzidos em português brasileiro.

Subdividido em sete subprojetos relativamente autônomos, esse projeto compartilha um mesmo ponto de partida: o corpus PLN-BR.

Esse corpus é formado por 103.080 mil textos de diversos cadernos (Casa, Ciência, Esporte, Moda, Política, Veículos etc.) do jornal Folha de São Paulo, compreendidos entre os anos de 1994 a 2005. Esse corpus originou-se das amostras disponibilizadas pelo projeto PLN-BR.

Desse corpus maior foram retirados os textos que constituem o corpus *Summ-it*, precisamente, 50 textos do caderno Ciências. Cada documento corresponde a um arquivo texto (ASCII) com tamanho entre 1 Kbytes e 4 Kbytes (de 127 a 654 palavras).

Os 50 textos originais do *Summ-it* foram, então, filtrados, isto é, caracteres de controle, títulos e autoria são removidos. Essas informações extraídas são fornecidas em arquivos auxiliares, seguindo o padrão internacional para codificação de corpus: o XCES. Isso e as outras informações que compõem a anotação do corpus *Summit* são reportadas na próxima seção.

## **2.2 Anotação do corpus**

Como dito, um corpus pode, geralmente, ser definido como um conjunto organizado de itens lingüísticos autênticos, organização essa externa e internamente. A organização externa diz respeito à espécie de corpus, ou seja, organização segundo, por exemplo, registro (corpus de texto jornalístico ou literário, ou ainda, corpus falado, dialetal) e tempo (por exemplo, corpus histórico). A organização interna relaciona-se, por outro lado, com decisões

sobre a configuração do corpus: o tratamento do material lingüístico – como, por exemplo, o que incluir/excluir do corpus, atomização (o que se considera uma palavra) e segmentação (critérios de divisão das sentenças) – e a categorização do material lingüístico, incluindo a parte de morfossintaxe, de semântica e discursiva. No que toca esse aspecto interno, surge uma noção importante: *treebank*. Constituirá um *treebank* um determinado corpus que tenha sido analisado automaticamente e/ou interpretado manualmente em formato específico de árvores.

Um formato bastante usado nesse sentido é um padrão internacional para codificação de corpus: o *Corpus Encoding Standard* (CES)<sup>26</sup>. Ele utiliza e adapta os padrões de diretrizes para codificação e intercâmbio de textos eletrônicos (TEI)<sup>27</sup> para codificar corpus. O CES é uma aplicação do SGML e possui uma versão mais atual em XML, o XCES. O *eXtensible Markup Language* (XML) é um padrão internacional para representação e intercâmbio de dados na Web que tem características e extensões úteis para a criação e manipulação de corpora anotados, entre elas (IDE, 2000; IDE et al., 2000):

- *XML Links*, que permitem endereçar os elementos XML tanto dentro de um mesmo documento como em outros documentos;
- linguagem *XPath* e *XPoint* que, através de predicados, permitem localizar elementos na estrutura de elementos (em árvore) e selecionar fragmentos do texto;
- *XSLT*, que pode ser usada para converter um documento XML em outro formato;

---

<sup>26</sup> Disponível em: <http://www.cs.vassar.edu/CES/>

<sup>27</sup> Disponível em: <http://www.tei-c.org/>

- *XML schemas* que estendem o poder dos DTDs (*Definition Type Document*), permitindo uma avaliação melhor tanto da forma quanto do conteúdo dos documentos XML, sendo que tanto DTD como *XML Schema* são metalinguagens que permitem comprovar a integridade dos dados em qualquer momento.

Esses dois padrões, XCES e XML, estão na base da construção do corpus *Summ-it*, na medida em que foi utilizando o XCES para codificar as informações bibliográficas de cada um dos 50 textos e o XML para representar os diferentes níveis de anotação.

### 2.2.1 Anotação XCES

Assim como o corpus PLN-BR, as informações bibliográficas dos 50 textos do *Summ-it*, tais como data de publicação, data de coleta, palavras-chave, autoria, tamanho do arquivo, número de palavras, sentenças e parágrafos, etc., estão estruturadas em XCES cujo esquema de utilização desdobra-se em quatro arquivos: um contendo dados bibliográficos e a classificação textual do texto original (Figura 5), um segundo com as estruturas dos parágrafos (Figura 6), um terceiro com uma marcação lógica (Figura 7) e outro com o texto indexado em sentenças – cada sentença recebe um identificador único (Figura 8). Destaca-se que o padrão XCES é uma marcação *stand-off*, ou seja, é aquela feita separadamente do corpus. Nela as etiquetas não são inseridas no corpus original, mas sim é criado um repositório de marcações com apontadores para elementos do corpus, os *treebanks*. Para que seja possível fazer esse tipo de marcação, os elementos do texto a serem marcados (palavras, sentenças etc.) devem estar indexados, para que possam ser referenciados nas marcações. Isso é alcançado pelo arquivo índice.

```

...
<biblStruct>
  <monogr>
    <title>Vespa parasita usa 'arma química' para escravizar aranha</title>
    <title>Droga secretada por larva provoca uma mudança no padrão de
      comportamento do animal hospedeiro</title>
    <author>CLAUDIO ANGELO</author>
    <biblNote>CIÊNCIA</biblNote>
  </monogr>
</biblStruct>
...

```

**Figura 5 – Arquivo de dados bibliográficos XCES**

```

...
<p id="p2">E não se trata de nenhum extraterrestre. Apesar do nome Hymenoepimecis
  sp., o tal invasor de corpos é só uma vespa.</p>

<p id="p3">O biólogo William Eberhard, da Universidade da Costa Rica, descobriu que as
  larvas desse inseto, ao parasitar a aranha Plesiometa argyra, provocam mudanças no
  comportamento da hospedeira.</p>
...

```

**Figura 6 – Arquivo da estrutura de parágrafos XCES**

```

...
<struct type="div" from="0" to="1700">
  <feat name="type" value="materia"/>
</struct>
...

```

**Figura 7 – Arquivo de marcação lógica XCES**

```

...
<struct type="s" from="0" to="130">
  <feat name="id" value="p1s1"/>
</struct>
<struct type="s" from="131" to="183">
  <feat name="id" value="p1s2"/>
</struct>
...

```

**Figura 8 – Arquivo índice XCES**

As principais vantagens da marcação *stand-off* em relação à marcação convencional, em que são inseridas etiquetas diretamente no corpus, são as seguintes (WYNNE, 2001):

- A integridade do texto não é comprometida pela anotação;
- Os arquivos de texto não se tornam muito grandes nem muito confusos;
- O trabalho colaborativo é facilitado;
- Múltiplas anotações (e anotações de anotações) se tornam mais fáceis.

### 2.2.2 Anotação automática dos programas PALAVRAS e Palavras Xtractor

Assim como a representação em XCES, que permite a separação entre os dados originais e anotações para que, por exemplo, se possa aplicar, futuramente, outros tipos de etiquetas num mesmo corpus, optou-se pelo XML para representar o resultado da anotação gramatical do corpus *Summ-it*.

Esse resultado é gerado pelo programa PALAVRAS, desenvolvido para o português por Eckhard Bick (2000). Ele realiza, automaticamente, as etapas de *tokenização*, processamento léxico-morfológico e análise sintática. O PALAVRAS faz parte de um grupo de analisadores do projeto VISL (*Visual Interactive Syntax Learning*)<sup>28</sup>, do *Institute of Language and Communication da University of Southern Denmark*. Esse robusto recurso computacional anota em formato de CG (*Constraint Grammar*)<sup>29</sup>. Ele recebe como entrada o conjunto de sentenças do corpus em forma ASCII e gera a análise morfossintática e semântica das sentenças em formato próprio.

---

<sup>28</sup> Disponível em: <http://visl.sdu.dk>

<sup>29</sup> Disponível em: [http://beta.visl.sdu.dk/constraint\\_grammar.html](http://beta.visl.sdu.dk/constraint_grammar.html)



De posse dessa análise, é empregada outra ferramenta, o Palavras Xtractor (GASPERIN et al., 2003), desenvolvida para converter a saída do PALAVRAS em três arquivos XML. O primeiro, denominado arquivo de *words*, é um arquivo básico de palavras que funciona como índice para os itens lexicais do texto. Ele codifica as palavras do texto em elementos <word> e atributos *id*, sendo que esse corresponde a um identificador único de cada palavra do texto. A Figura 9 ilustra um trecho de um arquivo de *words*.

O segundo arquivo contém a análise lexical das palavras do texto (*Part-of-Speech*), por isso ele é designado de arquivo de *PoS*. Nele estão anotadas informações gramaticais que compreendem as classes de palavras (nome, pronome, verbo, advérbio, adjetivo etc.) e seus paradigmas de flexões (número, gênero, pessoa, caso, tempo etc.), como na Figura 10 (referente ao sintagma nominal *a substância*), em que o elemento <n> indica a forma nome comum (*nouns*) e os atributos “F” (feminino) e “S” (singular) descrevem a flexão genérica e numérica respectivamente. Em adição, o arquivo de *PoS* oferece uma anotação semântica<sup>30</sup>, codificada por mais de 200 etiquetas semânticas (*semantic tags*)<sup>31</sup>. Embora o sistema conte com etiquetas semânticas para verbos e adjetivos, essa anotação abarca de modo mais efetivo os nomes – comuns (*nouns*) – e as entidades mencionadas (*named entities*) – nomes próprios, datas, horários etc. Por exemplo, na Figura 10, observa-se *secondary tags* que oferecem informações semânticas: *am*, *f* e *cm* exprimem, respectivamente, que a palavra apresenta o traço semântico de ser não contável, representar característica/propriedade (química) e ser concreto.

<sup>30</sup> Disponível em: [http://beta.visl.sdu.dk/visl/da/info/prototype\\_project.html?guest=1](http://beta.visl.sdu.dk/visl/da/info/prototype_project.html?guest=1)

<sup>31</sup> Disponível em: <http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html>

O terceiro arquivo contém as estruturas sintáticas das sentenças representadas por elementos *chunks*, por isso designa-se esse arquivo de *chunks*. Ele codifica a estrutura interna da sentença que pode apresentar *sub-chunks*. Por exemplo, na Figura 11 (referente ao sintagma nominal *a aranha*), o elemento *chunk* pai (atributo *id*="chunk\_255") constitui um sintagma nominal (atributo *form*="np") cuja função sintática é sujeito da oração (atributo *ext*="subj"). Por sua vez os *chunks* filhos (atributo *id*="chunk\_256" e *id*="chunk\_257"), são, respectivamente, um artigo (atributo *form*="art") que está numa relação de dependência (atributo *ext*="n") com *substância* (atributo *span*="word\_191"), nome núcleo do *chunk* pai (atributos *form*="n" e *ext*="h" respectivamente).

```
...
<word id="word_190">A</word>
<word id="word_191">substância</word>
<word id="word_192">atinge</word>
<word id="word_193">o</word>
<word id="word_194">sistema</word>
<word id="word_195">nervoso</word>
...
```

Figura 9 – Arquivo de *words*

```
...
<word id="word_190">
  <art canon="o" gender="F" number="S">
    <secondary_art tag="artd"/>
  </art>
</word>
<word id="word_191">
  <n canon="substância" gender="F" number="S">
    <secondary_n tag="am"/>
    <secondary_n tag="f"/>
    <secondary_n tag="cm"/>
  </n>
</word>
...
```

Figura 10 – Arquivo de *PoS*

```

...
<chunk id="chunk_255" ext="subj" form="np" span="word_190..word_191">
  <chunk id="chunk_256" ext="n" form="art" span="word_190">
    </chunk>
  <chunk id="chunk_257" ext="h" form="n" span="word_191">
    </chunk>
  </chunk>
...

```

Figura 11 – Arquivo de *chunks*

### 2.2.3 Anotação de correferência e anáforas

Na literatura, é possível encontrar vários trabalhos que apresentam sistemas de resolução automática de correferência (e anáforas) ou sistemas que contam com tal recurso para realizar tarefas de PLN. Para ilustrar isso, se pode apontar um evento internacional: *Discourse Anaphora and Anaphor Resolution Colloquium - DAARC*<sup>32</sup>. Aliás, a Resolução Anafórica ou Resolução de Anáforas figura como uma importante subárea do PLN. Precisamente, ela compreende as aplicações computacionais que visam mapear, de modo automático, os fenômenos relativos à coesão referencial (relações anafóricas, de correferência, casos de pronomes zero – elipses – etc.), assim como estudos relacionados a esse tópico. Para a avaliação, e de certo modo para o desenvolvimento das aplicações de Resolução Anafórica, é necessário um corpus padrão (*gold standard*), anotado manualmente. Para o inglês, são utilizados os corpora MUC-6 e ACE, disponibilizados pelo *Linguistic Data Consortium – LDC*<sup>33</sup>. Para o português, não existe ainda um corpus anotado com esse tipo de informação.

<sup>32</sup> O último DAARC está disponível em: <http://daarc2007.di.fc.ul.pt/>

<sup>33</sup> Disponível em: <http://www ldc.upenn.edu/>

Para preencher essa lacuna, foi realizada a anotação do corpus *Summ-it* com informações de correferência e anáforas dos seus textos.

A anotação do *Summ-it* seguiu as instruções para anotação de informações de correferência e de referências dêiticas, designadamente, elaboradas para o discurso escrito do português do Brasil (*Guidelines*, versão 2.7 – Anexo 2). A metodologia de anotação é fundamentada na discussão apresentada na Seção 1.3 e conta com o PALAVRAS (Seção 2.2.2) e a ferramenta de anotação *Multi-Modal Annotation in XML* - MMAX (MÜLLER e STRUBE, 2001).

A MMAX (*Multi-Modal Annotation in XML*) é específica para anotação de corpus, sendo seu principal enfoque a anotação de correferência. Neste trabalho, a ferramenta utiliza, seguindo a metodologia *stand-off*, o arquivo de *words* e o arquivo de *chunks* gerados pelo PALAVRAS. Além desses, ela usa um terceiro arquivo que recebe e armazena a anotação de correferência, o arquivo de *markable*, e um quarto que codifica o esquema de anotação, o *annotation scheme file*. Esse último é desenvolvido manualmente. Já o arquivo *markable* é gerado pela própria ferramenta MMAX. Ela codifica as marcações como elementos *markables* (ou seja, as unidades de interesse), associando-os a vários atributos.

Basicamente, o esquema de anotação seguiu os seguintes passos: seleção das unidades de interesse (*markables*), identificação das suas configurações lexicais, construção das cadeias de correferência, classificação dos *markables* e de seus relacionamentos sob a perspectiva das cadeias de correferência.

Na anotação de correferência (e anáforas), as unidades de interesse são os sintagmas nominais, considerando, evidentemente, as entidades mencionadas<sup>34</sup> e estruturas simples (por exemplo, de pronomes substantivos<sup>35</sup>), e desconsiderando as estruturas oracionais. Os *markables* foram gerados de modo semi-automático, isto é, primeiramente, as unidades de interesse foram selecionadas automaticamente a partir da anotação em formato XML do PALAVRAS e, posteriormente, esse resultado foi revisado. O processo de revisão se deu em dois momentos. Num primeiro momento, seguindo as instruções de anotação (*Guidelines*, versão 2.7), dois revisores inspecionaram cada um dos *markables*, assinalando prováveis desacertos. Num segundo momento, um terceiro revisor, em acordo com os outros dois, efetuou as correções. Todos os casos de correção foram registrados para futuramente aprimorar as instruções de anotação (*Guidelines*) e apurar o processo de seleção automática dos *markables*.

O segundo passo foi a identificação das configurações morfossintáticas dos *markables*, ou melhor, nos atributos *np\_form* (para os sintagmas nominais com núcleo nome) e *pro\_form* (para os sintagmas nominais com núcleo pronome), os *markables* foram rotulados com uma das 18 etiquetas oferecidas pelo esquema da anotação. Esse processo, assim como o anterior, ocorreu de modo semi-automático, após a geração automática, o resultado foi revisado manualmente.

Para os sintagmas nominais com núcleo substantivo comum, as etiquetas são: **def-np** (determinante artigo definido – *os pesquisadores*), **indef-np** (determinante artigo indefinido – *um filhote*), **dem-np** (determinante pronome demonstrativo – *essa medida*), **poss-np**

---

<sup>34</sup> Ou seja, sintagmas nominais com núcleo nome próprio.

(determinante pronome possessivo – *nossas pesquisas*), **int-np** (determinante pronome interrogativo – *que horas são?*), **num-np** (determinante numeral – *971 espécies*), **coord-np** (sintagmas nominais coordenados – *vinho e queijo*), **quant-np** (determinante quantificador – *várias respostas*) e **bare-np** (sintagma nominal sem determinante – *estrelas*).

Para os sintagmas nominais com núcleo substantivo próprio, as etiquetas são: **def-pn** (artigo definido – *o Brasil*) e **pn** (sintagma nominal sem determinante – *Brasil*).

Para os sintagmas nominais compostos exclusivamente por pronomes, as etiquetas são: **indef-pro** (pronome indefinido – *alguém*), **pes-pro** (pronome pessoal – *eles*), **dem-pro** (pronome demonstrativo – *isso*), **poss-pro** (pronome possessivo – *meu*), **int-pro** (pronome interrogativo – *quando*) e **num-ana** (numeral – *um*).

O terceiro passo, construção das relações das cadeias de correferência, assim como os próximos passos, foram realizados manualmente por uma equipe de 12 anotadores, sendo que cada texto foi anotado por dois membros dessa equipe. Obviamente, todos os passos, inclusive os três últimos, foram antecidos por uma leitura atenta do texto.

No terceiro passo, o anotador, efetivamente, montou as cadeias de correferência. Primeiramente, ele classificou cada um dos *markables* como correferente (*old*), anafórico associativo (*associative*), dêitico (*deictic*) ou nova no discurso (*new*)<sup>36</sup>. Posteriormente, os *markables* classificados como correferentes foram ligados a outros, sempre numa relação de identidade, usando o mecanismo *member* da MMAX. Para os *markables* classificados como

---

<sup>35</sup> Ou seja, pronome que não está junto com um nome, e sim, o substitui, podendo ser demonstrativo, indefinido,

anafóricos associativos, o anotador apontou o *markable* que serve de âncora semântica, utilizando o mecanismo *pointer* da MMAX.

Depois de mapear os relacionamentos de correferência e anafóricos, o anotador explicitou, no quarto passo, o tipo de relação de correferência, isto é, ele classificou, os *markables* correferentes, no atributo *is\_anaphoric*, em:

- *direct*: o nome núcleo do *markable* é idêntico a um dos outros nomes núcleos da cadeia de correferência (*uma proteína – a proteína*);
- *indirect*: o nome núcleo do *markable* é diferente a todos os outros da cadeia de correferência (*a aranha – o inseto*);
- *encapsulation*: o *markable* é uma anáfora encapsuladora (*a temperatura no núcleo alcançou 250°C – o aquecimento*). Nesses casos, o trecho retomado foi registrado por meio do mecanismo *comment* da MMAX.

O último passo foi refinar as informações sobre os relacionamentos anafóricos associativos, ou seja, além de estabelecer a relação associativa, o anotador a classificou no atributo *is\_bridging*. As opções para esse nível de anotação são:

- *element-of*: a expressão anafórica é um elemento de um grupo previamente introduzido (*as margens do Taquari – a margem esquerda*). Nos casos em que o elemento ocorre antes do conjunto, é utilizada a relação inversa, *element-of-inv*;
- *subset-of*: a expressão anafórica refere-se a um (sub)conjunto de uma entidade introduzida anteriormente no texto (*os bichos – os machos*);

- *part-of*: a expressão invoca parte de uma entidade mencionada anteriormente (*macieiras transgênicas – a maçã*). Quando a parte ocorre antes do todo, é empregada a relação inversa, *part-of-inv*;
- *entity-attribute*: a expressão refere-se a um atributo de uma entidade previamente mencionada (*o fóssil de raptor – o estado de conservação*);
- *possessor-thing*: a expressão âncora semântica apresenta uma relação de posse com a presente expressão, anafórica associativa (*a fábrica – o terreno*);
- *other-bridging*: outro tipo de relação anafórica associativa não definida pela anteriores (*grávidas com HIV – a contaminação do feto*).

Cabe salientar que o mapeamento das cadeias de correferência e respectivas classificações foram realizados em três momentos. Num primeiro momento, cada anotador fez, individualmente, sua leitura do comportamento das cadeias de correferência do texto – através da textura dos mecanismos *member* e *pointer* e da definição dos atributos *is\_anaphoric* (classificação das relações de correferência) e *is\_bridging* (classificação das relações anafóricas associativas). Num segundo momento, os anotadores reuniram-se com seu par (cada texto do corpus foi anotado por dois pesquisadores), a fim de discutir a anotação e dirimir possíveis dúvidas. Se necessário, os anotadores estavam autorizados a realizar alterações.

Atualmente, está ocorrendo a fase em que os pares de anotações são comparados para se obter um consenso. Essa concordância está sendo analisada por três juízes. Em linhas gerais, o consenso reúne os pontos de vista nos casos de anáforas associativas e ajusta as

---

<sup>36</sup> Isto é, uma nova entidade no discurso que não apresenta uma relação anafórica associativa.



situações de correferência em acordo com os anotadores, sendo que, quando necessário, foram realizadas reuniões com os anotadores.

Além disso, destaca-se que, para organizar a anotação do corpus, os textos foram divididos em quatro partes conforme mostra a Tabela 2. Essa tabela também reporta o tempo do primeiro momento de anotação, anotação individual sem discussão com o par ou com o grupo.

As anotações de correferência e anáforas estão armazenadas nos arquivos de *markables*. Um trecho desse documento XML é apresentado na Figura 12. Para ilustrar, nela, precisamente, no elemento <markable> com identificador "32", o *span* demarca o intervalo do arquivo de *words* (word\_103..word\_104) que corresponde à expressão em questão (*o combustível*), o atributo *np\_form* oferece a forma dessa expressão, uma descrição definida com núcleo nome comum (*def-np*), e o atributo *status* indica que esse sintagma nominal está envolvido em uma relação de correferência (*old*) com as expressões da cadeias de correferência identificada como *set\_27* (*o biodiesel, biodiesel, esse recurso, esse combustível e o biodiesel*). A Figura 13, arquivo de *words*, serve como suporte para ilustração anterior.

```
...
<markable id="markables_32" span="word_103..word_104" is_anaphoric="indirect"
  np_n="yes" np_form="def-np" status="old" member="set_27" />
<markable id="markables_33" span="word_106..word_110" np_n="yes"
  np_form="bare-np" />
<markable id="markables_34" span="word_109..word_110" is_bridging="element-of"
  np_n="yes" np_form="def-np" status="associative" pointer="markables_1" />
...
```

Figura 12 – Arquivo de *markables*

```

...
<word id="word_100">a</word>
<word id="word_101">produção</word>
<word id="word_102">de</word>
<word id="word_103">o</word>
<word id="word_104">combustível</word>
<word id="word_105">.</word>
<word id="word_106">Cálculos</word>
<word id="word_107">iniciais</word>
<word id="word_108">de</word>
<word id="word_109">o</word>
<word id="word_110">ministério</word>
...

```

**Figura 13 – Arquivo de *words***

Como recurso visual, esses dados foram convertidos para HTML. Em outras palavras, para cada anotação foi gerado um relatório no formato HTML (*HyperText Markup Language*). HTML é uma linguagem de marcação utilizada para produzir páginas na Internet, pois os documentos HTML podem ser interpretados por navegadores. Essa tecnologia é resultado da união dos padrões HyTime e SGML. O HyTime é um padrão para a representação estruturada de hipermídia (esse padrão é independente de outros padrões de processamento de texto em geral) e o SGML é um padrão de formatação de textos (não foi desenvolvido para hipertexto, mas tornou-se conveniente para transformar documentos em hiper-objetos e para descrever suas ligações). O Anexo 3 oferece um exemplo desse relatório HTML.

Os resultados da anotação de correferência apresentados aqui estão baseados no cálculo da média entre a anotação de dois anotadores para cada texto. Embora a anotação de correferência esteja concluída para as quatro partes, o consenso final está em fase de conclusão. De posse da anotação de correferência do *Summ-it*, reporta-se um total de 560

cadeias de correferência, tendo em média cada cadeia 6 membros – a menor com o número mínimo, dois membros, e a maior com 32 membros.

Códigos dos Textos	Kb	Anotador 1	Anotador 2	Anotador 3	Anotador 4	Anotador 5	Anotador 6	Anotador 7	Anotador 8	Anotador 9	Anotador 10	Anotador 11	Anotador 12
<b>PACOTE 1</b>													
2000 17088	2	48 min	38 min										
2000 17101	2			60 min	63 min								
2000 17112	2					45 min	25 min						
2005 28747	2							26 min	72 min				
2000 17108	2									22 min	90 min		
2004 26415	2											69 min	30 min
2000 17082	2	53 min		60 min									
2000 17109	2		42 min		70 min								
2000 17113	3					50 min				31 min			
2002 22023	3						45 min						
2003 24219	3							30 min				88 min	
2005 28756	3								84 min				44 min
<b>PACOTE 2</b>													
2002 22010	4	95 min											51 min
2002 22015	4		42 min								95 min		
2002 22029	4			90 min						45 min			
2003 24226	4				80 min			40 min					
2004 26423	4					87 min	48 min						
2004 26425	4								58 min			52 min	
2005 28755	4				64 min								35 min
2005 28764	4		48 min	90 min									
2005 28766	4					137 min				50 min			
2005 28774	4						70 min	40 min					
<b>PACOTE 3</b>													
2001 19858	3	45 min											25 min
2002 22027	3		57 min						89 min				
2003 24212	3			75 min						35 min			
2002 22005	3				35 min								
2005 28754	3					80 min	80 min						
2005 28752	3								93 min			88 min	
2004 26417	3				33 min								30 min
2005 28743	3	40 min		60 min									
2000 6391	2					28 min				15 min			
2001 6406	2						30 min						
2001 6414	2		22 min									36 min	
2002 6441	2		27 min							10 min			
2001 6410	1	20 min			20 min								
2001 6423	1					30 min							
2003 6472	1			30 min			20 min						
<b>PACOTE 4</b>													
2005 6507	1	35 min											15 min
2000 6380	2		74 min						68 min				
2000 6381	2			20 min						20 min			
2000 6389	2				25 min			15 min					
2001 6416	2					45 min	35 min						
2003 6457	2								72 min			58 min	
2003 6465	2				14 min								15 min
2004 6480	2	45 min		20 min									
2004 6488	2					13 min				7 min			
2004 6494	2						25 min	15 min					
2005 6514	2		69 min									62 min	
2005 6515	2		63 min						66 min				
2005 6518	2					41 min		20 min					

**Tabela 2 – Organização e tempos de anotação do corpus *Summ-it***

Referente ao segundo passo, a Tabela 3 apresenta a distribuição das configurações morfosintáticas dos *markables* – atributos *np\_form* e *pro\_form*. Corroborando estudos anteriores (Seção 1.3), observa-se que as descrições definidas com núcleo nome comum representam a classe mais numerosa do corpus (40,99%).

<i>np_form</i>	# (%)	<i>pro_form</i>	# (%)
def-np	2069 (40,99%)	pes-pro	154 (3,05%)
bare-np	1132 (22,43%)	dem-pro	35 (0,69%)
indef-np	384 (7,60%)	num-ana	27 (0,54%)
def-pn	384 (7,60%)	indef-pro	21 (0,42%)
pn	308 (6,10%)	int-pro	6 (0,12%)
num-np	156 (3,08%)	poss-pro	0 (0%)
quant-np	110 (2,18%)	<b>Total <i>pro_form</i></b>	243 (4,82%)
coord-np	98 (1,93%)		
dem-np	90 (1,78%)		
poss-np	73 (1,45%)		
int-np	2 (0,04%)		
<b>Total <i>np_form</i></b>	4804 (95,18%)		
<b>Total <i>markables</i>:</b>		<b>5047 (100%)</b>	

**Tabela 3 – Configurações dos *markables* do corpus *Summ-it***

Existe, ainda, para esse corpus, a anotação de estruturas retóricas com base na RST (MANN e THOMPSON, 1987). Uma apresentação breve desse nível de anotação é dada no Anexo 4. Todavia, esse nível de anotação está fora do escopo deste trabalho. A anotação RST foi desenvolvida com a coordenação da Dr<sup>a</sup> Lucia Helena M. Rino do grupo NILC/UFSCar.

## 2.3 Sumários

O corpus *Summ-it* conta, também, com sumários. Entre outros fins, os sumários visam atender pesquisas relativas à Sumarização Automática e sua relação com as cadeias de correferência e anáforas. Esse material foi construído e está organizado de acordo com as orientações da comunidade da área, registradas por Pardo e Rino (2001, 2003).

Os sumários estão dispostos em um repositório que contém 3 tipos de textos: *tarjados*, *sumários manuais* e *extratos automáticos*. Os textos tarjados correspondem aos textos-fonte (em formato ASCII) em formato Word (.doc) com marcações de destaque para algumas sentenças do texto. Esses arquivos possuem o mesmo nome dos textos-fonte acrescidos do prefixo *tarjado*. Os sumários manuais e os extratos automáticos também têm os nomes acrescidos de um prefixo: *sum* e *ext* respectivamente. Os textos tarjados e os sumários manuais foram elaborados por sumarizadores profissionais e sumários automáticos (extratos automáticos) foram gerados por uma ferramenta de sumarização automática, o GistSumm (PARDO, RINO e NUNES, 2003). Em suma, para cada texto do corpus *Summ-it*, há um texto tarjado, um sumário manual e um extrato automático. Logo, os textos tarjados e de sumários formam um conjunto de 150 arquivos.

### **2.3.1 Elaboração dos textos tarjados e sumários manuais**

A marcação dos textos tarjados e os sumários foram realizados por uma equipe de três sumarizadores profissionais. Todos eles, pós-graduados, trabalham a mais de dez anos na seleção, produção e sumarização de textos para o poder público. Na gestão pública, freqüentemente, são encomendadas varreduras documentais que são entregues em forma de sumários. Geralmente, a esses resumos são associadas imagens (fotos, gráficos, tabelas, plantas etc.). Todo esse material agiliza as decisões do poder público. Também é comum, na esfera pública, a encomenda de textos que apresentam à sociedade o ponto de vista de uma determinada secretaria estatal. Em nota, esses textos registram a “voz” do Estado no plano das relações públicas. Esses dois exemplos são algumas das atividades que fazem parte da rotina dos profissionais que elaboraram os sumários e as marcações do presente corpus.

Precisamente, solicitou-se a esse grupo de profissionais duas tarefas (Anexo 5): a construção de um texto tarjado para cada texto-fonte do corpus (Tarefa 1) e a elaboração de um sumário para cada texto-fonte do corpus (Tarefa 2). A fim de facilitar, organizar, a realização dessas tarefas, os textos-fonte foram separados em quatro partes, sendo que cada parte foi dividida em outras três partes, uma para cada sumarizador profissional. O sumarizador somente recebia uma nova parte mediante a entrega da anterior. As partes eram compostas pelos textos-fonte de responsabilidade do sumarizador em formato *doc*. A devolução do texto tarjado e do sumário também se dava nesse formato.

Posteriormente, o material recebido foi tratado e padronizado. O arquivo com marcações (texto tarjado) foi formatado de acordo com um modelo previamente definido (margens, estilo e tamanho da fonte, parágrafos, espaçamentos etc) e foi renomeado de acordo com o padrão de adição de prefixos. O sumário foi convertido para o formato ASCII e filtrado. Uma vez recebido, o material foi revisado e padronizado de acordo com o exposto anteriormente.

Para a realização da Tarefa 1, foi pedido ao sumarizador que: **i)** marcasse a(s) sentença(s) que lhe indicam a idéia principal de cada texto-fonte, isto é, aquela(s) que lhe dão a diretriz para a elaboração do respectivo sumário, e que **ii)** selecionasse entre as marcadas uma que seja a mais representativa da idéia principal do texto (no caso, de haver apenas uma sentença marcada em **i** essa passa automaticamente a atender esse pedido). No arquivo *doc*, as sentenças marcadas no pedido **i** estão destacadas pelo sublinhado e, no pedido **ii**, pelo negrito. Além disso, para facilitar a visualização, tanto as sentenças sublinhadas quanto as em negrito estão realçadas pela cor vermelha.

Na execução da Tarefa 2, solicitaram-se, exatamente, sumários informativos, ou seja, aqueles que a fidelidade ao original (o sumário apresenta sentidos que também podem ser alcançados pelo texto-fonte, jamais aquele contradizendo as idéias desse<sup>37</sup>) se sobressai como aspecto necessário e que, por isso, podem ser substitutos dos originais correspondentes. Em adição, ajustou-se, à grosso modo, o tamanho dos sumários em, aproximadamente, 25-30% do tamanho do texto-fonte correspondente. Sob a perspectiva de Sumarização Automática, isso equivale a fixar a taxa de compressão dos textos-fonte ao intervalo de 70-75%. Caso o sumarizador julgasse pertinente, outras variações de extensão poderiam ser consideradas – situação que não ocorreu.

Ao observar a realização das duas tarefas, foram notados, nos membros do grupo de sumarização, certos procedimentos que influenciaram no resultado fim da tarefa. Detectou-se, na definição da espécie de sumário desejado (sumários informativos), que a classificação adotada neste trabalho é equivalente à empregada pelos sumarizadores. Segundo eles, embora esse tipo de resumo seja um dos mais encomendados, nem todo texto pode ser sintetizado em forma de sumário informativo, com efeito, textos com estratégias argumentativas complexas e textos consideravelmente densos.

De acordo com as orientações apresentadas (Anexo 5), cada sumarizador primeiramente efetuou a Tarefa 1 e, posteriormente, a Tarefa 2. Entretanto, antes de iniciar as duas tarefas, todos os sumarizados, ao seu estilo, estudaram uma amostra de cinco textos do corpus coletados aleatoriamente, a fim de se familiarizarem com o gênero e o tipo de texto.

---

<sup>37</sup> Retornando os fatores de *consistência* e de *relevância* da seção Coerência Textual.

Também foi relatado pelo grupo que a Tarefa 1 é um procedimento básico para execução da Tarefa 2. A isso, soma-se um cuidado na textura das entidades do texto-fonte. Afirma-se isso, pois constatou-se que algumas expressões nominais eram assinaladas como o “melhor” representante das entidades salientes no discurso. Isso também, de certa forma, revela que estruturar, estabelecer relações é uma forma de atribuir consistência às escolhas para elaboração de um sumário. Conforme colocado pelo grupo de profissionais, a dificuldade de estabelecer relações de uma parte (trecho, expressão etc.) do texto com outras pode ser um indício de que algo está “solto” no todo (no texto), ou melhor, está num segundo plano de importância no discurso. Em suma, segundo os profissionais, a marcação das sentenças entre principais e secundárias, incluindo demarcação da estrutura (por exemplo, parágrafo de introdução e de fecho), é um recurso indispensável à sumarização.

As Tarefas 1 e 2 se deram em, aproximadamente, sete meses, incluindo elaboração do plano de ação, digitalização e formatação dos arquivos, envio e recebimento do material, reuniões e observações.

Com a finalidade de examinar a consistência da Tarefa 1, as sentenças tarjadas foram comparadas com as dos sumários correspondentes na busca de examinar o trânsito de sintagmas do original para o resumo. Esse procedimento está apoiado na idéia de que, como o sumário deve ser informativo, as sentenças marcadas no texto tarjado devem influenciar a edição do sumário na medida que conservam algum de seus itens lexicais. Foi comum observar, nos sumários, vestígios lexicais (sintagmas idênticos) dos trechos tarjados.

Isso é decorrente de uma peculiaridade dos sumarizadores profissionais: devido à natureza dos textos do corpus, os sumarizados procederam à edição dos sumários com o



mínimo de alteração, ou melhor, conservando o máximo de itens lexicais e construções frasais. Tal medida, conforme o grupo, preserva sensivelmente o sentido global desse gênero de texto, favorecendo o resultado final, ou seja, a produção de um sumário informativo em detrimento de um indicativo. Em consonância com essa perspectiva, este trabalho defende que cada item lexical, mesmo numa relação de sinonímia, representa uma instrução única para engendrar um sentido.

Por isso, as operações básicas de redução foram minimizadas, excetuando a de supressão, que está na base do processo de sumarização. Considera-se supressão a eliminação daquilo que se pode classificar como secundário, acessório ou redundante. Somam-se à supressão as operações de generalização e integração. Entende-se como generalização a operação que ocorre quando entidades do discurso são reunidas sob um rótulo superordenado, ou seja, relações de hiponímia. Já integração é uma transação de redução que opera por meio de *frames*, *scripts*, esquemas etc.

Voltando às Tarefas 1 e 2, destaca-se que ambas as tarefas foram sempre executadas por um mesmo membro do grupo de sumarização. Além disso, por arbítrio do grupo (procedimento padrão), os sumários foram elaborados por um outro membro do grupo e revisados por outro. Essa atitude promoveu várias modificações na edição final do respectivo sumário. Eles evitavam resumir num mesmo turno textos que tratavam de um mesmo assunto, pois, segundo eles, há chances consideráveis do processo de síntese ser prejudicado. Por exemplo, caso para um mesmo sumarizador estivesse programado 2 textos que tratassem de clonagem animal (fato que efetivamente ocorreu), cada um era trabalhado num dia. Em adição, verificou-se que, caso a Tarefa 1 ou 2 fosse interrompida, ela era retomada do início e

jamais do ponto de parada. Também registra-se que, em hipótese alguma, as Tarefa 1 e 2 foram intercaladas por outra atividade.

### **2.3.2 Geração dos extratos automáticos**

Os extratos automáticos do *Summ-it* foram gerados pelo GistSumm (PARDO, RINO e NUNES, 2003), um sumarizador extrativo projetado para mapear a idéia central de um texto. Para isso, ele combina métodos estatísticos simples. Com finalidade de simular o processo humano, ele, inicialmente, busca a sentença que “melhor expressa a idéia principal” para, então, selecionar as demais sentenças que irão fazer parte do extrato.

Pardo (2002b) reporta que as hipóteses do GistSumm para realizar essas tarefas são as seguintes: (a) pode-se identificar a sentença que apresenta a idéia central do texto-fonte aplicando métodos estatísticos simples; (b) pode-se gerar extratos coerentes a partir da justaposição das sentenças do texto-fonte que se relacionam com àquela que apresenta a idéia principal. Considera-se também que (a) pode ser confirmada caso a sentença escolhida não seja a que apresenta a idéia central, mas uma aproximação expressiva da mesma.

A partir dessas proposições, entende-se que o sumarizador automático adota uma granularidade sentencial, sendo a segmentação das sentenças orientada por algoritmos que detectam sinais de pontuação (por exemplo, ponto final e sinais de exclamação e interrogação) e letras capitulares. Juntamente com o fracionamento, ocorre a enumeração das sentenças segmentadas.

Essa enumeração é importante para um momento chamado por Pardo de “o ranqueamento das sentenças”, ou seja, a ordenação das frases a partir de seus pesos obtidos

pela aplicação dos métodos estatísticos. O ranqueamento pode ser organizado em cinco etapas – que são aplicadas a cada uma das sentenças do texto-fonte. Na primeira, as sentenças são convertidas em vetores (SALTON, 1988) cujas posições armazenam as suas palavras – um exemplo é fornecido na Figura 14. Com base nas orientações da comunidade da área (WITTEN, MOFFAT e BELL, 1994), ocorre uma segunda etapa, designada *case folding*. Nela as letras maiúsculas são alteradas para minúsculas – exemplo na Figura 14. Na terceira etapa, os termos armazenados nos vetores são substituídos por suas respectivas formas canônicas, isto é, recupera-se no léxico do sistema (NUNES et al., 1996) a forma básica de cada palavra do texto-fonte – exemplo na Figura 14. Depois, ocorre uma quarta etapa, denominada de remoção das *stopwords*, que é a desconsideração, para fins de cômputo, das palavras pertencentes a categorias fechadas, tais como preposições, artigos etc. A essa etapa estão associados três passos: i) as palavras iguais são unificadas em uma única posição do vetor (precisamente, a da primeira ocorrência do termo); ii) a frequência de cada palavra no vetor é armazenada junto às próprias palavras; iii) a frequência das *stopwords* é zerada. A Figura 14 fornece um exemplo de um vetor atualizado por cada processo. Segundo Pardo, esses processos não só facilitam o processamento computacional posterior como também apuram os resultados de métodos de sumarização estatística.

Na quinta etapa, é realizada a atribuição de pesos para as sentenças do texto-fonte por meio do método de palavras-chave (BLACK e JOHNSON, 1988). Esse segue a proposta de Luhn (1958), descrito na Seção 1.4, em que os pesos resultam do somatório das ocorrências de cada termo no texto-fonte. Por exemplo, se for considerado que, em certa sentença de um texto-fonte, a palavra *Santanaraptor* surge quatro vezes e as palavras *espécie*, *origem* e

*tiranossauro* ocorrem uma vez, a pontuação total dessa sentença se dará da seguinte forma:

$$(5*1) + (1*3) = 8.$$

Vetorização										
O	Santanaraptor	pode	ser	a	espécie	que	deu	origem	ao	tiranossauro
Case folding										
o	santanaraptor	pode	ser	a	espécie	que	deu	origem	ao	tiranossauro
Substituição por formas canônicas										
o	Santanaraptor	poder	ser	o	espécie	que	dar	origem	ao	tiranossauro
Remoção de <i>Stopwords</i>										
o	Santanaraptor	poder	ser	o	espécie	que	dar	origem	ao	tiranossauro
0	1	1	1	0	1	0	1	1	0	1

**Figura 14 – Tratamento das sentenças no momento de ranqueamento**

Nesse método, a sentença com a maior pontuação é selecionada como aquela que expressa a idéia principal do texto-fonte. A partir dela e do valor da taxa de compressão, as demais sentenças são recolhidas do texto-fonte, a fim de compor o extrato. Esse processo de seleção é regido pelos seguintes passos:

- i. Calcula-se a média da pontuação das sentenças do texto-fonte e declara-se essa média como sendo um *cutoff*, nota de corte para eliminar sentenças do texto-fonte;
- ii. Identificam-se as sentenças que apresentam pelo menos uma palavra cuja forma canônica coincida com uma das formas canônicas da sentença escolhida como aquela que expressa a idéia principal do texto-fonte;

- iii. Selecionam-se entre as sentenças elencadas no item (ii) as que possuam maior pontuação de *cutoff*.

Com a finalidade de respeitar a taxa de compressão especificada pelo usuário, o GistSumm pode excluir do conjunto do item (iii) as sentenças com menor pontuação, mesmo acima do *cutoff* definido.

O GistSumm, em adição, contempla técnicas de *Text Mining* e de localização das sentenças no texto – descritas na Seção 1.4. Por utilizar, além desses recursos, métodos estatísticos simples, o GistSumm pode ser facilmente adaptado para qualquer língua ocidental, desde que se ajuste seu léxico eletrônico e se modelem questões referentes ao gênero textual. Mais detalhes sobre o sistema pode ser encontrado em Pardo (2002b) e Pardo, Rino e Nunes (2003).

A par dos aspectos técnicos do GistSumm, iniciou-se, propriamente, a geração dos sumários automáticos, ou seja, para cada um dos 50 textos do corpus *Summ-it*, foi gerado um extrato automático. Eles foram produzidos com o método de palavras-chave numa taxa de compressão de 70%. Em harmonia com a comunidade científica da área, os textos que serviram de entrada para o sumarizador são arquivos ASCII do *Summ-it*, ou seja, arquivos textos sem títulos, subtítulos e outros dados bibliográficos (Seção 2.2.1).

De posse dos sumários gerados pelo GistSumm, no próximo capítulo, focaliza-se nos problemas de coerência acarretados pela falta de uma etapa de mapeamento dos relacionamentos anafóricos e de correferência.

### 3 SUMARIZAÇÃO, CORREFERÊNCIA E COESÃO REFERENCIAL

Com vistas a apresentar a principal aplicação para o corpus *Summ-it*, é reportado aqui um estudo que faz parte do projeto ProCaCoSa<sup>38</sup>, que conta com a parceria da NILC/UFSCar, responsáveis pelos trabalhos específicos da área de Sumarização Automática, e do Laboratório de Engenharia da Linguagem da UNISINOS, com experiência em trabalhos na área de Resolução Anafórica<sup>39</sup>. Nesse contexto, este estudo visa rastrear e diagnosticar alguns problemas desencadeados pela ocorrência de cadeias de correferência não resolvidas durante a seleção e estruturação do conteúdo de sumários gerados de modo automático. Cabe lembrar que, aqui, se emprega a denominação *cadeia de correferência* para designar uma série de eventos lingüísticos em que se observam relações de identidade, associação e/ou dependência referencial.

Mais precisamente, este estudo investiga, na Sumarização Automática, a viabilidade de procedimentos de pré e pós-edição baseados em informações sobre as cadeias de

---

<sup>38</sup> Processamento de Cadeias de Correferência para a Sumarização Automática de Textos em Português

<sup>39</sup> A subárea de Resolução Anafórica, ou Resolução de Anáforas, compreende o desenvolvimento de aplicações computacionais que mapeiem automaticamente relações anafóricas, de correferência e casos de pronomes zero (elipses).

correferência. Logo, tem-se o objetivo de averiguar se as informações sobre correferência e anaforicidade podem beneficiar os sistemas de sumarização automática.

O uso de conhecimento anafórico e de correferência não é novidade na Sumarização Automática – conforme a Seção 1.4. Alguns trabalhos utilizam esse conhecimento na pós-edição dos sumários, ou seja, após a aplicação de determinado algoritmo de sumarização, é efetuado um segundo processo para aperfeiçoar o resultado do primeiro baseado em informações de ordem correferencial e anafórica. Por exemplo, estudos mostram que sumários que contêm expressões anafóricas sem antecedentes acarretam, por via de regra, resumos incoerentes (MANI e MAYBURY, 1999). Outros trabalhos investem no uso do conhecimento sobre correferência e anáfora na pré-edição dos sumários, isto é, a informação sobre a coesão referencial é usada para discriminar elementos principais de secundários (AZZAM et al., 1999; CRISTEA et al., 2003).

O caráter inovador do trabalho reportado aqui está no uso das informações sobre as cadeias de correferência: as informações são utilizadas em sua totalidade, isto é, tanto os vínculos internos (aqueles que “dão forma” a cadeias de correferência) quanto os vínculos externos (aqueles pontos de contato entre cadeias – tais como, relações associativas). Investe-se nas informações sobre cadeias de correferência, pois se sabe da diversidade de aplicações computacionais que podem gerar de modo automático grande parte dessas informações – exemplo disso é o DAARC’2007<sup>40</sup>.

---

<sup>40</sup> O *Discourse Anaphora and Anaphor Resolution Colloquium* – DAARC’2007 está disponível em: <http://daarc2007.di.fc.ul.pt/>

Logo, se coloca como objetivo imediato fornecer subsídios para enriquecer os modelos de Sumarização Automática baseados em informação anafóricas e de correferência. Para isso, foram analisados sumários produzidos automaticamente, a fim de detectar problemas de coerência desencadeados pela ausência de resolução anafórica e de correferência. Em adição, foi verificado como as informações sobre as cadeias de correferência podem auxiliar na pré e pós-edição de sumários automáticos.

### **3.1 Estudo dos sumários automáticos sob a ótica das cadeias de correferência**

Conforme comentado na Seção 1.4, ainda que os métodos superficiais apresentem resultados interessantes para determinados gêneros textuais, há uma margem considerável desses métodos gerarem sumários incoerentes e/ou desconexos, pois freqüentemente essas técnicas não contam com recursos de rescrita dos segmentos extraídos, elas, para compor o sumário, apenas justapõem as sentenças extraídas, desconsiderando a coesão do texto-fonte, com efeito, os elos coesivos por correferência e anáfora.

Na busca de antecipar os problemas de interpretação de ordem correferencial e anafórica, este estudo concentrou sua atenção em como as cadeias de correferência foram afetadas pelo processo automático de sumarização.

Para esse experimento foram utilizados os recursos disponibilizados pelo *Summ-it*, notadamente, a anotação de correferência e anáforas (Seção 2.2.3) e o conjunto de sumários (Seção 2.3).



O método de exame dos sumários automáticos sob a ótica das cadeias de correferência pode ser formalizado em cinco passos. Num primeiro momento, efetuou-se a leitura atenta dos extratos automáticos. Na seqüência, foram rastreados possíveis problemas de interpretação, coerência e/ou coesão, sendo cada um deles assinalados – foi considerado um problema qualquer distúrbio que comprometa a qualidade do sumário automático, lembrando que o resultado automático esperado é, sempre, um sumário informativo (conforme a Seção 1.4).

Num terceiro momento, todos os casos assinalados anteriormente foram examinados minuciosamente, a fim de isolar os problemas de natureza referencial (correferência e anafórica) de outros problemas – por exemplo, aqueles de elos de junção (*transitional ties*<sup>41</sup>).

Uma vez isolados os distúrbios de ordem referencial, foi disparado um quarto passo: verificou-se se esses problemas são decorrentes do processo automático de sumarização ou se eles já se encontravam no texto-fonte correspondente. Em outras palavras, comparou-se o comportamento das cadeias de correferência do texto-fonte com o novo arranjo referencial (anafórico e de correferência) do sumário. Além da leitura atenta do texto-fonte, essa investigação contou com os relatórios HTML da anotação das cadeias de correferência (Seção 2.2.3), ilustrado no Anexo 3.

Depois de detectados, isolados e confirmados, os problemas de ordem referencial (anafórico e correferencial) foram submetidos a um agrupamento por similaridades, isto é, casos parecidos foram reunidos em grupos.

---

<sup>41</sup> Esses são os elos coesivos que ligam as orações entre si. Para Halliday e Hasan (1976) esta categoria (*conjunction*, na sua terminologia) abrange fatores de conectividade desempenhados por advérbios, conjunções e expressões preposicionais. Em acordo com os estudos de Frederick Crew (1986), que declaradamente se fundamentam em Halliday e Hasan, os elos coesivos dessa natureza pode ser classificados em: *consequence*, *likeness*, *contrast*, *amplification*, *example*, *concession*, *insistence*, *sequence*, *restatement*, *recapitulation*, *time* ou *place*.

No primeiro, foram reunidos os casos de referência adicional (*additioning reference*)<sup>42</sup>, ou seja, a adição de uma nova entidade sem que se encontrem, na expressão referencial, na cadeia de correferência ou no texto (sumário), pistas suficientes para a sua interpretação. Geralmente, nesses casos, a carência informacional chega às raias da não compreensão da sentença, quando não do próprio sumário. Em textos naturais, essa deficiência reflete uma falha na quantidade de informação compartilhada. Nas amostras gerados pelo GistSumm, sob a luz das informações sobre as cadeias de correferência, seguramente, se pode dizer que esse problema decorre da poda de expressões representativas da cadeia de correferência que engendram a respectiva entidade do discurso.

Esse distúrbio, no corpus, desencadeia graves problemas de interpretação, como pode ser percebido em (9). Nele é possível apontar pelo menos uma situação de referência adicional: expressão *a solução*. Ela não oferece recursos para engendrar uma entidade no discurso, pelo menos, não uma entidade que seja funcional ao sumário. É evidente que o sintagma nominal *a solução* é aceitável sintaticamente e, de certo modo, semanticamente, porém, discursivamente, ele agride, no mínimo, os fatores de informatividade e relevância (ver Seção 1.2.1), por isso, o item de referência *a solução* é compreendido como não funcional. Assim como no exemplo, os piores casos de referência adicional são os que envolvem anáforas encapsuladoras, sendo esses seguidos pelos epítetos, pronomes indefinidos, possessivos e pessoais do caso oblíquo (FINE, 1994). Em (9), cabe notar que a posição do distúrbio – início do sumário – pode, também, configurar outro agravante para a

---

<sup>42</sup> Nas palavras de Jonathan Fine (1994, p. 221-2): “If a participant is not specifically introduced but the speaker later attempts to use a personal or demonstrative reference to identify that participant, then uninterpretable additioning results. [...] Additioning references [...] are uninterpretable since speakers improperly assume that participants have been introduced or are retrievable by inference when they are not.”

interpretação do extrato, ele pode estimular o leitor a uma postura de desprezo, ou melhor, não aceitar o texto como coerente e coeso (Seção 1.2).

( 9 ) A solução foi sugerir um modelo alternativo em que, nos primórdios da formação de um sistema planetário, instabilidades gravitacionais no disco de gás e poeira já induziriam o surgimento dos envelopes dos gigantes gasosos numa velocidade pelo menos dez vezes maior.<sup>43</sup>

No segundo grupo, foram agrupados os casos de referência incerta (*unclear reference*), isto é, a presença de um item de referência com baixa carga informacional – nesses casos, a expressão referencial não deixa claro “a identidade” da entidade. Ao contrário da referência adicional, a referência incerta oferece na expressão referencial, na cadeia de correferência ou no texto (sumário), pistas suficientes para a sua “parcial” interpretação. Geralmente, quando parte da cadeia de correferência é preservada pela sumarização automática, um caso de referência incerta pode engendrar a entidade do discurso por catáfora<sup>44</sup>.

Pelo menos uma situação de referência pode ser detectada em (10): *o astro*. Embora a expressão *o astro* não deixe claro a identidade da entidade abordada no texto-fonte correspondente, é possível esboçar um “contorno” para a referência evocada, ou seja, é possível atribuir um significado relevante para o sentido global do texto. Apesar de *o astro* não apresentar o grau de informatividade da entidade engendrada pelo texto-fonte, seu significado atende, mesmo que parcialmente, os fatores de coerência, relevância e aceitabilidade, por isso, se considera uma situação de referência incerta menos nociva do que

<sup>43</sup> Sumário correspondente ao texto CIENCIA\_2002\_6441, corpus Summ-it, gerado automaticamente pelo GistSumm, método palavras-chave, taxa de compressão de 70%.

<sup>44</sup> Em contraste com anáfora, a catáfora advém da propriedade da linguagem de “levar adiante, para frente” a apresentação “plena” de uma entidade do discurso.

uma de referência adicional para a Sumarização Automática. Em textos naturais, geralmente, os casos de referência incerta são consequência de lapsos no desenvolvimento de um tópico discursivo ou na introdução de um comentário (AITCHINSON, 1987). Entretanto, ao comparar os casos presentes nos 50 sumários automáticos com as respectivas informações sobre as cadeias de correferência dos textos-fonte, pode-se afirmar que, de modo similar aos casos de referência adicional, as situações de referência incerta ocorrem pela exclusão de expressões significativas da cadeia de correferência que engendram a entidade correspondente.

( 10 ) Antes que algum espertalhão resolva aproveitar a notícia para dar uma de Orson Welles (que aterrorizou Nova York em 1938 ao anunciar pelo rádio uma suposta invasão marciana, numa encenação da obra "Guerra dos Mundos", de H.G. Wells), é bom avisar: o **astro** vai passar a uma distância do Sol \_ cerca de mil anos-luz\_ que é até pequena, se comparada às dimensões da Via Láctea, mas não o suficiente para causar dano à Terra ou a qualquer corpo do Sistema Solar.<sup>45</sup>

Uma situação de ordem correferencial não contabilizada como um problema para a interpretação dos resumos foi a repetição de instruções lingüísticas. Ao observar o arranjo correferencial dos sumários em relação ao comportamento das cadeias de correferência do texto-fonte correspondente, descobriu-se que 8% dos sumários examinados apresentam apenas relações diretas (de correferência ou anafórica), isto é, as relações indiretas foram podadas pelo processo automático de síntese. Assim como ocorre nos textos naturais, quando

---

<sup>45</sup> Sumário referente ao texto CIENCIA\_2002\_22015, corpus Summ-it, gerado automaticamente pelo GistSumm, método de palavras-chave, taxa de compressão de 70%.

essa situação é detectada no resultado automático, o texto (sumário) torna-se “pobre” lexicalmente, deixando a desejar no que toca a construção das entidades evocadas, já que elos coesivos indiretos são preteridos a favor da repetição sistemática dos mesmos lexemas.

Excetuando essa situação, registrou-se no exame dos 50 resumos gerados pelo GistSumm o seguinte resultado: aproximadamente 76% dos sumários apresentam algum problema de ordem referencial (de correferência ou anafórica) que agrediu a interpretação (coerência) do resumo automático.

A análise que gerou esses resultados permite dizer que as cadeias de correferência podem ser empregadas tanto na pré-edição quanto na pós-edição dos sumários gerados pelo GistSumm. Na pré-edição, as informações sobre as cadeias poderiam orientar as escolhas para composição do sumário, por exemplo, todas as manifestações lingüísticas de uma mesma entidade poderiam ser computadas com um valor único, isto é, não só os relacionamentos diretos mas também os indiretos poderiam receber um mesmo peso para a busca da(s) sentença(s) que “melhor expressa(m) a idéia principal”. Outra hipótese de uso das informações sobre relacionamentos anafóricos e de correferência que o experimento com os extratos gerados pelo GistSumm oportunizou é considerar como mais salientes as sentenças em que se detecta um trânsito maior de cadeias de correferência, ou seja, aquelas em que se localiza um maior número de instruções lingüísticas envolvidas em algum relacionamento interfrásico de correferência ou anafóricos (incluindo os associativos).

Na pós-edição, uma sugestão de melhoria seria incluir, na sumarização automática, uma etapa de comutação de expressões referenciais “pobres” lexicalmente. Em outras palavras, os resumos automáticos poderiam ser tratados com informações sobre as cadeias de

correferência: depois de gerados os extratos, todas as expressões nominais poderiam ser consultadas na anotação sobre correferência e anaforicidade para checar se há na corrente de instruções lingüísticas uma expressão mais representativa da entidade evocada.

Trabalhos recentes com cadeias lexicais em Sumarização Automática (por exemplo, STEINBERGER, POESIO, JEZEK, 2007) orientam selecionar a primeira ocorrência da cadeia para substituição. Entretanto, nos sumários automáticos examinados, observou-se que a melhor operação seria selecionar para substituição o sintagma nominal mais completo lexicalmente e não necessariamente o primeiro elemento da cadeia. Em outras palavras, a melhor escolha resulta de um balanço entre a informatividade e a concisão, por exemplo, um item de referência intermediário da cadeia de correferência. Numa simulação realizada com os extratos automáticos, verificou-se que a substituição pelo primeiro item da cadeia de correferência não solucionou a maioria dos distúrbios de ordem referencial. Todavia, a comutação pelo item com o maior número de itens lexicais da cadeia de correferência foi, na grande maioria dos casos com problemas, o reparo necessário. Ao entender que há uma dificuldade considerável em avaliar computacionalmente os casos que necessitam de tratamento, a pós-edição sugerida aqui poderia se configurar como um processo de exame e tratamento para problemas relativos à coesão referencial. Em outras palavras, pode-se partir do princípio de que todas as expressões nominais dos sumários são passíveis de problemas de ordem referencial, logo, todas deveriam ser confrontadas com suas respectivas cadeias de correferência, a fim de verificar se elas são as melhores representantes do fluxo de correferência (anafórico) da entidade evocado, em caso negativo a substituição deveria ser realizada.

Para ilustrar esse procedimento, retoma-se o exemplo (10), precisamente a expressão *o astro* que configura, ao lado de *a notícia*, uma situação que compromete a interpretação do sumário automático – via coesão referencial.

Primeiramente, ao identificar a expressão *o astro* como um sintagma nominal (anotação gramatical – Seção 2.2.2) do extrato automático, verifica-se se esse item de referência pertence a uma das cadeias de correferência do texto-fonte correspondente. Em caso negativo, move-se a conferência para a próxima expressão. Em caso positivo, percorre-se a respectiva cadeia de correferência (Figura 15), checando se a expressão em questão (*o astro*) é o melhor representante da cadeia – tendo como parâmetro de “melhor” nesse caso o maior número de itens lexicais. Em caso positivo, é mantida a expressão em questão e inicia todo o processo na próxima expressão. Em caso negativo, substitui-se a expressão pela melhor representante da cadeia. Em (10), a expressão *o buraco negro GRO-J 1655-40* é colocado como substituto de *o astro*. O resultado desse processo de tratamento é apresentado em (11).

um deles
o astro
esse objeto
o buraco negro GRO-J 1655-40
ele
o GRO-J 1655-40
aquele astro
o buraco negro
o astro

**Figura 15 – Exemplo de cadeias de correferência**

( 11 ) Antes que algum espertalhão resolva aproveitar [...], é bom avisar: **o buraco negro GRO-J 1655-40** vai passar a uma distância do Sol \_ cerca de mil anos-luz\_ que é até pequena, se comparada às dimensões da Via Láctea, mas não o suficiente para causar dano à Terra ou a qualquer corpo do Sistema Solar.

Nesse exemplo, é possível reconhecer a importância de não se fixar um candidato para substituição como, por exemplo, sempre o primeiro item da cadeia – no exemplo anterior: *um deles*. Em (11), percebe-se os benefícios operar com um critério de permuta mais flexível, como a escolha de um nome próprio ou a instrução lingüística com mais itens lexicais.

Ao término deste estudo, nota-se que as informações sobre correferência e anáforas podem ser úteis para Sumarização Automática. Com efeito, os recursos de resolução anafórica que contemplam as cadeias de correferência podem auxiliar, na sumarização extrativa, a realização lingüística (pós-edição) e a seleção das unidades (pré-edição). Mesmo com limitações, os atuais sistemas de resolução anafórica podem mapear satisfatoriamente cadeias de correferência e anáforas. Para língua portuguesa, experimentos iniciais, com as etiquetas semânticas do PALAVRAS, já foram realizados, visando esse tipo de mapeamento de modo automático (COELHO et al., 2006; VIEIRA et al., 2006). Há, também, alguns trabalhos de classificação automática entre expressões correferentes e não correferentes que podem ser facilmente adaptados para a obtenção de uma representação correferencial e anafórica em forma de cadeia (COLLOVINI, COELHO e VIEIRA, 2005; COLLOVINI e VIEIRA, 2006).



## CONSIDERAÇÕES FINAIS

Este trabalho apresentou o corpus *Summ-it* com seus diferentes níveis de anotação e seus sumários. Ele, constituído de 50 textos do corpus PLN-BR, foi anotado, seguindo padrões internacionais para codificação de corpus, tais como o modelo de marcação *stand-off* e a representação em XML.

O primeiro nível de anotação foi representado em XCES. Ele reúne dados bibliográficos (títulos, autoria, palavras-chave, datação etc.) e relativos à organização dos parágrafos e das sentenças dos textos (como um índice das sentenças).

Para geração do segundo nível de anotação (informações lexicais) e terceiro nível de anotação (informações sintáticas) foram utilizados os programas PALAVRAS (BICK, 2000) e Palavras Xtractor (GASPERIN et al., 2003). Nesses níveis de marcação, é possível encontrar de informações flexionais a informações semânticas (concreto, abstrato, contável, comestível etc).

No quarto nível de anotação, há informações sobre a estrutura retórica dos textos, empregando para isso a *Rhetorical Structure Theory* - RST (MANN e THOMPSON, 1987) e a ferramenta RSTTool (O'DONNELL, 2000).

A partir da experiência de diferentes projetos, ANACORT<sup>46</sup>, COMMON-REFs<sup>47</sup>, TeXto<sup>48</sup>, CROWS<sup>49</sup> e VENEX<sup>50</sup>, foi elaborada a anotação com informações sobre correferência, anaforicidade e referências dêiticas – o quinto nível de anotação do corpus *Summ-it*. Esse nível de anotação contou com a MMAX (MÜLLER e STRUBE, 2001), ferramenta específica para anotação de corpus, sendo seu principal enfoque a anotação de correferência. O esquema de anotação, em linhas gerais, seguiu os seguintes passos: seleção das unidades de interesse (*markables*), identificação das suas configurações lexicais, construção das cadeias de correferência e anáforas, classificação dos *markables* e de seus relacionamentos sob a perspectiva das cadeias de correferência.

Além disso, o corpus *Summ-it* conta com sumários manuais e automáticos. Para cada texto do corpus foi elaborado um resumo por um sumarizador profissional (sendo esse validado por outro) e foi gerado um extrato pela ferramenta GistSumm (PARDO, RINO e NUNES, 2003). Em adição, foi produzido, manualmente, um texto tarjado para cada texto do corpus. Para construção dos textos tarjados, foi pedido ao sumarizador que **i)** marcasse a(s) sentença(s) que lhe indicam a idéia principal de cada texto-fonte, isto é, aquela(s) que lhe dão a diretriz para a elaboração do respectivo sumário, e que **ii)** selecionasse entre as marcadas uma que seja a mais representativa da idéia principal do texto (no caso, de haver apenas uma sentença marcada em **i** essa passa automaticamente a atender esse pedido).

---

<sup>46</sup> Anotação automática de co-referência textual.

<sup>47</sup> Um modelo computacional unificado para o tratamento de referências.

<sup>48</sup> Acesso a informações em bases textuais.

<sup>49</sup> Construção de ontologias para a Web Semântica.

<sup>50</sup> Projeto colaborativo entre a Università' di Venezia e a Essex University.

De posse do corpus *Summ-it* anotado e seus sumários, foi realizada a principal atividade deste trabalho de conclusão (tal como previsto pelo projeto ProCaCoSa<sup>51</sup>): o rastreamento e diagnóstico de alguns problemas desencadeados pela ocorrência de cadeias de correferência não resolvidas durante a seleção e estruturação do conteúdo de sumários gerados automaticamente. Em suma, esse exame mostra que dos 50 extratos gerados pelo GistSumm 76% apresentam algum problema de ordem referencial (de correferência ou anafórica) que agrediu a interpretação (coerência) do extrato automático. Seja por adição (*additioning reference*) ou pela falta de clareza (*unclear reference*), observou-se que os distúrbios (co)referenciais ocorrem, nos extratos examinados, freqüentemente, pela exclusão de expressões significativas de uma cadeia de correferência. Isso foi, particularmente, agravado em situações que envolviam anáforas encapsuladoras, já que, esse fenômeno está, freqüentemente, associado à organização tópica (KOCH, 2002). Com efeito, a carência informacional, nos casos de referência adicional, chegou às raias da incompreensão do sumário automático. Embora menos nocivos, os casos de referência incerta também acarretaram problemas consideráveis de interpretação quando somados a outros tipos de problemas, como aqueles relativos aos elos de junção (*transitional ties*).

Por isso, se recomenda adicionar aos processos de sumarização automática que adotam a abordagem empírica uma etapa de resolução anafórica. Experimentos iniciais com as etiquetas semânticas do PALAVRAS já foram realizados (COELHO et al., 2006; VIEIRA et al., 2006). Há, também, alguns trabalhos de classificação automática entre expressões correferentes e não correferentes (COLLOVINI, COELHO e VIEIRA, 2005; COLLOVINI e VIEIRA, 2006) que podem orientar os processos automáticos de sumarização.

---

<sup>51</sup> Processamento de Cadeias de Correferência para a Sumarização Automática de Textos em Português

Em especial, esse estudo permitiu observar que as cadeias de correferência podem ser empregadas tanto na pré-edição quanto na pós-edição dos sumários automáticos. Na pré-edição, as informações sobre as cadeias podem nortear as escolhas para composição do sumário, por exemplo, todas as manifestações lingüísticas (co)referentes a uma mesma entidade poderiam receber um mesmo peso para a busca da(s) sentença(s) que “melhor expressa(m) a idéia principal”. Também é possível considerar como mais salientes as sentenças em que se detecta um trânsito maior de relacionamento interfrásico de correferência ou anafóricos (incluindo casos associativos). Já, na pós-edição, os resumos automáticos podem ser tratados com informações sobre as cadeias de correferência, ou seja, depois de gerados os extratos, todas as expressões nominais poderiam ser consultadas na anotação sobre correferência e anaforicidade para checar se há na corrente de instruções lingüísticas uma expressão mais representativa da entidade evocada.

Em continuidade ao que foi exposto neste trabalho, está prevista uma dissertação de mestrado em Computação Aplicada que automatizará o processo de revisão de coerência pela verificação das cadeias de correferência. Em adição, está em desenvolvimento um trabalho de conclusão da Ciência da Computação para resolução automática de correferência, empregando os recursos apresentados aqui. Também se pretende expandir, em uma dissertação de mestrado em Lingüística Aplicada, a elaboração de heurísticas de resolução anafórica baseadas na anotação semântica do PALAVRAS associada às relações de correferência direta (cujo tratamento automático já ocorre de modo satisfatório) e/ou o estudo das cadeias de correferência e anáforas no contexto de sumarização multidocumentos, principalmente, nas tarefas que envolvam a pré-edição.

Além disso, dois experimentos estão programados envolvendo os resultados deste trabalho: um comparativo dos métodos apresentados aqui de pré-edição com os baseados na *Veins Theory* (CRISTEA, IDE e ROMARY, 1998) e a modelagem de uma aplicação computacional que trate as referências dêiticas a partir das informações codificadas nos arquivos XCES, que contêm informações sobre o local, datação etc.

Cabe destacar que, evidentemente, há uma série de trabalhos referentes à anotação RST que já estão em andamento pela equipe do NILC/UFSCar e que, em breve, atuarão ao lado dos recursos baseados nas informações sobre as cadeias de correferência e anáforas.

Concomitantemente a esses estudos e implementações, serão adicionados outros níveis de anotação ao corpus *Summ-it*. Em incubação, há a anotação semântica *FrameNet* (FILLMORE e BAKER, 2001)<sup>52</sup> sob a coordenação da Dr<sup>a</sup> Rove Luiza de O. Chishman.

Apesar de pequeno, o *Summ-it* é rico em anotação lingüística, por isso, como produto final (que, certamente, será estendido), ele será disponibilizado pelo portal do projeto PLN-BR.

---

<sup>52</sup> O *FrameNet* é um recurso lexical de grande porte que divide as palavras em *frames*: unidades que descrevem um significado independentemente da sua categoria morfossintática. Cada *frame* reflete um predador e os seus possíveis papéis semânticos como argumentos, que são chamados de *Frame Elements*. Além disso, a *FrameNet* indica um conjunto de palavras que compartilham aquela estrutura semântica especificada, as chamadas de *lexical units*. Ele também especifica relações semânticas de subframe, herança, uso etc. entre *frames*. Essas e outras características do *FrameNet* determinam a riqueza e conseqüentemente o grau de utilidade do recurso em diversas tarefas.

## REFERÊNCIAS BIBLIOGRÁFICAS

AITCHINSON, Jean. *Words in the mind: an introduction to the mental lexicon*, Oxford: Basil Blackwell, 1987.

ALLEGRI, R. F.; HARRIS, P. e DRAKE, M. Evaluación neuropsicológica en las demencias. *Revista Neurológica Argentina*, Bueno Aires, v. 24, p.11-15, 2006.

AZZAM, S.; HUMPHREYS, K.; GAIZAUSKAS. R. Using coreference chains for text summarization. *Proceedings of the workshop on coreference and its applications*, New Brunswick: Association for Computational Linguistics, p.72-89, 1999.

BASTOS, L. K. *Coesão e coerência em narrativas escolares*. São Paulo: Fontes, 1994.

BAXENDALE, P. B. Machine-made index for technical literature – an experiment. *IBM Journal of research and development*, Washington, v.2, p. 354-361, 1958.

BEAUGRANDE, R. A. de e DRESSLER, W. U. *Introduction to text linguistics*. Londres: Longman, 1981.

BEAUGRANDE, R. *Text, discourse and processo: a multidisciplinary science of texts*. Londres: Longman, 1980.

BICK, E. *The parsing system PALAVRAS: automatic grammatical analysis of portuguese in a constraint grammar framework*. Arhus University: Arhus, 2000.

BLACK, W. J.; JOHNSON, F. C. A practical evaluation of two rule-based automatic abstraction techniques. *Expert systems for information management*, Manchester, v.3, p. 33-52, 1988.

BOGURAEV, B. e KENNEDY, C. Salience-based content characterisation of text documents. *Proceedings of the intelligent scalable text summarization workshop - ACL/EACL'97*, Madri, p. 2-9, 1997.

BROWN, G.; YULE, G. *Analisis del discurso*. Madri: Visor, 1983.

BRUN, A.; ENGLUND, B.; GUSTAFSON, L.; PASSANT, U.; MANN, D. M. A.; NEARY, D. e SNOWDEN, J. S. Clinical and neuropathological criteria for frontotemporal dementia. *Journal of neurological and neurosurgical psychiatry*, Londres, v.57, p.416-418, 2006.

CARLSON, L. e MARCU, D. *Discourse tagging reference manual*. Technical Report ISI-TR-545, 2001.

CHAROLLES M. Introduction aux problèmes de la cohérence des textes. *Langue française*, Paris: Lh, v. 38, p.7-42. 1978.

CHAROLLES M. Towards an heuristic approach to text-coherence problems. *Coherence in natural language texts*, Hamburgo: Buske, v. 1, p. 45-98, 1983.

CHAROLLES M. Coherence as a principle in the regulation of discursive production. *Connexity and coherence, analysis of text and discourse*, Berlin: Gruyter, p.3-16, 1989.

CLARK, H. H. Bridging. *Thinking: readings in cognitive science*. Londres: Cambridge University Press, p.56-70, 1977.

CLOCKSIN, W. e MELLISH, C. Programming in prolog. *Springer-Verlag*, Berlim, p. 144-149, 1981.

COELHO, J. C. B.; COLLOVINI, S. C. e VIEIRA, R. Estudo de corpus para classificação de expressão anafóricas da língua portuguesa. *3º Workshop em tecnologia da informação e da linguagem humana (TIL'2005)*, São Leopoldo: UNISINOS, 2005.

COELHO, J. C. B.; MULLER, V. M.; COLLOVINI, S. C.; VIEIRA, R.; RINO, L. H. M. Resolving nominal anaphora. 7<sup>th</sup> *Workshop on the computational processing of written and spoken portuguese* - PROPOR, 2006, Itatiaia. Lecture Notes in Artificial Intelligence. Berlin: Springer, v. 3960, p.160-169, 2006.

COLLOVINI, S.; COELHO, J.C.B.; VIEIRA, R. Classificação automática de expressões anafóricas em textos da língua portuguesa. *Anais do encontro nacional de inteligência artificial* – ENIA, 2004. São Leopoldo, p.57-64, 2005.

COLLOVINI, S.; VIEIRA, R. Learning portuguese discourse-new references. *IFIP 19<sup>th</sup> World computer congress*, 2006, Santiago do Chile. Artificial intelligence in theory and practice - IFIP 19<sup>th</sup> World computer congress, TC-12 IFIP AI 2006 Stream. Berlin: Springer, v. 217, p. 267-276, 2006.

COLLOVINI, S.; CARBONEL, T; FUCHS, J. T.; COELHO, J. C. B.; RINO, L e VIEIRA, R. Summ-it: um corpus anotado com informações discursivas visando à sumarização automática. *5<sup>o</sup> Workshop em tecnologia da informação e da linguagem humana* – TIL, 2007, Rio de Janeiro: IME, 2007.

CREW, F. Cohesion: inter and intra sentence errors. *Journal of learning*, v.22, nº. 8, p. 338-341, 1986.

CRISTEA, D.; IDE, N. e ROMARY, L. Veins theory: a model of global discourse cohesion and coherence. *Proceedings of the coling/ACL' 1998*, Montreal, p.281-285, 1998.

CUMMINGS, J. L. e BENSON, D. F. *Multidisciplinary clinical approach*. 2<sup>a</sup> ed., Stoneham: Butterworth Heinemann, 2005.

EARL, L. L. Experiments in automatic abstracting and indexing. *Information storage and retrieval*, Bruxelas, n.6, p. 313-334, 1970.

EDMUNDSON, H.P. New methods in automatic extracting. *Journal of the ACM*, Ottawa, n.16, p.264-285, 1969.

ELHADAD, M. FUF: the universal unifier user manual, version 5.0. *Journal department of computer science – columbia university*, Nova York, n.12, v.1, 1991.



FÁVERO, L.L. *Coesão e coerência textuais*. São Paulo: Ática, 2000.

FILLMORE, C.J. e BAKER, C.F. Frame semantics for text understanding. *Proceedings of Word-Net and Other Lexical Resources Workshop*, NAACL, Petersburgo, p 59-64, 2001

FINE, J. *How language works: cohesion in normal and nonstandard communication*. Norwood/New Jersey: Ablex Publishing Corporation, 1994.

FRAURUD, K. Definiteness and the processing of NPs in natural discourse. *Journal of Semantics*, v.7, p. 395-433, 1990.

GASPERIN, C.; VIEIRA, R.; GOULART, R.; QUARESMA, P. Extrating XML syntactic chunks from portuguese corpora. *Traitement automatique des langues minoritaires - TALN*, Btaz-sur-mer, p. 84-96, 2003.

GIORA, R. A text-based analysis of nonnarrative texts. *Theoretical linguistics*, v.12, p.115-135, 1985.

GREGG, N. Cohesion: inter and intra sentence errors. *Journal of learning disabilities*, v. 19, nº 6, Jun/Jul, p. 338-341, 1986.

GRICE, H.P. (1975). *Logic and conversation*. *Syntax and semantics 3: speech acts*. Nova York: Academic Press, 1975.

GROSZ, B. e SIDNER, C. Attention, intentions, and the structure of Discourse. *Computational linguistics*, Viena, v. 12, No. 3, 1986.

GROSZ, B.J. Focusing and description language dialogues: elements of understanding. Proc. of the 21<sup>st</sup> Annual Meeting of the Assoc. para *Computational Linguistics*, Cambridge, Mass., 1981.

HALLIDAY M A K e HASAN, R. *Cohesion in english*. Londres: Longman, 1976

HALLIDAY, M.A.K. *An introduction to functional grammar*. Londres: Edward Arnold, 1985.

HAWKINS, J. A. *Definiteness and indefiniteness: a study in reference and grammaticality prediction*. Londres: Croom Helm, 1978.

HEARST, M. A. Untangling text data mining. *Proceedings of ACL '99, the 37<sup>th</sup> annual meeting of the ACL*, Maryland, v.1, p.387-395, 1999.

HODGES, J. R., SALMON, D. P. e BUTTERS, H. Semantic memory impairment in Alzheimer Disease: Failure of access or degraded knowledge? *Neuropsychologia*, Paris, v. 79, p. 301-324, 2006.

HUTCHINS, J. Summarization: some problems and methods. *The frontier of informatics*, Amsterdã, p. 151-173, 1987.

IDE, N.; BREW, C. Requirements, Tools, and Architectures for Annotated Corpora. *Proceedings of data architectures and software support for large corpora*. Paris: European Language Resources Association, p.1-5, 2000.

KAPLAN, R.M. e BRESNAN, J. Lexical-functional grammar: a formal system for grammatical representation. *The mental representation of grammatical relations - cambridge editions*, Maryland, p. 102-189, 1982.

KOCH, G. V. *A coesão textual*. São Paulo, Contexto, 1989.

KOCH, G. V. *Argumentação e linguagem*. São Paulo: Cortez Editora, 1984.

KOCH, G. V. *Desvendando os segredos do texto*. São Paulo: Cortez Editora, 2002.

KOCH, I. G. V. e TRAVAGLIA, L. C. *Texto e coerência*. São Paulo: Cortez, 1989.

KUPIEC, J.; PETERSEN, J. e CHEN, F. A trainable document summarizer. *Proceedings of the 18<sup>th</sup> annual international ACM SIGIR - conference on research & development in information retrieval*, Seattle, p. 68-73, 1995.

LARocca NETO, J.; SANTOS, A.D.; KAESTNER, C.A.A. e FREITAS, A.A. Generating text summaries through the relative importance of topics. *Lecture notes in artificial intelligence*,

*international joint conference 7<sup>th</sup> ibero-american conference on AI, 15<sup>th</sup> brazilian symposium on AI IBERAMIA-SBIA 2000*, Atibaia, vol. 1952, p. 300-309, 2000.

LÖBNER, S. Definites. *Journal of semantics*, v.4, p.279-326, 1985.

LUHN, H. P. The automatic creation of literature abstracts. *IBM journal of research and development*, Londres, v.2, p. 159-165, 1958.

MANI, I. e MAYBURY, M.T. (eds.). *Advances in automatic text summarization*. Massachusetts: Cambridge Editions, 1999.

MANN, W.C. e MATTHIESSEN, C.M. Nigel: a systemic grammar for text generation. *Systemic perspectives on discourse: selected papers from the ninth international systemic workshop*, Londres, p.92-122, 1985.

MANN, W.C. e THOMPSON, S.A. *Rhetorical structure theory: a theory of text organization*. Relatório técnico ISI/RS-87-190, 1987.

MORIN, E. *Lintelligence de la complexité*, Paris: Lh, 1999.

MÜLLER, C; STRUBE, M. MMAX: a tool for the annotation of multi-modal corpora. *Proceedings of the 2<sup>nd</sup> IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Seattle, v.1, p.52-67, 2001.

NUNES, M.G.V.; VIEIRA, F.M.V.; ZAVAGLIA, C.; SOSSOLETE, C.R.C. e HERNANDEZ, J. *A construção de um léxico da língua portuguesa do brasil para suporte à correção automática de textos*. Relatórios Técnicos do ICMC-USP, São Paulo, n. 42, 1996.

PAICE, C. D. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. *Information retrieval research*, Orlando, v.12, p. 18-32, 1981.

PARDO, T.A.S. *DMSumm: um gerador automático de sumários*. São Carlos: UFSCar, 2002a.

PARDO, T.A.S. e RINO, L.H.M. A summary planner based on a three-level discourse model. *Proceedings of the 6<sup>th</sup> NLPRS – natural language processing pacific rim symposium*, Tóquio, p. 533-538, 2001.

PARDO, T.A.S. *GistSumm: um sumarizador automático baseado na idéia principal de textos*. Relatórios Técnicos NILC-TR-02-13, n.63, 2002b.

PARDO, T.A.S.; RINO, L.H.M. e NUNES, M.G.V. GistSumm: a summarization tool based on a new extractive method. *Proceedings of the 6<sup>th</sup> workshop on computational processing of the portuguese language - written and spoken*, Faro, v.1, p.115-143, 2003.

PARDO, T.A.S; NUNES, M.G.V. e RINO, L.H.M. DiZer: an automatic discourse analyzer for brazilian portuguese. *XVII Brazilian symposium on artificial intelligence - SBIA'04*, São Luís, v.2, p. 27-41, 2004.

PARDO, T.A.S e RINO, L.H.M. *TeMário: um corpus para sumarização automática de textos*. Série de Relatórios Técnicos: NILC-TR-03-09, ICMC/USP, São Carlos, 2003.

PARDO, T. A. S. *Métodos para análise discursiva automática*. ICMCUSP: São Carlos, 2005.

PRINCE, E. F. The ZPG letter: subjects, definiteness, and information status. *Discourse description: diverse analyses of a fund-raising text*. John Benjamins, p. 295-325, 1992.

PRINCE, E. F. Toward a taxonomy of given-new information. *Radical pragmatics*. Nova York: Academic Press, p. 223-256, 1981.

O'DONNELL, M. Rsttool 2.4: a markup tool for rhetorical structure theory. *Proceedings of the international natural language generation conference (INLG'2000)*, Mitzpe Ramon, p.55-69, 2000.

RATH, G.J.; RESNICK, A. e SAVVAGE, R. The formation of abstracts by the selection of sentences. *American documentation*, Austin, v. 12, n. 2, p. 139-141, 1961.

RAU, L. F. e BRANDOW, R. Domain-independent summarization of news. *Seminar report of summarizing text for intelligent communication seminar*. Mônaco-Ville, p.81-97, 1993.

REITER, E. e DALE, R. *Building Natural Language Generation Systems*. Maryland: Cambridge University Editions, v.1, 2000.

RINO, L.H.M. e SCOTT, D. Automatic generation of draft summaries: heuristics for content selection. *Proceedings of the third international conference on the cognitive science of natural language processing*, Berna, p.73-91, 1994.

RINO, L.H.M. *Modelagem de discurso para o tratamento da concisão e preservação da idéia central na geração de textos*. São Carlos: UFSCar, 1996.

RUSSELL, B. Descriptions. *Introduction to mathematical philosophy*. George Allen & Unwin Publishers. Londres: Routledge, 1919 [republicado em 1993].

SACHS, J. *The global economy in the information age: reflections on our changing world*. Pensilvânia: Pennsylvania State University, 2002.

SALTON, G. *Automatic text processing*. Estocolmo: Addison-Wesley, 1988.

SARDINHA, T. B. Lingüística de corpus: histórico e problemática. *D.E.L.T.A.*, Vol. 16, nº 2., p. 29-38, 2000.

SCOTT, D.R. e SOUZA, C.S. Getting the message across in rst-based text generation. *Current research in natural language generation*, Pequim, v.3, p. 47-73, 1990.

SIDNER, C. L. *Towards a computational theory of definite anaphora comprehension in english discourse*. MIT, 1979.

STEINBERGER, J.; POESIO, M e JEZEK, K. Two uses of anaphora resolution in summarization. *Information processing and management*, doi:10.1016/j.ipm., 2007.

STRAND, K. *A Taxonomy of Linking Relations*. Manuscrito, 1997.

VAN DIJK, T. A. *La ciencia del texto: un enfoque interdisciplinario*. Tradução Sibila Hunzinger. Barcelona: Paidós, 1992. Tradução de Tekstwetenschap. Een interdisciplinaire inleiding.

VAN DIJK, T. A. *Studies in the pragmatics of discourse*. Paris: Mouton Publishers, 1981.

VIEIRA, R. *Definite description processing in unrestricted text*. Edinburgh: University of Edinburgh, 1998.

VIEIRA, R.; BICK, E.; COELHO, J. C. B.; MULLER, V. M.; COLLOVINI, S.; SOUZA, J. de; RINO, L. Semantic tagging for resolution of indirect anaphora. 7<sup>th</sup> SIGdial Workshop on discourse and dialogue, 2006, Sydney. *Proceedings of the 7<sup>th</sup> SIGdial workshop on discourse and dialogue*. Cambridge: ACL, p.72-78, 2006.

VIEIRA, R.; GASPERIN, C. V. e SALMONALT, S. Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. *Anaphora processing: linguistic, cognitive and computational modelling*. Amsterdã: Benjamins, 2005.

WITTEN, I.H.; MOFFAT, A. e BELL, T.C. Managing gigabytes. *Van nostrand reinhold*. Helsinki, p. 32-37, 1994.

WYNNE, M. Writing a corpus cookbook. *Proceedings of IRCS workshop on linguistic databases*. Philadelphia, p.254-262, 2001.

YULE, G. Interpreting anaphora without identifying reference. *Journal of semantics*, v. 1, p.315-322, 1981.

## **ANEXOS**

## **Anexo 1 - Relatório das atividades realizadas pelo autor desta monografia**

Este texto visa reportar, de modo objetivo, as tarefas que realizei no presente trabalho. Todavia, antes, cabe registrar que atuo no Laboratório de Engenharia da Linguagem desde 2003. Nesse período, participei, como bolsista, do projeto TeXto - *Acesso a informações em bases textuais* e do projeto DIRPI - *Desenvolvimento e Integração de Recursos para Pesquisa de Informação*. Atualmente, sou membro (colaborador) das equipes dos projetos CROWS - *Construção de ontologias para a Web Semântica*, ProCaCoSa - *Processamento de Cadeias de Correferência para a Sumarização Automática de Textos em Português* e PLN-BR - *Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil*.

Em continuidade, apresentei-me para registrar a construção do corpus *Summ-it* (Capítulo 2) e efetuar uma das análises previstas pelo projeto ProCaCoSa (precisamente, um exame dos sumários produzidos pelo GistSumm sob a ótica das cadeias de correferência, a fim de detectar problemas de coerência decorrentes da ausência de resolução de anafórica – Capítulo 3).

Visto que a construção do *Summ-it* envolveu, e ainda envolve, muitos especialistas, elenco a seguir, em ordem cronológica, as tarefas que efetivamente realizei. Inicialmente, trabalhei na preparação das instruções de anotação de correferência, anaforicidade e referências dêiticas (*Guidelines* – Anexo 2), incluindo, evidentemente, uma fase de testes em experimentos do projeto TeXto. Cabe destacar que os *Guidelines* estão em constante atualização. Também, referente à etapa preparatória de anotação, participei da organização da



anotação (cronograma, organograma, logística etc.) e implementei o esquema de anotação de correferência (anáforas e referências dêiticas) em XML, o *annotation scheme file* da MMAX.

Feito isso, eu e outro membro do laboratório efetuamos, para cada um dos 50 textos do corpus, uma inspeção detalhada dos *markables* gerados automaticamente, com o objetivo de verificar se esses estavam dentro dos padrões de anotação de correferência. Precisamente, foram revisadas 5047 unidades. Na sequência, conferimos a configuração de cada um dos *markables* (de acordo, com esses mesmos padrões). Aliás, em todas as tarefas, registramos os ajustes manuais.

Eu também fiz parte do grupo de anotação de correferência, ou seja, fui um dos 12 anotadores. Encerrada a primeira fase de anotação, juntei-me ao grupo responsável pelo consenso.

Em adição, administrei a construção dos textos tarjados e sumários manuais, incluindo elaboração das instruções gerais e contratação dos sumarizadores profissionais. A parte logística dessa tarefa também foi de minha responsabilidade – por exemplo, envio dos textos-fonte digitalizados e impressos, recolha dos textos tarjados e sumários, formatação e digitalização do material coletado etc. Também acompanhei o processo de construção dos sumários manuais e dos textos tarjados, realizando observações e registros para futuros estudos. Uma vez encerrada a produção dos textos tarjados e dos sumários informativos, organizei esse material nos padrões da área.

Atualmente, acompanho outros dois membros da equipe no encerrando do consenso final da última parte do corpus. Também estamos reunindo os registros da construção do

corpus *Summ-it*, incluindo a parte sobre os sumários (manuais e automáticos), sob a forma de relatório técnico.

De posse do corpus anotado e do conjunto de sumários realizei um trabalho de análise das questões centrais ao projeto ProCaCoSa, ou seja, das inter-relações entre as cadeias de correferência (e anáforas) e dos efeitos sobre elas causados pelos processos extrativos de sumarização.

## Anexo 2 - Instruções para anotação de relações anafóricas e referência dêitica

Versão 2.7 – setembro de 2006

### Laboratório de Engenharia da Linguagem – LEL

Universidade do Vale do Rio dos Sinos – UNISINOS

Av. Unisinos, 950 – 93.022-000 – São Leopoldo, RS – Brasil

[www.inf.unisinos.br/~renata/laboratorio/labor.htm](http://www.inf.unisinos.br/~renata/laboratorio/labor.htm)

## Introdução

Este documento contém instruções para anotação de informações anafóricas e dêiticas, designadamente, elaboradas para o discurso escrito do português. O modelo de anotação, aqui, apresentado é baseado em estudos realizados pelos projetos ANACORT (Vieira, 2001), TeXto (Vieira, 2002) e VENEX (Poesio, 2005) e conta com o uso de uma ferramenta de anotação, a *MMAX* (Muller and Strube, 2000). Para efeito de síntese, este documento foi organizado sem discriminar os aspectos teóricos dos operacionais, sendo que as seções 1, 2, 3, 4, 5 e 6 correspondem a etapas de anotação.

## 1. Seleção das unidades de interesse (markables)

Nós estamos preocupados, aqui, em anotar, as relações anafóricas e as referências dêiticas realizadas por sintagmas nominais, que podem ter como núcleo um nome (exemplo 1) ou um pronome (exemplo 2). Locuções articuladoras, tais como, “nesse sentido”, “por essa razão”, “como resultado”, “além disso”, “isto é”, devem ser desconsideradas.

(1) *Quarta-feira, o Brasil inicia a nova campanha de turismo. Segundo o ministro, Olívio Dutra, isso impulsionará significativamente o setor. Ele, contudo, espero resultados a médio prazo.*

(2) *Ela estava responsável com alguém pelo esquema de viagens.*

É importante observar que sintagmas nominais podem empregar uma palavra de outra classe gramatical (adjetivos, verbos etc.) como nomes – por exemplo, “o pensar” em (3) e “o maior” em (4).

(3) *O pensar é uma atividade psíquica consciente e pode ser organizado ou não.*

(4) *Maria, ao comprar dois sorvetes, sabia que o maior era o meu.*

Além disso, sintagmas nominais podem apresentar outros sintagmas nominais internos – exemplo 5 (todos devem ser selecionados).

(5) *As regiões no interior do estado incluem diversos pontos turísticos.*

*As regiões no interior do estado*

*o interior do estado*

*o estado*

Outras observações:

- os apostos não são segmentados. Por exemplo, no sintagma nominal “o ministro, *Olívio Dutra*,” (exemplo 1), não deve ser marcado o sintagma nominal interno, em função apositiva, “*Olívio Dutra*”.

[o ministro, *Olívio Dutra*]<sub>mark1</sub>

Da mesma forma, os apostos sem marca explícita também não devem ser segmentados, por exemplo, em “a química *Nilva Ré-poppi*”, não deve ser marcado o sintagma nominal interno, “*Nilva Ré-poppi*”.

[a química *Nilva Ré-poppi*]<sub>mark1</sub>

- Quando ocorrerem apostos formados por sintagmas nominais internos que representam novas entidades, como “a empresa” no exemplo, “Antonio Calmon, diretor da empresa”, deve-se anotar uma unidade completa:

[Antonio Calmon, diretor da empresa]<sub>mark1</sub>

e outra unidade representando o sintagma nominal interno, “a empresa”.

[Antonio Calmon, diretor d[*a empresa*]<sub>mark2</sub>]<sub>mark1</sub>

No exemplo, “A Motorola, gigante internacional do setor de telecomunicações,” também devem ser marcados os sintagmas nominais internos que representam outras entidades.

[A Motorola, gigante internacional d[*o setor de [telecomunicações]*]<sub>mark3</sub>]<sub>mark2</sub>]<sub>mark1</sub>

- Quando ocorrem sintagmas nominais seguidos de parênteses em função de aposto, como em “feronômios (substâncias produzidas pelos insetos para se comunicar ou servir de atrativo sexual)”, deve-se anotar uma unidade completa:

[feronômios (substâncias produzidas pelos insetos para se comunicar ou servir de atrativo sexual)]<sub>mark1</sub>

e outra(s) unidade(s) representando o(s) sintagma(s) interno(s) que representam outra(s) entidade(s):

[feronômios (substâncias produzidas pel[*os insetos*]<sub>mark2</sub> para se comunicar ou servir de [*atrativo sexual*]<sub>mark3</sub>)]<sub>mark1</sub>

Observação: os casos que apresentam URL não devem ser considerados apostos - não representam a mesma entidade que a expressão anterior - como em “a Energy Resource 2000 (*www.energy.org*)”, a URL “*www.energy.org*” não deve ser anotada – não é considerada um markable na anotação:

[a Energy Resource 2000]<sub>mark1</sub>

- listas, como, “mamíferos, anfíbios, répteis e aves”, formam uma unidade completa

[mamíferos, anfíbios, répteis e aves]<sub>mark1</sub>

e uma unidade para cada nome

[[mamíferos]<sub>mark2</sub>, [anfíbios]<sub>mark3</sub>, [répteis]<sub>mark4</sub> e [aves]<sub>mark5</sub> ] **mark1**

- Nomes: construções em que dois nomes estão coordenados, tais como “vinho **e** queijo”, “inverno **ou** verão”, formam três unidades, uma completa:

[vinho e queijo]<sub>mark1</sub>

[inverno ou verão]<sub>mark1</sub>

e outras duas para cada nome:

[[vinho]<sub>mark2</sub> e [queijo]<sub>mark3</sub> ] **mark1**

[[inverno]<sub>mark2</sub> ou [verão]<sub>mark3</sub> ] **mark1**

- Modificadores: um sintagma nominal formado por modificadores coordenados, como por exemplo, “*amarelas e vermelhas*” em: “bolsas *amarelas e vermelhas*”, não devem ser separados. Exemplos:

[aspecto *ecológico e social*]<sub>mark1</sub>

[a diversidade genética animal e vegetal]<sub>mark1</sub>

[plantas silvestres e medicinais]<sub>mark1</sub>

Os exemplos citados não precisam ser separados, mas devem ser separados no caso de uma posterior referência explícita a **uma de suas partes** no decorrer do texto (por exemplo, no caso de ocorrer uma referência ao aspecto ecológico, à diversidade vegetal ou plantas silvestres nos exemplos descritos).

- Cláusulas relativas restritivas: quando um sintagma nominal apresentar na sua estrutura uma cláusula relativa restritiva, o pronome relativo (que, onde, em que, na qual etc.) não deve ser considerado, como por exemplo, o pronome relativo “que” em: “a NASA está estudando *os mamíferos que se adaptaram ao clima polar*, a fim de desenvolver novas tecnologias espaciais”:

[os mamíferos que se adaptaram ao clima polar]<sub>mark1</sub>

[o clima polar]<sub>mark2</sub>

Outro exemplo:

[uma falha geológica ativa *que pode explicar aquele terremoto*]<sub>mark1</sub>

[aquele terremoto]<sub>mark2</sub>

- Cláusulas relativas não restritivas: esses casos devem ser desconsiderados. Por exemplo, no enunciado “*Um dos autores do estudo, que está publicado na edição de hoje da revista britânica Nature, deixou claro o novo modelo implementado*”, o primeiro argumento (*Um dos autores do estudo, que está publicado na edição de hoje da revista britânica Nature,*) deve ser anotado da seguinte forma:

[*Um dos autores do estudo*]<sub>mark1</sub>

[a edição de hoje da revista britânica Nature]<sub>mark2</sub>

[a revista britânica Nature]<sub>mark3</sub>

Outro exemplo de como devem ser anotadas cláusulas relativas não restritivas pode ser visto na marcação, a seguir, do sintagma nominal “*a chamada falha normal, em que forças opostas levam blocos rompidos do solo a modificar de posição*”:

[a chamada falha normal]<sub>mark1</sub>

[forças opostas]<sub>mark1</sub>

[blocos rompidos de o solo]<sub>mark3</sub>

[o solo]<sub>mark4</sub>

[posições]<sub>mark5</sub>

Neste exercício de anotação, os sintagmas nominais são nossas unidades de interesse, denominadas a partir de agora, também, de *markables*. Cabe observar que os markables são gerados automaticamente e, posteriormente, revisados manualmente.

## 2. Indicação de correferência (member) e anáfora (pointer)

Considera-se, aqui, a referência (*referenciação*) como a propriedade da linguagem de evocar entidades, tais como, pessoas, animais, lugares, fatos etc, para o discurso. Instruções linguísticas usadas para referir são denominadas *expressões referenciais*.

Freqüentemente, é possível observar que essas expressões se ligam a outras expressões no texto, numa relação de identidade, associação ou dependência, constituindo, assim, *cadeias de correferência*. Quando a ligação entre as expressões se dá uma relação de identidade, elas são chamadas de *expressões correferentes* (ocorre a *correferência*). No exemplo:

(6) *O Eurocenter oferece cursos de Japonês na bela de Kanazawa. Os cursos têm quatro semanas de duração. As aulas do nível avançado incluem refeições típicas e passeios a pontos turísticos da cidade.*

observamos que na sentença inicial são introduzidas três referentes (“O Eurocenter”, “cursos de Japonês” e “a bela de Kanazawa”). Na sentença seguinte, a expressão “Os cursos” retoma “cursos de Japonês”. Nessa perspectiva, “cursos de Japonês” é antecedente de “Os cursos”, ou seja, duas expressões referenciais que fazem menção à mesma entidade, portanto expressões correferenciais.

- Quando um markable apresentar essa relação, assegure-se de adicioná-lo ao set apropriado, selecionando um antecedente com o mecanismo **member** da MMAX (excetuando os casos de encapsulation abordado a seguir).

Estendendo o conceito de correferência, é possível localizar a noção de anáfora, ou de anaforicidade. Em *lato sensu*, a anáfora se define como toda retomada de um elemento anterior em

um texto, mantendo-se a identidade referencial, como, por exemplo: “cursos de japonês” – “os cursos”, “a bela de Kanazawa” – “a cidade” (em 6). Já, em *stricto sensu*, o fenômeno anafórico pode não ser correferencial, o referente de uma expressão anafórica pode não ser explicitamente denotado por um mesmo referente anterior. Em (6), observamos que a expressão “As aulas do nível avançado” não é correferente a nenhum termo anterior, entretanto, apresenta parte do seu significado ancorado na expressão “cursos de Japonês”, tanto que o sentido pleno dessa expressão (As aulas do nível avançado do curso de japonês) é alcançado pela soma da expressão anafórica (“As aulas do nível avançado”) a sua âncora referencial (“cursos de Japonês”). Assim, a anáfora pode ser um fenômeno semântico de natureza inferencial e pode ser um fenômeno de correferência. Em síntese, uma expressão anafórica pode retomar uma referência anterior mantendo uma relação de identidade (i.e., anáfora correferencial), mas também pode ativar um novo referente cuja interpretação é dependente de outras expressões referenciais anteriormente presentes do texto (i.e., anáfora associativa).

- Quando um markable apresentar essa relação, certifique-se de apontar a expressão que serve de âncora textual, usando o mecanismo **pointer** da MMAX.

### 3. Atributos dos markables

Nomes dos Atributos	Descrição	Forma de Anotação
Comment	usado para inserir comentários da anotação.	Manual
np_form	os valores deste atributo são os tipos de sintagmas nominais, os quais foram baseados na lista usada no GNOME.	automática (revisão manual)
pro_form	os valores deste atributo são os tipos de pronomes, os quais foram baseados na lista usada no GNOME.	automática (revisão manual)
Member	Usado no MMAX para indicar cadeias de correferência constituídas por expressões que retomam a mesma entidade no discurso.	Mecanismo de anotação do MMAX
Pointer	usado no MMAX pra indicar uma referência associativa.	mecanismo de anotação do MMAX
Status	representa as relações possíveis entre as entidades do discurso. Os valores deste atributo podem ser <i>new</i> , <i>old</i> , <i>associative</i> e <i>deictic</i> .	automática (revisão manual)
Is_bridging	Quando “status=associative”, o atributo is_bridging indica o tipo de relação associativa expressa por um <i>pointer</i> .	manual
Is_anaphoric	Quando “status=old”, o atributo is_anaphoric especifica o tipo de relação entre a entidade do discurso e o seu antecedente ( <i>direct</i> , <i>indirect</i> , <i>pronominal</i> e <i>encapsulation</i> ).	automática (revisão manual)

## 4. Identificação de forma dos markables

Entendendo que as configurações morfosintáticas oferecem pistas valiosas para processamento da língua, nós definimos os atributos `np_form` e `pro_form`, em que o anotador distinguirá os sintagmas nominais com núcleo nome (**np-n** = yes) dos pronomes (**np-n** = no).

➤ **np\_form** (**np-n** = yes) - Sintagmas nominais com núcleo nome:

- **def-np**: sintagma nominal formado por núcleo nome comum e determinante artigo definido (o, a, os, as), exemplo: "o setor" (em 1), "As regiões no interior do estado" (em 5).
- **def-pn**: sintagma nominal composto por núcleo nome próprio e determinante artigo definido, exemplo: "o *Brasil*" (em 1). Datas iniciadas por determinante artigo definido podem ser classificadas como **def-pn**, exemplo: "os anos 60", "o século 19".
- **indef-np**: sintagma nominal constituído por núcleo nome e artigo indefinido (um, umas, uns, umas) exemplo: "uma casa".
- **dem-np**: sintagma nominal com núcleo nome e determinante pronome demonstrativo (esse, essas, aquele, isso etc.), exemplo: "essa última disputa", "aquele terremoto".
- **poss-np**: sintagma nominal constituído por determinante pronome possessivo (seu, suas, nossa etc.) e núcleo nome, exemplo: "sua casa", "nossa pesquisa".
- **int-np**: sintagma nominal constituído por determinante pronome interrogativo (quando, quem, qual etc.) e núcleo nome, exemplo: "que horas termina a aula?"
- **num-np**: sintagma nominal iniciado por um numeral (um, primeiro etc.) e núcleo nome, exemplo: "95 empresas", "10 mil trabalhadores".
- **quant-np**: sintagma nominal constituído por núcleo nome e seguido de quantificadores (todos os, todas as, a maioria, ambos etc.). Os casos de determinante pronome indefinido (algum, outra, certa, várias, etc.) devem ser classificados como **quant-np**.
- **coord-np**: sintagmas nominais coordenados, exemplos: "ceras, resinas e gomas", "Maria e João", "café ou chá".
- **bare-np**: sintagma nominal sem determinantes composto por núcleo nome comum, exemplo: "quarta-feira" (em 1), "viagens" (em 2).
- **pn**: sintagma nominal sem determinantes núcleo nome próprio, exemplo: "Maria" (em 4), "Porto Alegre". Datas também podem ser classificadas como **pn**, exemplo: "1732".

➤ **pro\_form** (**np-n** = no) - Pronomes:

- **indef-pro**: sintagma nominal formado exclusivamente por um pronome indefinido, exemplo: "alguém" (em 2).
- **dem-pro**: sintagma nominal constituído somente por um pronome demonstrativo, exemplo: "isso" (em 1).
- **pes-pro**: sintagma nominal formado exclusivamente por um pronome pessoal, exemplo: "Ele" (em 1), "Ela" (em 2).



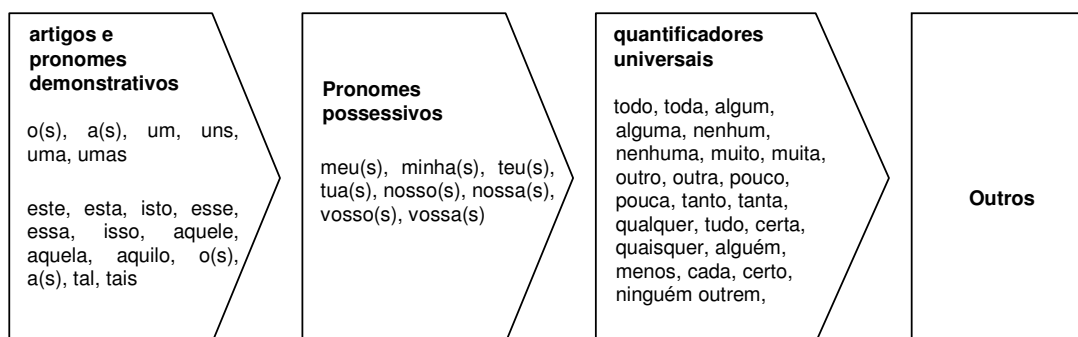
- **poss-pro:** sintagma nominal formado exclusivamente por pronome possessivo, exemplo: “meu”, “teu”.
- **int-pro:** sintagma nominal formado exclusivamente por um pronome interrogativo, exemplo: “quando”, “quem”.
- **num-ana:** sintagma nominal núcleo numeral, exemplo: “eu preciso de dois”.

Observações: Nos casos de sintagmas nominais com determinantes complexos (exemplos 7 e 8) deve-se seguir a hierarquia para classificação da forma proposta neste manual. Por exemplo: em (7) e (8), as classificações adequadas são def-np e dem-np respectivamente.

(7) *Todos os alunos passaram na disciplina de Cálculo.*

(8) *Esses meus amigos participaram da festa.*

#### Hierarquia para classificação da forma



Além disso, ao contrário de algumas análises morfossintáticas que consideram os numerais como candidatos a ocupar a função *determinante*, nós limitamos os numerais à função *modificador*. Por exemplo: “**as** primeiras experiências com inseticidas”, a classificação adequada é **def-np**.

## 5. Classificação das informações referenciais dos markables

Na gama de fenômenos relativos a coesão referência (referenciação, correferência, anáforas, elipses etc.), a casos em que a informação requerida para interpretação de uma expressão não é encontrada no texto, mas na situação comunicativa. Esses casos configuram a *referencia dêitica*, ou simplesmente, a *dêixis*. Resumidamente, pode-se dizer que a dêixis é a característica da linguagem humana que consiste em fazer um enunciado referir-se a uma situação definida, real ou imaginária, que pode ser: i) quanto aos participantes do ato de enunciação ou todo assunto da comunicação, exemplo: “nós” em (10); ii) quanto ao momento da enunciação, exemplo: “semana passada” (em 10); ou iii) quanto ao lugar onde ocorre a ação, estado ou processo, exemplo: “nesta edição” em (10).

(10) *Nesta edição, nós apresentamos o novo conversível XJ520 da Jaguar lançado na semana passada. Em primeira mão, explicaremos como funciona o motor de 620cv de potência a 7.600rpm.*

Em (10), as expressões “Nesta edição” e “nós” são ditas dêiticas, pois seus referentes só podem ser recuperados em relação à situação comunicativa. Com nosso conhecimento de mundo, nós sabemos que a primeira faz referência ao próprio objeto de leitura (a revista) e a segunda ao corpo editorial. Já o referente da expressão dêitica “a semana passada” somente pode ser recuperado de posse da data da revista.

Como dito anteriormente, nós estamos interessados aqui com as relações anafóricas nominais. Um tipo importante de relacionamento anafórico que nós não estudaremos será a elipse, como em:

(11) *Meu livro não está aqui, [ele] sumiu.*

A par do exposto, num primeiro momento, o anotador deve indicar o status dos markables:

#### ➤ **status**

- **new:** o sintagma nominal introduz um novo referente no discurso sem apresentar parte de seu sentido ancorado em uma expressão anterior. Exemplo: “o novo conversível XJ520 da Jaguar” (em 10).
- **old:** o sintagma nominal é uma anáfora correferencial, ou seja, retoma um referente já introduzido por uma expressão anterior. Exemplo: “cursos de japonês” – “os cursos”, “a bela de Kanazawa” – “a cidade” (em 10).
  - Se um markable estiver ligado a outro com **member** selecionar esse status.
- **associative:** o sintagma nominal é uma anáfora associativa, ou seja, introduz um novo referente no discurso cujo significado está ancorado em uma expressão anterior. Exemplo: “cursos de japonês” – “as aulas do nível avançado” (em 8), “o novo conversível XJ520 da Jaguar” – “o motor de 620cv de potência a 7.600rpm” (em 10).
  - Se um markable estiver ligado a outro por **pointer** marcar esse status.
- **deictic:** o sintagma nominal é uma referência dêitica. Exemplo: “Nesta edição” (10).

## 6. Classificação dos relacionamentos anafóricos correferenciais

Essa etapa de anotação é feita de modo automático após todas as etapas manuais. Nela, os casos, anteriormente, classificados como anafóricos correferenciais (status=“old”), são classificados no atributo `is_anaphoric` em:

#### ➤ **is\_anaphoric**

- **direct:** a expressão anafórica e seu antecedente apresentam núcleos idênticos. Exemplo: “cursos de japonês” – “os cursos” (em 8).
- **indirect:** a expressão anafórica e seu antecedente apresentam núcleos diferentes. Exemplo: “a bela de Kanazawa” – “a cidade” (em 8).
- **encapsulation:** a expressão anafórica (inclusive pronomes) retoma um trecho de texto maior que um sintagma, tais como sentenças ou mesmo parágrafos – por exemplo, em (12), “a operação” retoma “O Banco Central interveio ontem para segurar a cotação do dólar”.

- Nesse caso, não é necessário adicioná-lo a um **set**, entretanto, é preciso “comentar” qual trecho de texto foi retomado. Para isso, temos o atributo **comment** para adicionar esses comentários à anotação.

(12) *O Banco Central interveio ontem para segurar a cotação do dólar. A operação ocorreu quando a moeda havia alcançado R\$ 2,08.*

## 7. Classificação dos relacionamentos anafóricos associativos

Nesta etapa, é especificado, no atributo `is_bridging`, o tipo de relacionamento anafórico associativo. Assim, assegure-se de que os markables anafóricos associativos estão marcados com uma das relações seguintes:

### ➤ **is\_bridging**

- **element-of**: essa é a relação aplicada quando uma entidade do discurso evocada por uma expressão anafórica é um elemento de um grupo previamente introduzido – exemplos (13) e (14). Quando o elemento ocorre primeiro, deve-se usar a relação inversa: **element-of-inv** – exemplo (15). Atenção! Algumas relações associativas têm uma relação inversa, entretanto, deve-se sempre apontar da expressão anafórica para seu antecedente, nunca de outra maneira!!!

(13) *As turbinas do F-996 possuem sistemas distintos de segurança. A turbina traseira recebeu um sistema contra-incêndio.*

(14) *Hoje, o Museu Nacional inaugura uma nova galeria expondo dois valiosos vasos chineses. Segundo o curador, o vaso maior, que mede 90 cm, está avaliado em mais de cinquenta milhões de dólares.*

(15) *Por exemplo, o gás argônio é um elemento encontrado em diminuta proporção na atmosfera terrestre. Contudo, a partir de janeiro, a UNICAMP iniciará o rastreamento de todos os nobres incolores na estratosfera.*

- **subset-of**: essa relação é empregada sempre que uma expressão anafórica retoma um subconjunto de uma entidade introduzida anteriormente no texto. Essa relação também é utilizada para relacionamentos entre um tipo e seus subtipos. É possível perceber isso em (16) e (17).

(16) *Os símios não precisam de muito espaço e se alimentam de quase tudo o que existe na floresta: folhas, brotos, frutinhas. O inverno, porém, é a estação de fartura para os bugios pela fartura de pinhões.*

(17) *Para orientar as manobras das asas-deltas, o Pico-da-águia recebeu aparelhos do aeródromo Olavo Bilac. Perpendicularmente à extremidade sul, birutas foram instaladas.*

- **part-of**: essa relação deve ser marcada quando a expressão anafórica retoma uma parte de uma entidade já mencionado – exemplos (10). Quando a parte ocorre antes do todo, deve-se usar a relação inversa: **part-of-inv** – exemplo (18).

(18) *Uma nova torre chama a atenção dos cariocas. A Igreja Nossa Senhora Aparecida foi totalmente restaurada, sendo que a estrutura do sino foi aumentada dez metros.*

- **entity-attribute:** essa relação é empregada sempre quando a expressão anafórica retoma um atributo de uma entidade previamente mencionada – exemplo (19).

(19) *Devido aos novos padrões, as modelos estão se submetendo a severos programas de dieta. Conforme a agência Bellezza, o peso tem de ser menor do que é normalmente exigido.*

- **possessor-thing:** essa relação é marcada quando o antecedente possui a entidade evocada pela expressão anafórica associativa – exemplo (21).

(21) *Ontem, o ministro dos transportes foi convocado pela polícia para prestar esclarecimentos. Semana passada, foram encontradas, no apartamento, várias mercadorias contrabandeadas.*

- **other-bridging:** essa relação deve ser usada para outro tipo de relação não definido pelos anteriores. Por exemplo, os epítetos prestam-se particularmente a refletir a apreciação subjetiva do enunciador com relação ao discurso. Muitas vezes, as anáforas desse tipo constituem fatos de polifonia, quando a recategorização do referente é atribuída a uma outra voz que não a do enunciador, como se vê em:

(22) *Ao analisar os resultados do Sistema Nacional de Avaliação Básica do MEC, o ministro Paulo Renato Souza (Educação) afirmou que "a escola está cada vez menos interessante e motivadora, e o aluno cada vez mais dispersivo e indisciplinado". A pesquisa deste ano mostrou queda de aproveitamento nas escolas particulares. Para Paulo Renato, esse "efeito chatice" é provocado por duas razões centrais: a falta de reciclagem das escolas e a grande oferta de conhecimento fora da sala de aula, principalmente na Internet.*

Nem sempre é fácil distinguir um relacionamento part-of de um entity-attribute. Os testes lingüísticos baseados em construções lingüísticas gerais tais como "X de Y" ou "X tem Y" não são de muita ajuda: pode-se dizer que "casas têm janelas/portas/cômodos" e "casas tenha uma altura/largura/preço". Um critério razoavelmente útil é o status ontológico: as partes (part-of) tendem a ser objetos concretos, tais como portas, janelas, rodas, motores etc., e os atributos (entity-attribute) tendem a ser objetos mais abstratos, tais como altura/peso/velocidade. Os testes lingüísticos que usam verbos mais específicos podem também ser úteis: pode-se dizer dos atributos tais como a altura que a "altura é um atributo dos objetos", mas não se pode dizer que as partes dos objetos como: "as janelas são um atributo das casas". Inversamente, se pode dizer que as "janelas são partes das casas", mas não que "o comprimento é parte da casa". Também, o relacionamento possessor-thing não são aqueles entre um objeto e um de seus atributos ou suas partes. Se você não puder se decidir, marque estes casos como is\_bridging= "other-bridging" e use o atributo **comment**.

### Anexo 3 - Exemplo de relatório de cadeias de correferência

	Classificação	Sintagma
CADEIA : set 33		
word_1..word_19	---	Um ser que invade corpos e domina a mente alheia , forçando suas vítimas a fazer o_que ele ordena
word_18	---	ele
word_34..word_35	---	essa criatura
word_49..word_56	---	-Hymenopimecis sp .- o tal invasor de corpos
word_59..word_60	---	uma vespa
word_75..word_76	---	esse inseto
word_114..word_115	direct---old	a vespa
word_145..word_146	direct---old	a Hymenopimecis
word_235..word_236	indirect---old	a parasita
word_266	---	ela
word_322	---	parasita
CADEIA : set 34		
word_29..word_32	---	uma aranha de a Costa_Rica
word_81..word_84	direct---old	a aranha Plesiometa argyra
word_92..word_93	indirect---old	a hospedeira
word_99..word_100	direct---old	a aranha
word_130..word_131	---	sua anfitriã
word_141..word_142	direct---old	a aranha
word_166..word_167	indirect---old	o aracnídeo
word_187..word_188	indirect---old	a vítima
word_197..word_198	direct---old	a aranha
word_202	---	ela
word_224..word_225	direct---old	o aracnídeo
word_255..word_256	direct---old	a aranha
word_259	---	a
word_268..word_269	---	sua ex-hospedeira
word_307	---	aranhas
word_324	---	hospedeiro
CADEIA : set 23		
word_72..word_76	---new	as larvas de esse inseto
word_95..word_96	indirect---old	A larva
word_151..word_152	direct---old	A larva
word_248..word_249	direct---old	a larva
CADEIA : set 31		
word_89..word_93	---new	o comportamento de a hospedeira
word_305..word_307	---	comportamento de aranhas
CADEIA : set 25		
word_106..word_108	---new	a própria teia
word_210	---	teia
word_243..word_244	direct---old	a teia
CADEIA : set 29		
word_111..word_115	---new	o casulo de a vespa
word_232..word_236	direct---old	o casulo de a parasita
word_277..word_278	direct---old	o casulo
CADEIA : set 32		
word_133..word_135	---new	A relação espúria
word_316..word_324	---	uma interação química tão complexa entre parasita e hospedeiro
CADEIA : set 18		
word_178..word_179	---	uma droga
word_190..word_191	indirect---old	A substância
CADEIA : set 35		
word_263..word_264	encapsulation---old	a exploração
word_328..word_330	direct---old	A exploração alheia

### **Texto original CIENCIA\_2000\_17108 referente ao relatório anterior**

Um ser que invade corpos e domina a mente alheia, forçando suas vítimas a fazer o que ele ordena, não é mero personagem de ficção. Para uma aranha da Costa Rica, essa criatura existe. E não se trata de nenhum extraterrestre. Apesar do nome *Hymenoepimecis* sp., o tal invasor de corpos é só uma vespa. O biólogo William Eberhard, da Universidade da Costa Rica, descobriu que as larvas desse inseto, ao parasitar a aranha *Plesiometa argyra*, provocam mudanças no comportamento da hospedeira. A larva induz quimicamente a aranha a modificar o formato da própria teia para que o casulo da vespa possa se desenvolver. Não satisfeita com a manipulação, ainda mata e devora sua anfitriã. A relação espúria começa no abdome da aranha, onde a *Hymenoepimecis* injeta os ovos. A larva passa de 7 a 14 dias ali dentro, fartando-se do sangue do aracnídeo, até estar madura o suficiente. Então, libera uma droga ainda desconhecida na corrente sanguínea da vítima. A substância atinge o sistema nervoso da aranha. Dopada, ela passa a repetir um único padrão de teia, em vez de tecê-lo no formato circular tradicional. Sem saber, o aracnídeo está providenciando o suporte perfeito para o casulo da parasita. Na noite em que a teia fica pronta, a larva irrompe do corpo da aranha, matando-a. Para completar a exploração, ela devora sua ex-hospedeira. Só então começa a entrar no casulo, onde se transformará numa vespa adulta. "É uma descoberta e tanto", disse o psicólogo César Ades, da USP, especialista em comportamento de aranhas. "É a primeira vez que se vê uma interação química tão complexa entre parasita e hospedeiro", afirmou. A exploração alheia não tem limites. Nem mesmo no reino animal.

## Anexo 4 - Anotação RST

Em linhas gerais, pode-se dizer que *Rhetorical Structure Theory* – RST (Mann and Thompson, 1987), estabelece suas bases no princípio de coerência (Seção 1.2) e na idéia de que um texto apresenta uma estrutura retórica subjacente cuja organização permite alcançar o objetivo comunicativo do produtor. Essa estrutura é formada por unidades elementares do discurso EDUs (*Elementary Discourse Unit*), inter-relacionadas por meio de vínculos retóricos. Essas relações expressam os tipos de ligações existentes entre as EDUs, com vistas na organização coerente do texto.

Em RST, atribui-se às EDUs o papel de núcleo ou satélite. Numa ponta, a unidade nuclear (núcleo) sinaliza a informação principal, informação mais relevante. Na outra ponta, o satélite indica a informação adicional, sendo que essa influencia na interpretação da informação presente no núcleo. Além disso, a RST prevê situações em que as unidades são nucleares, ou melhor, há relações multinucleares – com mais de um núcleo e nenhum satélite.

Nesse quadro, as relações podem ser divididas em: hipotáticas e paratáticas (CARLSON e MARCU, 2001). As relações hipotáticas se apresentam em pares de EDUs com diferentes níveis de importância, havendo, rigorosamente, uma unidade nuclear e a outra satélite. As relações paratáticas ligam EDUs com um grau aproximadamente equivalente de importância. A primeira situação é denominada mononuclear e a segunda multinuclear.

A anotação RST do corpus *Summ-it* se deu nesses moldes com o auxílio de uma ferramenta de anotação: a RSTTool (O'DONNELL, 2000). A interface da RSTTool oferece um importante apoio gráfico, contemplando a visualização estrutural em árvore prevista pela

RST – como pode ser observado na Figura 16. Nessa representação cada segmento de uma relação é indicado por uma linha horizontal, sendo os segmentos nucleares assinalados por uma linha vertical. Excetuando as relações multinucleares, a seta aponta rigorosamente na direção satélite-núcleo.

Todo o processo de anotação foi orientado pelos *guidelines* elaborados por Mann e Thompson (1987) e Carlson e Marcu (2001). Com vistas de diminuir situações de dúvida, alguns incrementos foram feitos pelos anotadores.

Primeiramente, o texto foi segmentado em EDUs. Embora seja arbitrário o tamanho de uma EDU em RST, os *guidelines* sugerem que seja considerada a oração como unidade elementar. Tal sugestão foi adotada na anotação RST do corpus *Summ-it*.

Na sequência, foram estabelecidas as relações entre as EDUs, em que os vínculos devem expressar uma estrutura hierárquica. Para isso, a RSTTool, além de disponibilizar alguns conjuntos pré-definidos de relações RST, possibilita que seja adicionado novas relações ou mesmo novos conjuntos de relações. Na anotação do *Summ-it*, os anotadores utilizaram um conjunto de 32 relações, já empregado por Pardo (2005), que abarca 26 relações do conjunto de Mann e Thompson e seis do conjunto de Carlson e Marcu. Um dos motivos que motivaram a adoção desse conjunto é um objetivo imediato da presente anotação: servir ao analisador discursivo DiZer (PARDO, 2005). Esse sistema busca identificar automaticamente relações retóricas em textos de divulgação científica em português com base em *Machine Learn* e indicadores textuais. Logo, a anotação manual servirá aos experimentos de aprendizado e ao levantamento de “pistas” textuais. Mais detalhes podem ser encontrados em Collovini et al. (2007).



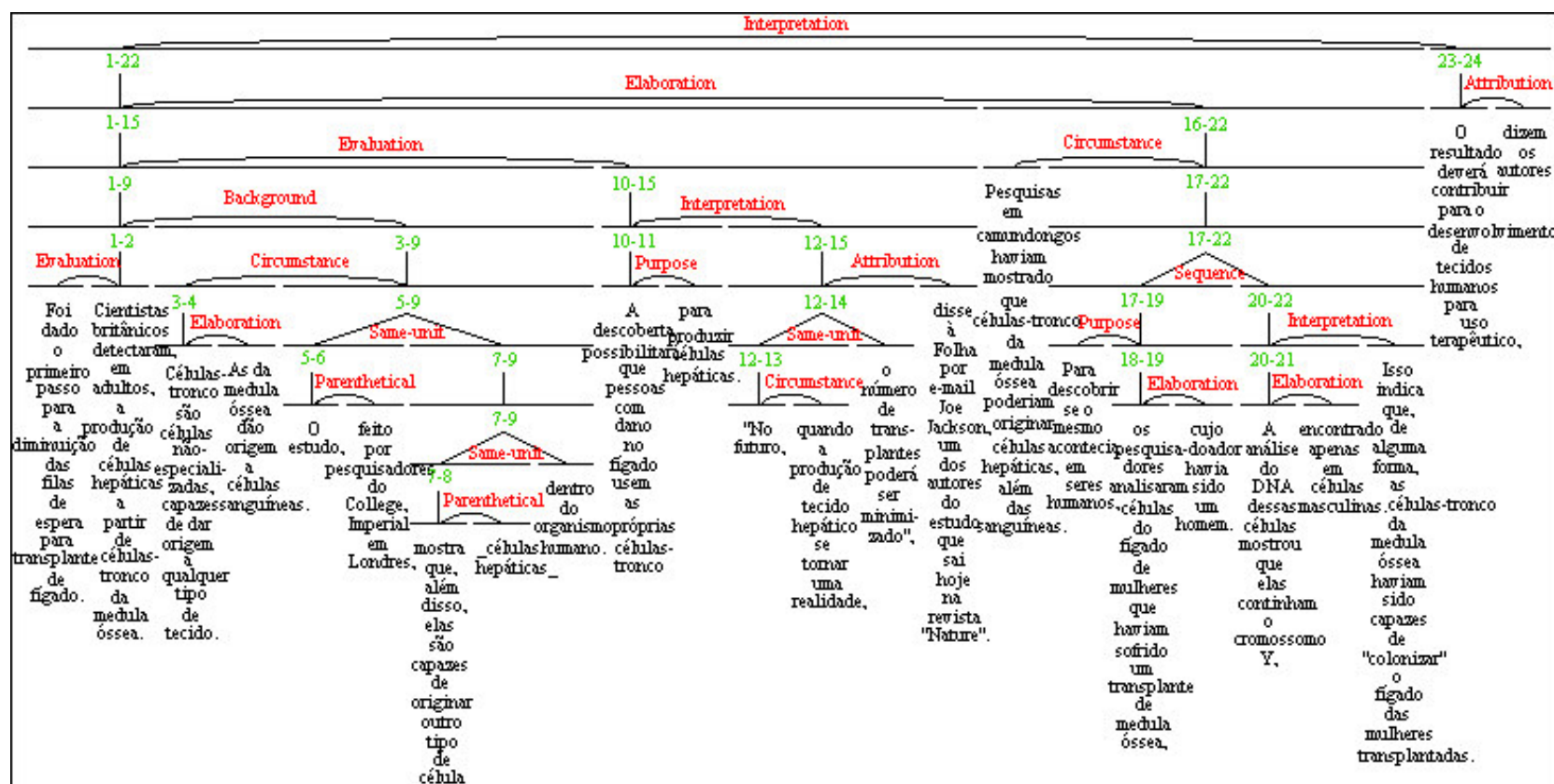


Figura 16 – Representação arbórea RST

## Anexo 5 - Orientações gerais para equipe de sumarizadores

### Nomenclatura geral

- Resumo/sumário: textos condensados de um texto original.
- Texto-fonte = texto original.
- Sumários indicativos: apresentam apenas os tópicos essenciais do texto-fonte, sem incluir detalhes de resultados, argumentos e conclusões; visando somente indicar ao leitor o que será encontrado no texto integral. Geralmente, esses sumários são empregados como indexadores de textos-fonte em, por exemplo, consultas eletrônicas a uma base de dados bibliográfica.
- Sumários informativos: contêm, além dos tópicos essenciais, informações principais de resultados, argumentos e conclusões, uma vez que pretendem ser uma reprodução condensada, de fato, do texto-fonte, podendo, inclusive, substituí-lo. Usualmente, esses sumários servem como substitutos de textos-fonte.
- Sumário crítico = resenha

### Tarefas solicitadas

#### Tarefa 1 – elaboração dos textos tarjados

- i. marcar com sublinhado a(s) sentença(s) que lhe indicam a idéia principal do texto-fonte em anexo, isto é, aquela(s) que lhe dão as bases para a elaboração do respectivo sumário;
- ii. marcar com **negrito** uma sentença, entre as marcadas em **i**, que seja a mais representativa da idéia principal do texto (no caso, de haver apenas uma sentença marcada em **i**, essa passa automaticamente a atender o item **ii**).

#### Tarefa 2 – elaboração dos sumários informativos

Elaborar, para o texto-fonte em anexo, um sumário informativo, ou seja, aquele em que a fidelidade ao original (o sumário apresenta sentidos que, também, podem ser alcançados pelo texto-fonte, sendo que, jamais, aquele contradiz as idéias desse) se sobressai como aspecto necessário e, por isso, ele pode ser substituto do original correspondente.

Destaca-se que o tamanho do sumário deve ser de aproximadamente 25-30% do tamanho do texto-fonte em anexo. Outras variações de extensão podem ser consideradas desde que justificadas por escrito.