Uma abordagem baseada em córpus e em sistemas de crítica para a construção de ambientes Web de auxílio à escrita acadêmica em português

Valéria Delisandra Feltrim

Orientadora: ***Profa. Dra. Maria das Graças Volpe Nunes***
Co-orientadora: ***Profa. Dra. Sandra Maria Aluísio***

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional.

"VERSÃO REVISADA APÓS A DEFESA"

Data da Defesa:	10/12/2004
Visto do Orientador:	

USP – São Carlos
Fevereiro/2005

À minha familia. Sem ela,
nada seria possível.

Agradecimentos

A Graça V. Nunes, por me apresentar ao maravilhoso mundo da Lingüística Computacional. Seu profissionalismo certamente deixará marcas em minha formação. Acima de tudo, quero agradecer-lhe por sua amizade, incentivo e apoio em tantos momentos críticos.

A Sandra M. Aluísio, por ser um modelo de entusiasmo e amor à pesquisa. Em especial, quero agradecer-lhe por estar sempre disponível a me ajudar na interpretação dos modelos estruturais e no árduo trabalho da anotação manual de córpus.

A Jorge M. Pelizzoni, uma das pessoas mais notáveis que já conheci, por me socorrer sempre com suas idéias brilhantes e por não me deixar desanimar nunca. Serei infinitamente grata pela chance de trabalharmos juntos. Tê-lo como amigo é uma benção.

A Simone Teufel, minha supervisora no *Computer Lab*, Universidade de Cambridge, por me guiar nos caminhos do *Argumentative Zoning*. Sua experiência e tato na condução do estágio fizeram os meses que trabalhamos juntas dos mais produtivos e gratificantes.

A todas as pessoas que colaboraram para a conclusão deste trabalho: A Lucas Antiqueira, pela sua dedicação na implementação do *SciPo*. A Ivair Castelan e a Marcela F. Fossey, pelo auxílio na anotação do CorpusDT. A Lucia Specia, por revisar esta tese, mesmo estando em outro continente. Agradeço também a todos os alunos que participaram dos experimentos de avaliação do *SciPo*.

A todas as pessoas do NILC, com as quais tive o prazer de conviver nos últimos anos.

À Fapesp, à Capes e ao CNPq, pelo auxílio financeiro que possibilitou a realização deste trabalho.

Finalmente, agradeço a todos os meus amigos, pelos momentos inesquecíveis que passamos juntos e pelo apoio incondicional. Em especial, a Napoleão F. Valadares, por todo carinho e por agüentar a "montanha russa" de emoções vivida nos últimos tempos.

Publicações resultantes deste Trabalho

Feltrim, V.D., Teufel, S., Nunes, M.G.V., Aluísio, S.M. (*To appear*). Argumentative Zoning applied to critiquing novices' scientific abstracts. In James G. Shanahan, Yan Qu and Janyce Wiebe (Eds.) *Computing Attitude and Affect in Text.* Dordrecht, The Netherlands: Springer.

Feltrim, V.D., Pelizzoni, J.M., Teufel, S., Nunes, M.G.V., Aluísio, S.M. (2004). Applying Argumentative Zoning in an automatic critiquer of academic writing. In *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA 2004)*, Lecture Notes in Artificial Intelligence, 3171, Springer, p. 214-223.

Feltrim, V.D., Teufel, S. (2004). Automatic critiquing of novices' scientific writing using Argumentative Zoning. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text*, Yan Qu and James G. Shanahan and Janyce Wiebe (Eds.), SS-04-07, AAAI Technical Report Series, Menlo Park, CA:AAAI Press. p. 55-58.

Feltrim, V.D., Antiqueira, L., Nunes, M.G.V., Aluísio, S.M. (2003). A construção de uma ferramenta de auxílio à escrita de Resumos acadêmicos em português. In *Proceedings of ENIA 2003*, Campinas: SBC. p. 2399-2404.

Feltrim, V.D., Aluísio, S.M., Nunes, M.G.V. (2003). Analysis of the rhetorical structure of computer science abstracts in Portuguese. In *Proceedings of the Corpus Linguistics 2003*, Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (Eds.), UCREL Technical Papers, Vol 16, Part 1, Special Issue. p. 212-218.

Feltrim, V.D., Nunes, M.G.V., Aluísio, S.M. (2002). Um corpus de textos científicos para apoiar a construção de um sistema de auxílio à escrita técnica. In *Cadernos de Resumos do 12o InPLA*, São Paulo, SP: LAEL/PUC-SP. p. 209-210.

Resumo

Ferramentas computacionais destinadas a auxiliar a escrita na língua inglesa têm se tornado cada vez mais comuns, incluindo tanto sistemas dedicados ao pós-processamento do texto como sistemas mais completos, que atuam desde a geração e organização de idéias até a escrita. Entre esses sistemas há aqueles que apóiam especificamente a escrita de textos acadêmicos, um gênero mais tratável computacionalmente dada a sua rigorosa padronização. Em contraste, para o português não há sistemas semelhantes, mas apenas ferramentas de pós-processamento, como corretores ortográficos, revisores gramaticais, dicionários para consulta, entre outras, de aplicação geral. Nesta tese de doutorado, focalizamos o problema da escrita em português de teses e dissertações, visando uma modelagem computacional que segue uma abordagem baseada em córpus e em sistemas de crítica, para a criação e avaliação de ferramentas de apoio à criação de textos desse gênero. Para viabilizar o projeto, restringimos nossa análise às seções Resumo e Introdução, notadamente mais padronizadas, e à área de Ciência da Computação. A integração dos recursos e ferramentas desenvolvidos resultou no protótipo *SciPo* (*Scientific Portuguese*), implementado como um ambiente Web, que contempla tanto a composição *top-down* do texto (partindo do planejamento estrutural para a escrita propriamente dita) quanto *bottom-up* (partindo de um rascunho já escrito). Para avaliar o *SciPo* no seu contexto de uso, foram realizados experimentos com usuários reais, visando verificar a eficácia do modelo subjacente ao *SciPo*. Os resultados desses experimentos demonstraram que tal modelo de fato tem potencial no auxílio à escrita acadêmica.

Abstract

Aiding tools for writing in English have become more and more common, ranging from post-processing-only systems to more comprehensive ones, acting at various stages of the writing process, such as generation and organization of ideas and writing proper. Among these tools, there are those specific to the academic genre, notably more machine-tractable due to its rigorous standardization. In contrast, there are no such systems for Portuguese, but only post-processing tools, like spell-checkers, grammar-checkers and electronic dictionaries. In this work we focus on the issue of writing theses in Portuguese, aiming at a model for the creation and evaluation of writing tools for this genre, following an approach based on corpora and critiquing systems. For the sake of feasibility, we restrict our analysis to the Abstract and Introduction sections, acknowledgedly even more standardized, and to the Computer Science domain. The integration of the resources and tools developed resulted in the *SciPo* (*Scientific Portuguese*) prototype, a web-based application supporting composition either top-down (i.e. from structure planning to writing proper) or bottom-up (i.e. submitting a draft to be critiqued). In order to evaluate *SciPo* in use, we carried out experiments with real users with an emphasis on verifying the effectiveness of our underlying model. The results of these experiments show that it in fact has potential for helping academic writing.

Sumário

Sumário xi

Lista de Figuras xvii

Lista de Tabelas xxi

1 Introdução 1

2 A Estruturação de Textos Acadêmicos 7

2.1 Estrutura Esquemática Global 8

2.2 Detalhando o Resumo 9

2.3 Detalhando a Introdução 12

2.3.1 O Modelo de Swales 13

2.3.2 O Modelo de Weissberg e Buker 14

2.3.3 O Modelo de Booth, Colomb e Willians 14

2.3.4 O Modelo de Aluísio 15

2.3.5 Comentários sobre os Modelos 17

2.3.6 Detalhando os Elementos da Introdução 17

2.4 Detalhando os Materiais e Métodos 25

2.5 Detalhando os Resultados 27

2.6 Detalhando a Conclusão 28

2.7 Considerações Finais 30

3 Ferramentas de Auxílio à Escrita **31**

3.1 Writing Environment - WE 32

3.2 Writer's Assistant 33

3.3 Composer 34

3.4 Academic Writer 36

3.5 Abstract Helper 38

3.6 Mover 40

3.7 AMADEUS 41

3.7.1 Ferramenta de Referência 42

3.7.2 Ferramenta de Suporte 45

3.7.3 Ferramenta de Crítica 47

3.7.4 Ferramenta Tutorial 49

3.8 Comentários sobre as Ferramentas 50

3.9 Considerações Finais 51

4 Detecção Automática da Estrutura Esquemática de Textos Acadêmicos **55**

4.1 Critique 56

4.1.1 O Analisador Simbólico 57

4.1.2 Os Analisadores Probabilísticos 58

4.1.3 Avaliação Intrínseca 59

4.2 Mover 60

4.2.1 Avaliação Intrínseca 61

4.3 Argumentative Zoning (AZ) 61

4.3.1 Avaliação Intrínseca 63

4.4 Comentários sobre os Sistemas 65

4.5 Considerações Finais 66

5 Levantamento de Requisitos e Análise de Córpus **69**

5.1 Pesquisa com Potenciais-Usuários 70

5.1.1 Descrição do Questionário 70

5.1.2 Resultados da Pesquisa 70

5.1.3 Considerações sobre os Resultados 77

5.2 Coleta, Anotação e Análise do Córpus 77
5.2.1 Os Textos Coletados 78
5.2.2 Anotação Eletrônica do CorpuDT 79
5.2.3 Análise da Estrutura do CorpusDT 80
5.2.4 Anotação e Análise da Estrutura Esquemática dos Resumos e Introduções 81
5.2.5 Considerações sobre os Resumos e Introduções 87
5.3 Considerações Finais 87

6 Proposta de um Ambiente Computacional de Auxílio à Escrita Acadêmica 89

6.1 Apoio à Estruturação do Texto 90
6.1.1 As Bases de Exemplos 93
6.1.2 As Regras e o Processo de Crítica 95
6.1.3 As Regras e Medidas de Similaridade e a Recuperação de Exemplos Similares 96
6.2 Apoio à Estruturação a partir do Texto Pronto 99
6.2.1 Detecção Automática da Estrutura Esquemática de Textos Acadêmicos em Português 101
6.2.2 Material de Treinamento 101
6.2.3 Descrição dos Atributos Utilizados 103
6.2.4 Extração dos Valores dos Atributos 106
6.2.5 Descrição do Classificador Estatístico 114
6.2.6 Avaliação Intrínseca do *AZPort* 114
6.2.7 Considerações sobre o *AZPort* 117
6.3 Apoio à Escrita do Texto 118
6.4 Avaliação do Ambiente *SciPo* 119
6.4.1 Experimento 1 120
6.4.2 Experimento 2 122
6.5 Considerações Finais 123

7 Conclusões e Trabalhos Futuros 125

Referências Bibliográficas **129**

A Questionário Aplicado aos Potenciais Usuários de uma Ferramenta de Auxílio à Escrita Acadêmica **135**

B XML Schema para a Anotação do CorpusDT **139**

C Regras de Crítica Utilizadas no Ambiente *SciPo* **143**

C.1 Regras de Crítica para Resumos 143

C.1.1 Críticas de conteúdo 143

C.1.2 Críticas de ordem 144

C.1.3 Sugestões de conteúdo 146

C.1.4 Sugestões de ordem 147

C.2 Regras de Crítica para Introduções 148

C.2.1 Críticas de conteúdo 148

C.2.2 Críticas de ordem 149

C.2.3 Sugestões de conteúdo 151

C.2.4 Sugestões de ordem 152

D Lista de Expressões Regulares Utilizadas pelo *AZPort* **155**

E Questionários Utilizados na Avaliação do Ambiente *SciPo* **165**

E.1 Experimento 1 165

E.2 Experimento 2 167

F Orientações para a Anotação Manual do Corpus de Resumos Acadêmicos **171**

F.1 Antes da Anotação 172

F.2 Processo de Anotação 172

F.3 Descrição das Categorias 173

F.3.1 Contexto (C) 173

F.3.2 Lacuna (L) 173

F.3.3 Propósito (P) 174

F.3.4 Metodologia (M) 176

F.3.5 Resultado (R) 177
F.3.6 Conclusão (Co) 178
F.3.7 Estrutura (E) 179
F.4 Últimas observações 179

Lista de Figuras

2.1 Movimento Geral-Específico-Geral através da estrutura global do texto . . 9

2.2 Componentes do plano padrão para textos acadêmicos 9

2.3 Modelo de resumo típico 11

2.4 Modelo de resumo curto 12

2.5 Modelo de introdução CARS 13

2.6 Modelo para a estrutura de introduções 14

2.7 Modelo para a estruturação de introduções 15

2.8 Esquema detalhado para introduções 16

2.9 Modelo simplificado para introduções 18

2.10 Movimento Geral-Específico seguido na seção de contextualização 18

2.11 Modelo de ordenação dos possíveis elementos da seção de materiais e métodos 26

2.12 Categorias de materiais 26

2.13 Movimento Específico-Geral seguido na Conclusão 29

2.14 Possíveis passos para o início de uma conclusão 29

3.1 Ilustração da visualização de um esquema estrutural no Composer 36

3.2 Menu principal do Academic Writer 37

3.3 Tela principal do Mover 41

3.4 Exemplos de expressões padrão para introduções 43

3.5 Arquitetura da ferramenta de Referência 44

3.6 Arquitetura da ferramenta de Suporte 46

3.7 A abordagem de críticas 48

3.8 A arquitetura da Ferramenta de Crítica 49

4.1 Esquema de anotação utilizado no sistema Critique 56

4.2 Esquema de anotação utilizado pelo Mover 60

4.3 Esquema de anotação utilizado pelo AZ 63

5.1 Respostas de estudantes iniciantes e experientes com relação à escrita acadêmica 71

5.2 Respostas de estudantes iniciantes e experientes com relação à revisão dos textos produzidos 72

5.3 Respostas de estudantes iniciantes e experientes quanto ao tipo de texto utilizado como modelo para a escrita do seu próprio texto 72

5.4 Respostas de estudantes experientes e iniciantes quanto ao grau de dificuldade encontrado na escrita de cada seção 73

5.5 Respostas dos estudantes experientes quanto a relevância das funções de uma ferramenta computacional de auxílio à escrita 74

5.6 Respostas dos estudantes iniciantes quanto a relevância das funções de uma ferramenta computacional de auxílio à escrita 75

5.7 Respostas de estudantes iniciantes e experientes quanto aos aspectos operacionais de uma ferramenta de auxílio à escrita acadêmica 76

5.8 Modelo utilizado na anotação dos Resumos do CorpusDT 83

5.9 Modelo utilizado na anotação das Introduções do CorpusDT 86

6.1 Relação entre as três macro-funções do ambiente *SciPo* 90

6.2 Visão geral do processo de apoio à estruturação do texto 91

6.3 Apresentação de um exemplo de Resumo 92

6.4 Exemplo de resumo codificado em XML 94

6.5 Exemplo de estrutura codificada em Prolog 94

6.6 Processo de crítica implementado no ambiente *SciPo* 96

6.7 Resultado da aplicação das Regras de Crítica de Resumo à estrutura apresentada na Figura 6.5 97

6.8 Exemplo de resultado da recuperação de exemplos similares 100

6.9 Processo de apoio à estruturação a partir do texto redigido 100

6.10 Esquema de anotação utilizado no AZPort 101
6.11 Exemplos de expressões padrão 104
6.12 Etapas do processo de determinação dos atributos 106
6.13 Etapas do processamento sintático 111
6.14 Classificador Naive Bayes 114
6.15 Visão geral do processo de apoio à escrita do texto 118
6.16 Exemplo de formulário de edição e de inserção de texto reutilizável 119
6.17 Exemplo de uso do revisor ortográfico e gramatical 120

F.1 Esquema de classificação com as etiquetas que devem ser utilizadas na anotação 171

Lista de Tabelas

3.1 Adequação do tipo de ferramenta quanto ao conhecimento do usuário . . . 42

3.2 Resumo das ferramentas de auxílio à escrita investigadas 52

4.1 Performance de cada sistema individual (*Precision*, *Recall* e *F-Measure*) . . 59

4.2 Performance da *baseline*, do melhor sistema individual e do sistema de votação (*Precision*, *Recall* e *F-Measure*) 60

4.3 Matriz de Confusão: anotação manual vs. anotação Mover 62

4.4 Atributos utilizados pelo AZ 64

4.5 Desempenho do AZ e das *baselines*, calculado em comparação com um anotador humano 64

5.1 Número de trabalhos em cada área de pesquisa 78

5.2 Totais de palavras em dissertações e teses 79

5.3 Totais de palavras divididos por área de conhecimento 79

5.4 Dados sobre o tamanho em palavras dos Resumos 81

5.5 Dados sobre o tamanho em palavras das Introduções 82

5.6 Freqüência de componentes observada no córpus de Resumos 83

6.1 Resumo do conjunto de atributos 103

6.2 Exemplos de expressões regulares no formato `flex` e possíveis expressões reconhecidas 109

6.3 *Tagset* utilizado pelo MXPOST 110

6.4 Resultados da avaliação do processo sintático 113

6.5 Resultados da avaliação do processo sintático em termos de *Precision*, *Recall* e *F-Measure* 113

6.6 Distribuição das categorias 115

6.7 Desempenho do *AZPort* e das *baselines*, calculado em comparação a um anotador humano 115

6.8 Matriz de Confusão: anotação manual vs. anotação automática 115

6.9 *Precision*, *Recall* e *F-Measure* por categoria 116

6.10 Potencial de distinção de cada atributo em termos do Kappa entre um anotador humano e o *AZPort* 117

6.11 Resumo das ferramentas de auxílio à escrita investigadas, juntamente com o *SciPo* 123

Capítulo

1

Introdução

Sabe-se que produzir um bom texto acadêmico não é uma tarefa fácil, mesmo para escritores experientes. A sobrecarga cognitiva imposta sobre o escritor é bastante alta (Sharples & Pemberton, 1992), pois além de lidar com a complexidade inerente a qualquer processo de escrita, o escritor tem ainda que estar atento às convenções específicas do gênero acadêmico[1], referentes tanto ao tipo de informação a ser incluída no texto quanto à forma de apresentação dessas informações. Embora essas convenções apresentem um alto grau de padronização, muitas vezes não são claras o suficiente para o escritor, principalmente quando iniciante, o que dificulta a escrita e acarreta uma série de problemas no texto produzido, como o uso incorreto de padrões lexicais e de tempos verbais e organização textual ineficiente.

Nas duas últimas décadas, várias foram as ferramentas e ambientes de software produzidos para auxiliar o processo de escrita, acadêmica ou não, visando, sobretudo, melhorar a qualidade do texto produzido. Alguns sistemas oferecem um suporte mais abrangente, apoiando várias fases do processo de composição do texto, como a geração e organização de idéias. Como exemplos de tais sistemas, podemos citar o Academic Writer (Broady & Shurville, 2000), o Composer (Pemberton, Shurville, & Hartley, 1996), o Writer's Assistant (Sharples, Goodlet, & Clutterbuck, 1994) e o WE (Smith & Lansman, 1988). Outros sistemas oferecem um suporte mais específico, em geral de pós-processamento do texto, visando corrigi-lo tanto do ponto de vista gramatical, como da clareza, estilo e concisão das

[1]Neste trabalho, chamamos de "gênero acadêmico" aquele que abrange a divulgação de trabalhos acadêmicos em uma comunidade científica. Exemplos de textos desse gênero são os artigos para conferências ou periódicos, dissertações, teses, relatórios técnicos, etc.

idéias apresentadas, conforme levantamento descrito em Fontana *et al.* (1993). Exemplos de tais sistemas são os corretores gramaticais, como o Grammatik (RSI, 1992) e o Correct Grammar (WTG, 1991), os ambientes de referência e auto-instrução, como o HELPDISK! 2.6 e o English Language 2.0[2], e corretores de estilo, como o StyleWriter[3], além de vários dicionários e thesauri eletrônicos. Para o português, podemos citar o corretor gramatical ReGra (Martins *et al.*, 1998), que também incorpora características de corretor estilístico e é distribuído como parte das ferramentas *Office* da Microsoft.

Tais ferramentas de pós-processamento oferecem várias informações estatísticas sobre o texto, tais como o número e tamanho médio de parágrafos, sentenças e palavras do texto, e também estatísticas que avaliam o grau de legibilidade do texto. Além disso, oferecem regras para a correção da ortografia, da gramática, de adequação do estilo do texto a certo gênero (comercial, científico, jornalístico) ou a um padrão específico de uma companhia ou editora.

Embora esse tipo de auxílio seja extremamente útil para usuários com dificuldades de adequação à norma-padrão da língua, existe uma limitação no uso dessas ferramentas, pois elas proporcionam somente mecanismos de correção de um texto acabado. Entretanto, nem sempre um usuário é capaz de produzir um texto que possa vir a ser simplesmente melhorado. Além disso, as ferramentas de pós-processamento não resolvem uma dificuldade comum a muitos escritores: a produção do primeiro rascunho. Uma possível estratégia para minimizar esse problema seria a reutilização de sentenças, parágrafos e trechos maiores, como as seções de um artigo, compilados a partir de textos relevantes ao gênero em questão (Aluísio, 1995; Silva, 1999; Barros, 2000). Uma vertente importante de ferramentas que adotam tal estratégia é aquela formada por ferramentas de auxílio à escrita em língua estrangeira. Para os falantes do português, um gênero em que é evidente a necessidade de ferramentas como essas para a produção de textos é o acadêmico, já que a divulgação mais relevante desses trabalhos se dá prioritariamente em revistas especializadas e mundialmente veiculadas, que adotam o inglês como língua padrão. Um exemplo de ambiente dessa natureza é o AMADEUS - Amiable Article Development for User Support (Oliveira Jr., Caldeira, & Fontana, 1992; Fontana *et al.*, 1993; Aluísio, 1995; Aluísio & Oliveira Jr., 1995, 1996; Aluísio & Gantenbein, 1997; Aluísio *et al.*, 2001).

O AMADEUS é um ambiente acadêmico de auxílio e ensino da escrita técnica em inglês como língua estrangeira, que conta com várias ferramentas inter-relacionadas: Ferramenta de Referência, Ferramenta de Suporte, Ferramenta de Crítica e Ferramenta Tutorial, sendo que cada uma dessas ferramentas se aplica especificamente a um conjunto de problemas caracterizado por diferentes níveis de dificuldade no inglês e diferentes graus de experiência na escrita técnica. Outros sistemas voltados ao auxílio à escrita para não-nativos em inglês

[2]http://www.mantex.co.uk/software.htm [18-02-2002]
[3]http://www.editorsoftware.com [18-02-2002]

são o Composer (Pemberton, Shurville, & Hartley, 1996; Shurville, Hartley, & Pemberton, 1997), que não chegou a ser finalizado, e o Abstract Helper (Narita, 2000a,b). O projeto do Composer previa ferramentas de planejamento (ou geração de idéias) e composição de textos, e foi fundamentado em teorias cognitivas do processo de escrita. O Abstract Helper é composto fundamentalmente por uma base de exemplos de abstracts que podem ser reutilizados em vários níveis (estrutural, sentencial e lexical) e por facilidades para a recuperação desses exemplos. Comparando-se o Composer ao AMADEUS, pode-se dizer que o primeiro enfatiza a organização das idéias, enquanto o segundo enfatiza a definição de estratégias para a organização do texto escrito. Por outro lado, o Abstract Helper e o AMADEUS têm em comum a ênfase na reutilização de exemplos, mas enquanto o primeiro não oferece auxílio no planejamento do texto, o último apresenta capacidades avançadas e mais refinadas de especificação estrutural, incluindo um módulo de crítica (Silva, Pelizzoni, & Aluísio, 1998; Silva, 1999) da estrutura especificada pelo usuário.

Se, por um lado, as ferramentas destinadas a auxiliar a escrita na língua inglesa tornam-se cada vez mais comuns, por outro lado, ainda não são encontrados similares para a língua portuguesa. Só recentemente têm surgido ferramentas lingüísticas de pós-processamento, como corretores ortográficos, revisores gramaticais, dicionários para consulta, entre outras. Contudo, nenhuma dessas ferramentas, que focalizam o produto, agrupa todas as funcionalidades desejáveis de um ambiente de auxílio à escrita contemplando as várias etapas do processo de escrita, como as já existentes para o inglês.

A experiência adquirida pelo grupo atuante no NILC[4] desde 1993 na pesquisa e desenvolvimento de aplicações (revisores ortográfico e gramatical, sumarizadores de texto, tradutores, etiquetadores morfossintáticos, etc) e recursos lingüísticos (léxicos, córpus, etc), notadamente para o português brasileiro, somada à experiência no desenvolvimento do ambiente de auxílio e ensino da escrita técnica AMADEUS, tem motivado a investigação de ferramentas de auxílio à escrita do português que sejam mais direcionadas a um gênero particular de escrita (ao contrário dos ambientes de processadores de texto), tornando-se potencialmente mais úteis para uma determinada classe de usuários.

Não seria, no entanto, de grande utilidade uma simples adaptação do AMADEUS para a língua portuguesa, por dois fatores principais. Primeiramente, o AMADEUS visa auxiliar escritores não-nativos, enquanto uma ferramenta voltada para escrita em português se destinaria a escritores nativos, que podem ter necessidades diferentes. Em segundo lugar, o AMADEUS trata artigos científicos e é notória a solicitação para que pesquisadores brasileiros publiquem seus trabalhos científicos em revistas internacionais, portanto, em língua inglesa. Contudo, há uma classe de documentos acadêmicos que necessariamente devem ser escritos em português, cuja organização e redação têm sido objetos de estudo de profissionais da área de Letras (Rodrigues, 1998; Silva, 2002). Trata-se das teses e

[4]Núcleo Interinstitucional de Lingüística Computacional (www.nilc.icmc.usp.br).

dissertações acadêmicas.

É inegável que ferramentas de auxílio ao processo (composição) e produto (escrita) desses documentos seriam de grande valia tanto para os estudantes-autores de teses e dissertações, quanto para os orientadores que exigem qualidade naquilo que é o veículo oficial de divulgação de um trabalho científico desenvolvido, em geral, ao longo de vários anos. Além disso, o uso de tais ferramentas pode favorecer o aprendizado da escrita acadêmica pelo estudante-autor, que assim passa a ter maior consciência do processo envolvido na investigação científica, já que a escrita é a explicitação desse processo.

Dessa forma, nesta tese de doutorado, focalizamos o problema da escrita em português de teses e dissertações, visando uma modelagem computacional que segue uma abordagem baseada em córpus e em sistemas de crítica, para a criação e avaliação de ferramentas de apoio à criação de textos desse gênero. Além disso, objetivamos apoiar a estruturação do texto e a realização lingüística dessa estrutura de forma flexível, deixando o usuário livre para escolher entre dois modos de trabalho, a saber: (i) um processo *top-down*, que parte do planejamento estrutural para a escrita propriamente dita, como ocorre no AMADEUS; ou (ii) um processo *bottom-up*, em que se submete um texto já escrito à análise (detecção e crítica) automática da estrutura. Na verdade, tratam-se de pontos de partida distintos para um mesmo processo cíclico de refinamento, já que a estrutura detectada e criticada em (ii) pode ser aprimorada por meio dos recursos disponíveis em (i).

Como qualquer sistema de auxílio à escrita, e em particular à escrita acadêmica, para ser eficaz, deve levar em conta as particularidades da área de conhecimento em que está inserido o trabalho científico, é preciso definir um recorte no domínio de estudo. Neste trabalho, optamos por tratar textos da área de Ciência da Computação. Além disso, uma vez que muito esforço é necessário para o tratamento de cada uma das seções de um texto, mesmo em um domínio específico, conforme mostram duas das ferramentas desenvolvidas no projeto AMADEUS (que tratam apenas a introdução), decidimos lidar com duas seções de texto reconhecidamente importantes e problemáticas, além de mais padronizadas (Swales, 1990): o Resumo e a Introdução. A escolha dessas seções também se deve aos resultados obtidos em uma pesquisa realizada com estudantes de pós-graduação, em que tais seções foram apontadas entre as que oferecem maior grau de dificuldade na escrita (Seção 5.1, Cap. 5). Apesar de ter sido estabelecido esse recorte, vale ressaltar que a metodologia empregada suporta um processo incremental e evolutivo, possibilitando o tratamento futuro de todas as seções previstas para um texto acadêmico.

Seguindo a linha de pesquisa baseada em córpus, herdada do AMADEUS, para apoiar o processo *top-down* de especificação estrutural, foi necessária a compilação de um córpus especialmente dedicado a essa pesquisa, bem como a definição de modelos de anotação apropriados. A anotação manual do córpus incluiu a identificação das diferentes seções, da função argumentativa dos diversos trechos e das expressões padronizadas, entre outros.

Esse Córpus de Dissertações e Teses (CorpusDT) serviu tanto para a análise de padrões de escrita específicos da comunidade focalizada, como para a realização de uma análise crítica, visando identificar os principais problemas que poderiam ser minimizados pelo auxílio de uma ferramenta computacional.

Complementarmente, para permitir a análise estrutural automática de textos prontos necessária ao processo *bottom-up*, foram investigados sistemas de identificação da estrutura esquemática ou retórica[5] de textos (Burstein, Marcu, & Knight, 2003; Anthony & Lashkia, 2003; Teufel & Moens, 2002). Essa investigação culminou na implementação de um classificador estatístico, denominado *AZPort*, que foi treinado com os resumos do CorpusDT para detectar a estrutura esquemática de resumos acadêmicos em português. Embora esta tese trate as seções Resumo e Introdução, apenas o Resumo foi utilizado na concepção e implementação do *AZPort*, por constituir um estudo de caso suficiente para a avaliação inicial da viabilidade do uso desse tipo de classificador em uma ferramenta de auxílio à escrita. Além disso, o Resumo contém componentes semelhantes aos da Introdução, embora de forma condensada e, dessa forma, espera-se que a extensão desse classificador para introduções possa ocorrer de forma escalável, a partir do trabalho já realizado.

A integração dos recursos e ferramentas desenvolvidos ao longo deste trabalho resultou num protótipo denominado *SciPo* (abreviação para *Scientific Portuguese*), que visa auxiliar especialmente escritores iniciantes na escrita de resumos e introduções de teses ou dissertação em português, no domínio da Ciência da Computação. O protótipo *SciPo* foi implementado como um ambiente Web e contempla tanto a composição *top-down* do texto (funcional para Resumos e Introduções), quanto a composição *bottom-up*, ou seja, partindo de um rascunho já escrito (atualmente funcional apenas para Resumos, conforme mencionado).

Para avaliar o protótipo *SciPo* no seu contexto de uso, foram realizados experimentos com usuários reais com o objetivo de verificar o quanto o modelo subjacente ao *SciPo* (incluindo seus recursos e ferramentas) é de fato efetivo no auxílio à escrita acadêmica em português. Os resultados desses experimentos demonstraram que tal modelo tem potencial no auxílio à escrita acadêmica.

Dessa forma, podemos resumir as contribuições deste trabalho de doutorado em quatro pontos principais:

- na investigação do discurso acadêmico em português, especificamente na área de Ciência da Computação, incluindo a construção de córpus e a extração de modelos e padrões para nortear o tratamento computacional da língua nesse contexto;

[5]Neste trabalho, entende-se por estrutura esquemática ou retórica de um texto o conjunto de elementos (ou componentes retóricos) que descrevem a organização (funcional) do texto, relacionando suas partes de forma a atribuir-lhe sentido.

- na investigação de classificadores capazes de detectar a estrutura esquemática de textos acadêmicos em português, com base em um conjunto de componentes retóricos pré-estabelecidos, com potencial uso em outras aplicações, como sumarização automática e anotação retórica de córpus;

- na validação da proposta deste doutorado via a implementação de um ambiente de auxílio à escrita acadêmica, que poderá ser estendido num processo análogo e incremental;

- no oferecimento de um ponto de partida de onde deverá surgir uma família de novos desafios na área de escrita acadêmica em português apoiado por computador.

Em todos os casos, acreditamos que este trabalho constitui uma iniciativa inédita em se tratando da escrita acadêmica na língua portuguesa.

No próximo capítulo (Capítulo 2), é apresentada a estruturação de textos acadêmicos em português, abordando não só as seções tratadas neste trabalho (Resumo e Introdução), mas também outras seções que fazem parte da estrutura padrão de um texto acadêmico. No Capítulo 3, são apresentadas as ferramentas de auxílio à escrita descritas na literatura, em especial, o AMADEUS, enfocando os problemas tratados e as soluções adotadas por cada uma. Os sistemas de detecção automática da estrutura esquemática investigados são apresentados no Capítulo 4, com enfoque nos métodos de aprendizagem utilizados e nos resultados obtidos. No Capítulo 5, são detalhadas duas etapas cruciais da metodologia de desenvolvimento deste trabalho: o levantamento de requisitos para um ambiente de auxílio à escrita acadêmica junto a potenciais usuários e a análise de córpus. No Capítulo 6 é apresentado o protótipo *SciPo*, um ambiente computacional de auxílio à escrita acadêmica em português, detalhando seus componentes e recursos, incluindo o classificador *AZPort*, assim como os resultados da avaliação desse protótipo. Finalmente, no Capítulo 7 são apresentadas as conclusões deste trabalho e suas futuras extensões.

Capítulo

2

A Estruturação de Textos Acadêmicos

"Escrever é parte inerente ao ofício do pesquisador" (Feitosa, 1991). O trabalho do cientista não se esgota nas descobertas que faz. É de sua responsabilidade comunicar os seus resultados, suas descobertas, suas criações. Sendo assim, a escrita acadêmica caracteriza-se como um processo de comunicação muito importante, pois uma descoberta científica torna-se reconhecida através das publicações de seus resultados. Entretanto, escrever não implica em, necessariamente, comunicar com eficiência. Um trabalho escrito sem cuidado, sem a preocupação em guiar o leitor proporcionando-lhe uma leitura esclarecedora e agradável, não está comunicando como deveria, até porque ninguém se dispõe a ler um trabalho confuso e de leitura desestimulante. Dessa forma, para que um trabalho seja bem escrito e, conseqüentemente, comunique bem o seu propósito, é preciso atentar a uma série de fatores, tanto do seu conteúdo quanto da sua estrutura. Na literatura, muito já foi discutido sobre a pesquisa científica como um processo ordenado e documentado, sendo que a maioria dos autores discute o que o conteúdo de um texto acadêmico deve relatar e as etapas que devem ser seguidas para se chegar à informação que será incluída em um trabalho, discutindo-se a metodologia da pesquisa com alguma informação sobre a estrutura (Eco, 2000; Tachizawa & Mendes, 2000; Vieira, 1999; Severino, 1996; Feitosa, 1991; Castro, 1981; Barrass, 1979; Salomon, 1977; Hegenber, 1976; Rey, 1972). Porém, poucos trabalhos discutem detalhadamente a estrutura esquemática de um texto acadêmico em português. Estruturas esquemáticas (ou superestruturas) foram definidas por Kintsch e van Dijk (1978)[1] como sendo os com-

[1]Kintsch, W., van Dijk, T.A. (1978). Toward a model of text comprehensionand production. Psychological Review, 85, pp. 363-394 apud Aluísio (1995).

ponentes do discurso que definem sua forma e ajudam a identificar seu gênero. Alguns autores tratam a estruturação do texto acadêmico mais profundamente, como Huckin & Olsen (1991); Weissberg & Buker (1990); Swales (1990); Trimble (1985), todos trabalhos voltados para o inglês. Apesar da diferença de língua, a estrutura esquemática discutida nesses trabalhos pode ser aplicada também para o português. Embora se saiba que existem muitos fatores que podem influenciar a composição de um texto, como fatores culturais (Hartley & Paris, 1996; Vander Linden & Scott, 1995; Paris & Evans, 1994), estudos como de Taylor & Tingguang (1991) indicam que apesar de existirem diferenças na organização e na elaboração dos componentes da estrutura esquemática, devido a diferenças entre nacionalidades, culturas e disciplinas, os trabalhos acadêmicos compartilham uma mesma estrutura. De fato, analisando-se os trabalhos citados, fica claro que as diferenças de idioma não interferem na estrutura mais global de um texto acadêmico. O conteúdo apresentado a seguir é uma compilação das observações sobre a estruturação esquemática de textos acadêmicos que foram discutidas nos trabalhos citados anteriormente.

2.1 Estrutura Esquemática Global

Segundo Rey (1972), antes de se começar a escrever deve-se considerar a estrutura do texto que se pretende redigir. Para isso, deve-se estabelecer um esquema que permita expor as idéias de maneira sistemática e lógica, reunindo em cada item assuntos correlatos, sem risco de omitir ou de repetir as mesmas coisas ao longo do trabalho. Um trabalho científico pode enquadrar-se, em geral, dentro de um esquema que já se tornou clássico pela simplicidade, pelo desenvolvimento metódico e por abranger aspectos essenciais de uma comunicação científica desse gênero. Todos os autores pesquisados (Eco, 2000; Tachizawa & Mendes, 2000; Vieira, 1999; Severino, 1996; Feitosa, 1991; Castro, 1981; Barrass, 1979; Salomon, 1977; Hegenber, 1976; Rey, 1972; Huckin & Olsen, 1991; Weissberg & Buker, 1990; Swales, 1990; Trimble, 1985) concordam quanto à estrutura esquemática que um texto acadêmico deve seguir. Essa estrutura pode ser enunciada como Introdução — Desenvolvimento — Conclusão, sendo que o Desenvolvimento pode desdobrar-se nas seções de Materiais e Métodos e Resultados, ou ainda Materiais e Métodos, Resultados e Discussão. Essa forma de estruturação tem como objetivo apresentar o texto a partir do contexto no qual ele está inserido. Para isso, cada um dos componentes da estrutura desempenha um papel bem definido, que será discutido mais adiante. Em linhas gerais, essa estrutura deve guiar o leitor e fazer com que ele siga o movimento geral-para-específico, realizado na Introdução, e específico-para-geral, realizado na Conclusão, conforme ilustrado na Figura 2.1.

Note que a Figura 2.1 apresenta um outro componente chamado Resumo. O Resumo aparece de uma forma destacada na figura por ser um componente independente do res-

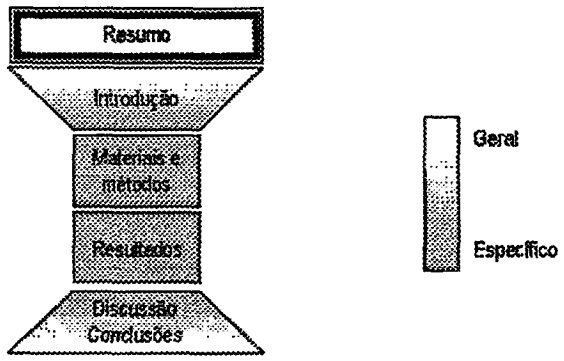


Figura 2.1: Movimento Geral-Específico-Geral através da estrutura global do texto (Weissberg & Buker, 1990)

tante da estrutura. Devido ao fato de conter informações relativas a todo o texto, tanto gerais como específicas, o Resumo não segue o movimento da estrutura global apresentada. Sendo assim, pode-se definir um plano padrão para estruturação esquemática de textos acadêmicos, contendo os seguintes componentes: Resumo, Introdução, Materiais e Métodos, Resultados, Discussão e Conclusão. A Figura 2.2 indica em linhas gerais o que deve ser incluído em cada um desses componentes.

Resumo	*Sucinta informação das principais descobertas*
Introdução	*O que fez o autor? Por quê?*
Materiais e Métodos	*Como fez?*
Resultados	*O que foi encontrado?*
Discussão	*Interpretação dos resultados*
Conclusões	*Conclusões do trabalho*

Figura 2.2: Componentes do plano padrão para textos acadêmicos (Barrass, 1979)

É importante ressaltar que, enquanto a estrutura esquemática global é ditada por um esquema mais fixo, a estrutura mais detalhada não é (Smith & Lansman, 1988). Sendo assim, serão discutidas nas próximas seções as particularidades de cada um dos componentes e quais os respectivos papéis dentro da estrutura esquemática global.

2.2 Detalhando o Resumo

O resumo é um componente independente, isto é, não interfere no movimento geral-específico-geral que o texto como um todo deve realizar. Trata-se de um componente complementar que pode preceder o texto a fim de dar ao leitor uma visão geral dos principais assuntos tratados. No entanto, essa característica não o torna um componente menos importante, já que muitos leitores se limitam a ler o título e o resumo de um

trabalho (Barrass, 1979). Sendo assim, o resumo deve ser redigido com muito cuidado, de forma a ser completo, interessante e informativo, dispensando a consulta ao restante do texto para que o leitor tenha a idéia do que trata o trabalho e, ao mesmo tempo, estimulando o interesse pela leitura do texto completo. A Sumarização Automática (SA) é uma subárea de PLN que se dedica exclusivamente ao estudo de resumos (tanto os acadêmicos quanto de outros gêneros de texto), de modo que muito já foi discutido nessa área sobre essa seção. Um dos aspectos dos resumos que é discutido em SA é a sua informatividade. De acordo com esse aspecto, um resumo pode ser classificado em três categorias (Mani, 2001; Feitosa, 1991; Rey, 1972):

- Informativo (ou autocontido), quando contém todas as informações apresentadas no texto e dispensa a leitura desse último;
- Indicativo, quando não dispensa a leitura do texto completo, pois apenas descreve a natureza, a forma e o propósito do trabalho, cuja matéria não é passível de uma apresentação abreviada. Pode ser visto como um índice, que indica os assuntos tratados no restante do trabalho.
- Crítico, quando formula julgamento sobre o trabalho.

Dependendo da função do resumo, um ou outro tipo é mais adequado. Alguns autores (Feitosa, 1991; Vieira, 1999; Oliveira, 2001) indicam que o resumo constituinte de um texto acadêmico deve ser preferencialmente informativo, descrevendo quais os propósitos, resultados e conclusões do trabalho, em vez de indicativo, dizendo, por exemplo, que o resultado se encontra no texto. Embora outros autores não explicitem o caráter informativo do resumo de um texto acadêmico, a preferência por um resumo informativo pode ser notada nos modelos de resumo sugeridos.

Os resumos de quase todas as áreas de estudo são escritos de uma maneira muito similar (Weissberg & Buker, 1990). Os tipos de informação incluídos e a ordem em que aparecem são muito convencionais, de modo que podem ser enunciados como modelos de resumo. Tais modelos objetivam guiar o escritor no sentido do tipo de informação que deve ser incluída em um bom resumo e da ordem que tais informações devem aparecer.

A seguir são apresentados dois exemplos de resumos, um indicativo e outro informativo. Esses exemplos são trechos autênticos extraídos do córpus construído como uma das etapas deste trabalho de doutorado[2].

Resumo indicativo:

> *"Um estudo empírico visando avaliar a eficácia em revelar erros, a dificuldade de satisfação e o custo de aplicação do critério Análise de Mutantes é apresentado neste*

[2]Esse córpus, chamado CorpusDT, é apresentado no Capítulo 5.

trabalho. Eficácia e custo também foram avaliados para os critérios Potenciais-Usos, comparando-se assim os resultados obtidos para esses critérios. A especificação e implementação de uma estratégia de minimização de conjuntos de casos de teste adequados ao critério Análise de Mutantes também é apresentada".

Resumo informativo:

"Este trabalho apresenta um protótipo de ferramenta, a Html2Hip, que proporciona um ambiente de importação e adaptação de documentos descritos segundo o padrão HTML (HyperText Markup Language) para a representação interna do SASHE (Sistema de Autoria e Suporte Hipermídia para Ensino), que se baseia na estruturação de objetos multimídia, segundo a hierarquia de classes proposta pelo MCA (Modelo de Contextos Aninhados). Além disso, este trabalho estendeu a capacidade do editor de nós de informação do tipo texto do protótipo anterior no que concerne ao processamento de arquivos-texto descritos pelo padrão RTF (Rich Text Format). Dessa forma, o SASHE tornou-se capaz de processar e organizar materiais instrucionais preparados em seu próprio ambiente, no ambiente WWW (World-Wide Web), bem como em processadores de texto comuns".

Vários autores apresentam modelos de resumo, entre eles Booth, Colomb, & Williams (2000), Huckin & Olsen (1991) e Weissberg & Buker (1990). Embora cada autor tenha a sua forma de expressar o modelo, existe um consenso sobre os elementos típicos e sua ordem. O modelo de Weissberg & Buker é apresentado aqui, na Figura 2.3, por ser o mais detalhado e abranger também os outros modelos citados.

1. Alguma informação de contextualização (*background*)
2. A principal atividade do estudo (seu propósito) e seu escopo
3. Algumas informações sobre a metodologia usada no estudo
4. Os resultados mais importantes do estudo
5. Uma afirmação de conclusão ou recomendação

Figura 2.3: Modelo de resumo típico (Weissberg & Buker, 1990)

Observando a Figura 2.3, fica fácil perceber que um resumo conterá informações de outras seções do texto. Dessa forma, é conveniente que o resumo seja elaborado após o término do trabalho (Barrass, 1979). Embora seja o componente que aparece primeiro, geralmente é o último a ser escrito. Vale notar que os cinco elementos citados no modelo de Weissberg & Buker direcionam para a composição de um resumo informativo, em que todos os elementos principais de um texto são citados. Esses elementos são semelhantes aos encontrados nas introduções, como será visto adiante (Seção 2.3), porém escritos de forma tão concisa quanto possível. Há ocasiões, no entanto, em que só uma escrita

concisa não basta, pois existe um limite pequeno de palavras a serem usadas que deve ser respeitado, como o caso de resumos para artigos publicados em revistas. Nesses casos, um tipo mais curto de resumo pode ser escrito por meio da eliminação ou combinação de alguns dos elementos citados anteriormente, conforme apresentado na Figura 2.4.

1. Propósito + metodologia
2. Resultados principais
3. Conclusão/Recomendação (opcional)

Figura 2.4: Modelo de resumo curto (Weissberg & Buker, 1990)

O resumo curto geralmente é escrito usando apenas dois ou três elementos dos cinco citados no modelo anterior, com a ênfase sendo colocada nos resultados do estudo. A informação de *background* nesse caso é retirada, sendo apresentados como primeiros elementos os propósitos e a metodologia utilizada, combinadamente. Em seguida deve ser feito um resumo dos resultados mais importantes do trabalho e finalmente pode ser colocada uma conclusão ou recomendação em uma ou duas sentenças. Essa conclusão ou recomendação é opcional e pode ser ocultada caso seja necessário.

2.3 Detalhando a Introdução

A introdução é o componente que inicia o movimento geral-específico-geral que o texto deverá seguir até o seu têrmino (Figura 2.1). Ela é a responsável por fazer com que o leitor seja guiado de forma coerente e agradável, partindo-se de um contexto geral até o assunto específico tratado pelo escritor. Dessa forma, ela própria segue o movimento geral-específico. Salomon (1977) diz que a função da introdução é apresentar o assunto do trabalho, colocando o tema abordado pelo escritor dentro de um quadro de referência teórica atualizado. No entanto, escrever uma introdução que realize tal função de maneira bem sucedida não é tarefa fácil, especialmente para escritores acadêmicos iniciantes. Mesmo escritores experientes admitem sentirem mais dificuldade em escrever a introdução do que o restante do trabalho (Swales, 1990).

Sendo assim, grande parte dos autores que discutem a estrutura esquemática de textos científicos dedica uma atenção maior ao componente Introdução. Enquanto alguns autores apresentam observações que podem ser tomadas como diretrizes no momento da escrita, outros apresentam modelos mais completos, descrevendo os elementos típicos encontrados em boas introduções, pois assim como os resumos, as introduções tendem a apresentar um padrão de elementos e de ordenação desses elementos.

Feitosa (1991) apresenta uma discussão interessante sobre a estrutura da introdução. A autora argumenta que pode haver situações em que o modelo típico de introdução deixa

de ser adequado e que é importante considerar o papel, ou seja, a função que a introdução deve desempenhar para que o processo de comunicação seja efetivo. Sendo assim, em vez de seguir um modelo pré-estabelecido de introdução, deve-se considerar que ela tem a função de levar o leitor a ler o texto pelo interesse nas informações nele contidas. Assim, a função da introdução é transmitir ao leitor informações que o deixem a par dos assuntos tratados no restante do trabalho e que o entusiasmem a continuar a leitura. Dessa forma, nos casos em que um esquema tradicional de introdução for adequado, então ele deve ser usado. Caso contrário, é melhor procurar construir uma introdução apropriada, para não haver risco de que a introdução se torne um elemento isolado do restante do trabalho.

Um esquema tradicional de introdução pode ser resumido em três elementos ou estágios: Contextualização Geral, Revisão Bibliográfica e Propósito do trabalho. Todos os modelos de introdução estudados baseiam-se nesses elementos, porém diferem no detalhamento de cada elemento e também na inclusão de outros elementos que complementam cada modelo.

Barrass (1979) descreve um modelo simples de introdução focalizando os três elementos citados. Segundo esse autor, a introdução deve trazer uma clara apresentação do problema focalizado, indicando o escopo do trabalho, isto é, estabelecendo o contexto no qual o trabalho se desenvolve (Contextualização Geral). Deve incluir uma breve referência a qualquer nota preliminar e outras investigações relevantes, próprias do autor ou alheias, para mostrar – se for o caso – de que modo o trabalho a ser relatado está baseado em trabalhos anteriores (Revisão Bibliográfica). Deve dizer também o que especificamente é tratado no trabalho e os motivos que levaram a sua realização (Propósito).

Outros autores apresentam modelos de introdução mais detalhados, em que são apresentados não só os elementos ou estágios do modelo, mas também os passos que compõem cada elemento/estágio. Como exemplos podemos citar os modelos de Swales (1990), Weissberg & Buker (1990), Booth, Colomb, & Williams (2000) e Aluísio (1995). A seguir é apresentado um breve resumo de cada um desses modelos.

2.3.1 O Modelo de Swales

Em Swales (1990) é apresentado um modelo para a estruturação de introduções chamado CARS – *Create a Research Space*. Esse modelo foi derivado da análise de um córpus de introduções das áreas da física, biologia/medicina e ciências sociais. É um dos modelos de introdução mais comentados (Aluísio, 1995; Huckin & Olsen, 1991; Weissberg & Buker, 1990) e é constituído de três movimentos, conforme apresentado na Figura 2.5, sendo que cada movimento é refinado em outros passos.

- **Movimento 1** Estabelecendo um território
  - **Passo 1** Mostrando proeminência e/ou
  - **Passo 2** Generalizando o tópico e/ou
  - **Passo 3** Revisando pesquisas anteriores
- **Movimento 2** Estabelecendo um nicho
  - **Passo 1A** Contra-argumentando ou
  - **Passo 1B** Indicando uma lacuna (gap) ou
  - **Passo 1C** Colocando questões ou
  - **Passo 1D** Continuando uma tradição
- **Movimento 3** Ocupando o nicho
  - **Passo 1A** Indicando propósito(s) ou
  - **Passo 1B** Descrevendo as principais características da pesquisa
  - **Passo 2** Indicando principais resultados
  - **Passo 3** Indicando a estrutura do artigo

Figura 2.5: Modelo de introdução CARS (Swales, 1990)

2.3.2 O Modelo de Weissberg e Buker

Weissberg & Buker (1990) apresentam um modelo básico, contendo cinco estágios para modelar introduções. Seu modelo também foi derivado da análise de córpus contendo textos das áreas das ciências exatas, humanas e biológicas. Como pode ser observado na Figura 2.6, cada estágio é novamente subdivido em passos. A numeração dos passos em seqüência indica que eles podem ser empregados dessa forma, enquanto que a indicação dos passos com números e letras indica exclusão mútua.

- **Primeiro Estágio:** Fornece um contexto para o problema a ser analisado
  1. Fatos relacionados à área geral da pesquisa
  2. Identificação da subárea
  3. Indicação do tópico
- **Segundo Estágio:** Revisa aspectos do problema já estudados por outros autores
  1. Citações onde a informação é enfatizada
  2. Citações sobre o trabalho de vários autores
  3. Nível de atividade da pesquisa na área
  4. Citações onde o autor é enfatizado
- **Terceiro Estágio:** Indica a necessidade de mais investigação na área
  - **1A.** A literatura revisada é inadequada, ou
  - **1B.** Conflito/Problema não resolvido: um *gap* na teoria ou na prática, ou
  - **1C.** Questões ainda não consideradas na área, ou
  - **1D.** Extensão/Melhoria do tópico
- **Quarto Estágio:** Indica o propósito ou objetivo do estudo
  - **1A.** Orientação ao artigo, ou
  - **1B.** Orientação à pesquisa

Quinto Estágio: (Opcional) Valor/Justificativas da realização do estudo

Figura 2.6: Modelo para a estrutura de introduções (Weissberg & Buker, 1990)

2.3.3 O Modelo de Booth, Colomb e Willians

Booth, Colomb, & Williams (2000) apresentam um modelo menos discutido do que o de Swales e o de Weissberg & Buker, porém bastante interessante, pois confirma os três movimentos básicos da introdução, os quais os autores chamam de pontos de vista. Para Booth et al. uma introdução pode ser vista como um conjunto de três pontos de vista chamados de *Base Comum*, *Ruptura* e *Resolução*, que aparecem quase sempre nessa ordem. A Figura 2.7 mostra o refinamento desse modelo.

1 – Base comum:	**Tipos de Abertura** ° Uma declaração geral. ° Um acontecimento ou caso. ° Uma citação ou fato estimulante. **Contexto** ° Compreensão compartilhada sobre o estado atual do problema ou antecedentes tidos como certos.
2 – Ruptura:	**Objeção:** *mas, contudo, por outro lado,* etc. **Declaração do problema** ° CONDIÇÃO de ignorância, pouca compreensão, etc. ° CUSTO/BENEFÍCIO de deixar a condição não resolvida ou de solucioná-la.
3 – Resolução:	**Declaração da resposta** ° Proposição principal ou proposição de lançamento

Figura 2.7: Modelo para a estruturação de introduções (Booth, Colomb, & Williams, 2000)

No modelo apresentado, CONDIÇÃO representa a má compreensão ou ignorância sobre um problema, isto é, uma lacuna no conhecimento, um conflito não explicado ou uma discrepância, uma falta de conhecimento ou entendimento. O CUSTO/BENEFÍCIO de uma condição pode ser tanto os custos da ignorância da solução de tal problema como os benefícios de sua solução. A "proposição principal" enuncia a solução para a condição. A "proposição de lançamento" promete que a solução será apresentada no decorrer do texto. No entanto, tal proposição deve ir além de simplesmente introduzir o tópico. Ela deve sugerir os esboços conceituais da solução e anunciar um plano (metodologia utilizada). Nesse caso, a proposição principal é apresentada na Conclusão.

Embora a ordem apresentada no modelo seja comum, pode haver casos em que o escritor deseje alterá-la. Os autores fazem três sugestões:

- A *Base Comum* é opcional, e pode ser omitida em casos específicos;
- A *Ruptura* normalmente contém tanto o CUSTO quanto a CONDIÇÃO, mas, se os leitores estão familiarizados com o problema tratado, pode conter apenas um deles;

- A *Resolução* deve declarar uma proposição principal ou uma proposição de lançamento, de preferência a primeira;
- Um planejamento explícito do texto (*Outline*) pode ser colocado depois da proposição.

2.3.4 O Modelo de Aluísio

Aluísio (1995) definiu um esquema detalhado para a estruturação de introduções baseada em dois modelos tradicionais, o de Swales e o de Weissberg & Buker, e na análise de um córpus de introduções de artigos da área da Física Experimental. Como resultado da análise de córpus foi proposto um modelo de oito componentes subcategorizados em estratégias, que cobre todos os componentes dos modelos de Swales e de Weissberg & Buker e adiciona o componente "Metodologia". Além disso, são deixadas explícitas as várias estratégias utilizadas pelos componentes para desenvolver suas mensagens características. A Figura 2.8 apresenta o esquema detalhado para introduções da área da Física Experimental descrito em (Aluísio, 1995) e (Aluísio & Oliveira Jr., 1996).

No modelo acima, têm-se estratégias enumeradas em cada componente. Aquelas que podem ser escolhidas mais de uma vez na escrita de uma introdução são identificadas somente com numerais, embora certas combinações possam não ser razoáveis. Os números seguidos de letras indicam exclusão mútua.

O esquema proposto não impõe uma ordem convencional para os componentes, pois sua função é apresentar as informações que aparecem em uma introdução. Entretanto, é muito razoável que eles apareçam nessa ordem.

2.3.5 Comentários sobre os Modelos

Como se pode perceber, os modelos apresentados são coerentes entre si. Eles apresentam similaridades tanto em relação aos componentes quanto à ordem em que aparecem, embora não exista um mapeamento direto de um modelo para o outro. Apesar da similaridade, algumas particularidades podem ser observadas em cada modelo.

O modelo de Weissberg & Buker, por exemplo, dá uma ênfase maior à contextualização da grande área de trabalho do que os modelos de Swales e de Booth, Colomb & Willians. Esses autores focalizam mais o tópico de pesquisa, evitando uma grande contextualização. Esse padrão se mostra mais adequado no caso de introduções menores ou em introduções destinadas a um público-alvo mais homogêneo em termos de conhecimento da área. Quando o tamanho da introdução não for limitado ou o público alvo for heterogêneo, o modelo de Weissberg & Buker parece ser mais adequado por conter um estágio de contextualização mais abrangente.

Componente 1 – Contextualização

1. Introduzir o tópico de pesquisa a partir da área de pesquisa
2. Familiarizar termos ou objetos ou processos
3. Argumentar sobre a proeminência do tópico/área

Componente 2 – Revisão da literatura

1. Revisão histórica da área
2. Tendências atuais na área
3. Citações organizadas da área geral para o tópico
4. Progresso na área
5. Requisitos para o progresso na área
6. Estado da arte
7. Citações e *gaps* cíclicos
8. Citações agrupadas por abordagem

Componente 3 – Necessidade de mais investigação/ Tipos de *gaps*

1. Existência de conflitos ou problemas não resolvidos
2. Restrições em trabalhos anteriores
3. Questões ainda não consideradas

Componente 4 – Propósito da pesquisa

1. Indicar o propósito principal
   - 1A. Resolver um conflito entre autores
   - 1B. Apresentar uma nova abordagem ou metodologia ou técnica
   - 1C. Apresentar melhorias/avanços em um tópico da literatura
   - 1D. Apresentar uma extensão de um trabalho anterior do autor
   - 1E. Apresentar uma abordagem alternativa
   - 1F. Apresentar um trabalho comparativo
2. Especificar o propósito
3. Introduzir mais propósitos

Componente 5 – Metodologia

1. Indicar critérios ou condições
2. Descrever materiais e métodos
3. Justificar escolhas por métodos e materiais

Componente 6 – Principais Resultados

1. Apresentação de resultados
2. Comentários sobre os resultados

Componente 7 – Valor da Pesquisa

Componente 8 – Estrutura do trabalho

- 1A. Indicar as seções do artigo
- 1B. Apresentar a lista dos tópicos a serem abordados

Figura 2.8: Esquema detalhado para introduções (Aluísio & Oliveira Jr., 1996)

O modelo de Aluísio deixa a cargo do escritor a decisão do tipo de estratégia que será usada em cada componente da introdução. No caso da contextualização, pode ser tanto uma estratégia mais abrangente como "Introduzir o tópico de pesquisa a partir da área de pesquisa" quanto uma estratégia mais restrita ao tópico como "Argumentar sobre a proeminência do tópico/área". Um dos fatores que proporciona essa flexibilidade é o fato de o modelo não ter como objetivo impor uma estrutura, mas simplesmente apresentar os componentes típicos de uma introdução. Outro fator que contribui para que o modelo seja flexível é que ele engloba tanto as estratégias do modelo de Swales quanto as do modelo

de Weissberg & Buker.

Um diferencial do modelo de Aluísio é a introdução do componente Metodologia, que não aparece explicitamente nos outros modelos estudados, embora alguns autores, como Huckin & Olsen (1991), sugiram que esse componente exista na introdução. Esse componente é mais característico das introduções de trabalhos que envolvem pesquisa experimental, em que os materiais e métodos utilizados são enfocados com uma relevância maior.

Quanto à ordem dos componentes, tanto Swales quanto Weissberg & Buker observam que existem outras formas de se organizar os componentes em uma introdução. Weissberg & Buker comentam que nem sempre os autores arranjam os estágios de uma introdução na ordem exata proposta pelo modelo, mas que o plano por eles fornecido é comum e o mais fácil para escritores iniciantes.

2.3.6 Detalhando os Elementos da Introdução

Partindo do modelo proposto em Aluísio (1995), em que uma introdução pode ser estruturada a partir de oito componentes ordenados numa seqüência "geral para específico" (Figura 2.8), e das recomendações feitas em Weissberg & Buker (1990), foi adotada uma versão simplificada de modelo de introdução dividido em oito estágios, conforme mostrado na Figura 2.9.

Estágio 1.	Estabelece um contexto que ajuda os leitores a entenderem como a pesquisa se situa num campo de estudo maior
Estágio 2.	É feita uma revisão bibliográfica, ou seja, são apresentados aspectos do problema que já foram estudados por outros pesquisadores
Estágio 3.	Indica a necessidade de mais investigação na área
Estágio 4.	Indica os objetivos/propósitos do estudo
Estágio 5.	(opcional) Descreve a metodologia utilizada na realização do trabalho
Estágio 6.	(opcional) Descreve os principais resultados encontrados
Estágio 7.	(opcional) Dá uma justificativa para se empreender o estudo em questão, afirmando o valor do trabalho
Estágio 8.	(opcional) Indica a estrutura do trabalho

Figura 2.9: Modelo simplificado para introduções

Nas subseções seguintes, são detalhados cada um dos oito componentes citados, listando algumas diretrizes e observações importantes para que o componente realize sua função dentro da introdução de forma satisfatória. É importante destacar que grande parte das informações mais detalhadas foi retirada de Weissberg & Buker (1990).

Primeiro Estágio: Contextualização

O contexto de uma introdução deve ser escrito de forma a fornecer aos seus leitores o *background* necessário para o entendimento do tópico particular da pesquisa em relação a uma área de estudo geral. Para que isso seja feito, pode-se começar com afirmações amplamente aceitas sobre a área na qual se está trabalhando. Então, passo a passo, o leitor deve ser guiado para ficar mais próximo do tópico de pesquisa específico do trabalho. Isso pode ser feito em apenas algumas sentenças ou em vários parágrafos, dependendo do tipo de trabalho a ser escrito e do estilo de escrita do autor. A Figura 2.10 ilustra o movimento "geral para específico" seguido nesse estágio.

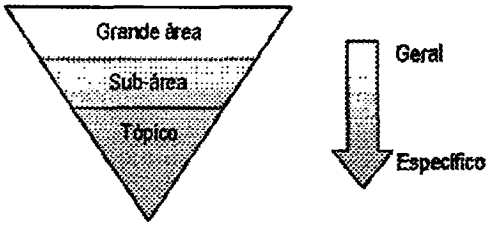


Figura 2.10: Movimento Geral-Específico seguido na seção de contextualização (Weissberg & Buker, 1990)

Weissberg & Buker (1990) apresentam uma analogia interessante, em que o estágio de contextualização é pensado como um processo em que primeiro se estabelece um "universo" para os leitores; então se isola uma "galáxia" dentro desse universo; e finalmente, guia-se os leitores para uma "estrela" na galáxia. Essa "estrela" é o tópico específico de pesquisa. Esse processo pode ser resumido da seguinte maneira:

- Comece com afirmações aceitas do fato relacionado à área geral de pesquisa (seu universo);
- Dentro da área geral, identifique uma subárea que inclua o tópico de pesquisa (sua galáxia);
- Indique o tópico (sua estrela).

Para que o leitor seja guiado suavemente através das idéias apresentadas nesse primeiro estágio, os escritores geralmente conectam as sentenças fazendo uso de informação conhecida e informação nova, alternadamente. Isso é feito colocando-se informação já conhecida pelo leitor no começo das sentenças e deixando a informação nova para o final da sentença. Veja no exemplo a seguir[3], como a segunda sentença é conectada à primeira por meio desse mecanismo.

[3]Todos os exemplos apresentados neste capítulo são trechos autênticos e foram extraídos do CorpusDT.

"Dentre as técnicas de verificação e validação, o teste é, sem dúvida, a atividade mais utilizada. Essa atividade [**info. conhecida**] *apresenta-se bastante onerosa podendo, em alguns casos, consumir 40% dos custos de desenvolvimento do software".* [**info. nova**]

O trecho a seguir exemplifica o movimento de contextualização descrito anteriormente. É importante notar como o leitor vai sendo guiado de um contexto bem geral para um contexto mais específico.

"O crescimento do mercado de software a cada dia acarreta o aumento do uso de técnicas de desenvolvimento, muitas vezes informais. [**contexto mais geral**] *A manutenção de tais softwares torna-se problemática, uma vez que a documentação associada ao software, na maioria das vezes, não está de acordo com o código implementado. Dessa forma, quando diante da manutenção do produto, o engenheiro de software encontra uma documentação informal e incompleta, que não reflete o software existente.* [**contexto intermediário**] *Nesse contexto é que se encontra a Engenharia Reversa de Software, com o propósito de recuperar as informações de projeto perdidas durante a fase de desenvolvimento, e de documentar o real estado do software".* [**contexto mais especializado**]

Segundo Estágio: Revisão Bibliográfica

No primeiro estágio da introdução se estabelece um contexto para o tópico de pesquisa em questão. No segundo estágio se revê os resultados encontrados por outros pesquisadores e que já tenham sido publicados na área de pesquisa de interesse. Por esta razão, o segundo estágio é comumente chamado de Revisão Bibliográfica. Esse estágio constitui-se essencialmente de uma coleção organizada de referências ou citações de outros trabalhos, os quais estão listados em uma seção separada no final do trabalho (normalmente chamada *Referências Bibliográficas*).

O estágio de revisão bibliográfica serve basicamente a três funções importantes. Primeiro, esse estágio dá continuidade ao processo iniciado no primeiro estágio, de dar ao leitor informação de *background* necessária para entender o estudo em foco no trabalho. Segundo, ele garante ao leitor que o autor do trabalho está familiarizado com as pesquisas relevantes que têm sido realizadas na sua área de pesquisa. Terceiro, esse estágio estabelece o estudo relatado como um elo na corrente de pesquisa que está desenvolvendo e aumentando o conhecimento em um campo de atuação.

Em alguns tipos de trabalhos, é comum existir uma seção de revisão bibliográfica escrita separadamente da introdução (Barrass, 1979). No caso específico de teses e dissertações, pode ser conveniente incluir um levantamento da literatura relevante anteriormente

publicada como uma seção independente (em um ou mais capítulos). Isso coloca a tese/dissertação em um quadro de referência e expõe o conhecimento existente que poderá servir de base para futuras investigações.

Tanto as referências bibliográficas apresentadas na introdução, como a seção "Revisão Bibliográfica" (se existir) precisam dar ao leitor o quadro indispensável em que se situa a investigação do trabalho sendo relatado e fornecer os elementos necessários para o desenvolvimento da argumentação. Essa revisão da literatura não deve ser uma seqüência impessoal de resumos de outros trabalhos; mas deve demonstrar que os trabalhos foram examinados e criticados objetivamente (Tachizawa & Mendes, 2000).

Seja na introdução, seja como um capítulo separado, a revisão bibliográfica será composta por um conjunto de citações de outros autores, de forma que é preciso atentar a dois fatores: o foco dado às citações e a ordem em que elas aparecem. A seguir, é discutido qual o enfoque que se deve dar a uma citação e como se pode ordenar um conjunto de citações de forma coerente.

Foco da Citação

Quando se cita trabalhos de outros autores, pode-se escolher focalizar a informação mostrada por aquele autor ou focalizar o próprio autor. O primeiro tipo de foco é chamado de "proeminente na informação", porque a informação é apresentada com especial importância. O nome do autor e a data da publicação são então colocados entre parênteses no fim da sentença. Dessa forma, fontes de informação mais completas podem ser encontradas na lista ordenada alfabeticamente de referências, no final do trabalho. Um tipo alternativo de citação proeminente na informação usa números entre parênteses ao invés do nome do autor e da data. O número se refere à lista de referências enumerada no final do trabalho. Geralmente, citações proeminentes na informação são utilizadas no início da revisão bibliográfica e referem-se a fatos da área geral do estudo em questão aceitos por toda a comunidade científica. Essas citações podem também aparecer no primeiro estágio (Contextualização). Veja o exemplo:

> *"Na engenharia reversa, o sistema geralmente é o ponto inicial do processo" (Chikofsky, 1990).*

Conforme a revisão bibliográfica prossegue, as citações passam a se referir a estudos mais proximamente relacionados ao estudo em foco no trabalho. Nesse tipo de citação é dada maior ênfase ao nome do autor. Nesse caso, o nome do autor passa a ser o foco da sentença, seguido pela data ou pelo número da citação entre parênteses, e então seguido pela informação em si. Esse tipo de citação é chamado de citação "proeminente no autor". Em geral, quando se está reportando resultados obtidos por outros autores em

seus trabalhos, utiliza-se o passado como tempo verbal enquanto que, quando se reporta teorias, opiniões e recomendações, utiliza-se o presente. O exemplo abaixo ilustra o uso de uma citação proeminente no autor.

> *"Rugaber (1992) afirma que a maior parte do esforço de desenvolvimento de software é gasto na manutenção de sistemas existentes e não no desenvolvimento de sistemas novos"*

Ordem das Citações

O grupo de citações da revisão bibliográfica também deve ser ordenado em uma seqüência lógica, visando facilitar o entendimento pelo leitor. Em Weissberg & Buker (1990) são sugeridas três estratégias para o arranjo das citações, sendo que cada estratégia se adapta melhor a um caso específico.

A estratégia mais comumente usada é a de arranjar as citações da revisão bibliográfica na ordem das mais indiretamente relacionadas para aquelas mais diretamente relacionadas ao estudo em questão, seguindo assim a estratégia utilizada no primeiro estágio que guia o leitor do geral para o específico.

Existem casos, porém, em que as outras estratégias são mais eficientes. Em uma revisão bibliográfica na qual se descreve a história da pesquisa em uma determinada área, o arranjo das citações em ordem cronológica pode ser mais indicado. Já no caso de existir um grande número de citações a serem incluídas em uma revisão bibliográfica, como em uma tese ou dissertação, por exemplo, pode-se agrupá-las de acordo com as diferentes abordagens ao problema tomadas pelos diferentes autores. As citações referentes a cada abordagem podem então ser ordenadas cronologicamente ou da geral para a específica, ficando a critério do autor.

Terceiro Estágio: Indicação da Lacuna

O terceiro estágio finaliza a revisão bibliográfica apontando uma lacuna (*gap*), isto é, uma área de pesquisa importante que não foi investigada por outros autores ou que não tenha sido suficientemente desenvolvida. Normalmente, esse estágio é constituído por apenas uma ou duas sentenças. Abaixo são apresentadas três alternativas para se escrever a(s) sentença(s) do terceiro estágio.

- Pode-se indicar que a literatura previamente estudada, descrita no segundo estágio, é inadequada porque um aspecto importante da área de pesquisa tem sido ignorado por outros autores, isto é, um determinado problema não está resolvido.

- Pode-se indicar que há um conflito não resolvido entre os autores dos estudos prévios relacionados ao tópico de pesquisa em foco, isto é, existe uma controvérsia. Essa controvérsia pode ser um desentendimento teórico ou metodológico.
- Pode-se indicar que o exame da literatura sugere uma extensão do tópico, ou levanta uma nova questão de pesquisa não considerada previamente por outros pesquisadores em seu campo de atuação.

Dessa forma, indicando algum tipo de falha deixada pelos estudos anteriores, o terceiro estágio prepara o leitor para focalizar o estudo em questão no trabalho e, de certa forma, justifica a realização do estudo. O exemplo abaixo mostra como o escritor pode indicar a existência de uma lacuna.

*"**No entanto, existe pouca informação disponível** sobre como escrever uma tese ou um artigo em português".*

↑ Indicação da Lacuna ↑ Tópico em questão

Repare como é sinalizado no início da sentença que uma lacuna na informação dada até então será anunciada. Nesse exemplo, isso se dá pelo uso do *No entanto.* Outros sinalizadores podem ser utilizados, como os que aparecem no modelo de Booth, Colomb & Willian (*mas, contudo, por outro lado, etc.*)

Quarto Estágio: Apresentação do Propósito

O quarto estágio serve para mostrar, o mais concisamente possível, o objetivo específico do trabalho que está sendo escrito. Dessa forma, este estágio de anúncio do propósito segue diretamente o terceiro estágio, pois é a resposta à necessidade de pesquisa adicional na área de estudo em foco, expressa no estágio anterior.

A apresentação do propósito deve estar diretamente ligada à questão da pesquisa na qual está baseado o estudo. Apesar de não ser obrigatória a inserção explícita dessa questão, o propósito deve ser escrito de modo que o leitor possa inferir qual é a questão por trás do estudo relatado.

A apresentação de propósito pode ser escrita seguindo-se uma de duas alternativas de orientação. As alternativas são:

- A orientação da apresentação do propósito pode ser dirigida ao próprio trabalho, isto é, pode se referir ao artigo, tese, dissertação ou relatório que vai comunicar a informação sobre a pesquisa em questão, conforme mostrado no exemplo abaixo. Note que o tempo verbal utilizado é o presente.

"O principal objetivo desta dissertação é o projeto e implementação de um mecanismo de tratamento de exceções para a construção de sistemas orientados a objetos confiáveis". [Orientado ao trabalho]

- A orientação do propósito pode ser dirigida à atividade de pesquisa, em outras palavras, ao próprio estudo em vez do trabalho escrito, como pode ser visto no exemplo abaixo. Note que o tempo verbal utilizado nesse caso é o passado.

  "O principal objetivo deste trabalho de mestrado foi a investigação de uma estrutura adequada de hiperdocumento para apoiar a documentação requerida durante o processo de engenharia reversa de software". [Orientado à pesquisa]

Quinto Estágio: Breve Descrição da Metodologia

Embora esse seja um estágio opcional, é recomendável que se diga logo na introdução qual foi a metodologia usada para o desenvolvimento do trabalho. No entanto, não se deve descrever toda a metodologia, pois isso deverá ser apresentado na seção Materiais e Métodos, na qual a metodologia utilizada é descrita passo a passo. Um exemplo de citação de metodologia é apresentado abaixo.

> *"A engenharia reversa foi desenvolvida baseando-se no método de engenharia reversa Fusion RE/I, e os produtos obtidos foram inseridos em uma hiperbase".*

Sexto Estágio: Citação dos Principais Resultados Encontrados

Um trabalho científico não deve criar um suspense sobre os seus resultados, dessa forma é recomendável que se apresente na introdução os principais resultados alcançados, embora esse estágio seja visto como opcional por alguns autores. Um exemplo de uma citação de resultados é mostrado abaixo.

> *"A partir das análises realizadas foi estabelecido um conjunto de diretrizes que visam auxiliar o processo de desenvolvimento de uma simulação distribuída conservativa".*

Sétimo Estágio: Afirmação do Valor

O sétimo estágio é usado para justificar a pesquisa relatada, baseado em algum possível valor ou benefício que o estudo possa trazer para outros pesquisadores do campo de pesquisa ou para as pessoas que trabalham em situações práticas relacionadas a esse campo.

Essa afirmação de valor não é incluída em todas as introduções. No caso específico de teses e dissertações, esse estágio pode incluir uma lista concisa das principais contribuições

do trabalho para a ampliação do conhecimento da área investigada. Esse estágio também deve aparecer quando se escreve a introdução de uma proposta de tese. A afirmação do valor é também comumente incluída em relatórios de pesquisa escritos para descrever um projeto conduzido com dinheiro proveniente de fontes externas. Em artigos escritos em jornais e revistas técnicas esse estágio é quase sempre omitido.

Existem dois pontos de vista a partir dos quais a afirmação de valor pode ser escrita:

- A afirmação de valor pode ser escrita do ponto de vista dos benefícios práticos que podem resultar da aplicação dos resultados da pesquisa.
- Ou pode-se escrever a afirmação de valor para enfatizar a importância teórica do estudo no avanço do estado do conhecimento em uma área de pesquisa específica.

O exemplo abaixo mostra uma afirmação de valor orientada à prática. Não foram encontrados no córpus exemplos de afirmação de valor orientada à importância teórica. Acreditamos que isso se deve ao fato de os textos pertencerem a uma área científico-tecnológica (Ciência da Computação), em que é mais comum destacar a importância prática da pesquisa.

> *"O sistema Animbs permite a associação de uma geometria ao MBS sendo simulado e utiliza os dados produzidos pela simulação do SD/FAST para criar uma animação do comportamento do sistema mecânico e, dessa forma, melhorar a análise de dados feita pelos usuários do SD/FAST".* [Orientação Prática]

Oitavo Estágio: Estrutura do Trabalho (Outline)

No final da introdução, o autor pode dar ao leitor um roteiro do que será encontrado no restante do trabalho. Esse componente é chamado de *outline* O *outline* pode ser tanto uma indicação das seções do artigo (ou capítulos de uma tese/dissertação) como um apanhado dos assuntos abordados no trabalho na ordem em que aparecem. Abaixo segue como exemplo parte de um *outline* de uma dissertação.

> *"Este trabalho está organizado em 7 capítulos. No Capítulo 2 é feita uma revisão sobre sensores, detalhando os aspectos relacionados à robótica e as técnicas comumente empregadas para medição de distâncias. Uma revisão sobre fusão de sensores é feita no Capítulo 3. (...) No Capítulo 6 são descritos os experimentos realizados, analisados os resultados obtidos com cada algoritmo e comparados os diversos métodos aplicados para a fusão dos sensores. A conclusão deste trabalho é apresentada no Capítulo 7".*

2.4 Detalhando os Materiais e Métodos

Normalmente, depois da introdução, é apresentada uma segunda seção de texto chamada de Materiais e Métodos, principalmente em trabalhos que envolvem pesquisa experimental. Nessa seção são descritos os passos que foram seguidos para conduzir o estudo e os materiais que foram usados em cada passo. Essa seção é especialmente útil aos leitores que querem saber como a metodologia utilizada influenciou os resultados, ou aos leitores que estão interessados em replicar ou estender o trabalho descrito.

Segundo Barrass (1979), a seção de materiais e métodos deve incluir detalhes suficientes capazes de assegurar que a repetição da investigação por alguém com experiência na mesma área leve à obtenção de dados similares. Barrass também observa que quando o trabalho científico for uma tese, o autor deve indicar de que modo abordou o problema e qual foi sua contribuição para a resolução desse problema.

A parte principal da seção Materiais e Métodos é a descrição dos passos procedimentais tomados no estudo e dos materiais utilizados em cada passo. Entretanto, existem outros elementos comumente descritos nessa seção. Esses elementos não seguem um padrão fixo, assim como a ordem na qual eles aparecem também não é fixa, com exceção dos procedimentos e materiais, que sempre aparecem e sempre estão nessa ordem. Apesar dessa flexibilidade, existe uma ordem convencional que pode ser seguida e que pode ser um bom modelo, conforme mostrado na Figura 2.11.

Embora todos os elementos da Figura 2.11 possam estar presentes, o foco se encontra nos procedimentos e materiais. A descrição dos passos (procedimentos) que foram seguidos durante a execução do estudo deve ser escrita de forma clara para que o leitor possa repetir sem problemas os seus passos. Naturalmente, a melhor maneira de descrever um procedimento é apresentado-o passo-a-passo ou cronologicamente.

Possíveis Elementos
Overview do experimento (Projeto)
População / Amostra
Locação
Restrições / Condições Limites
Técnica de Amostragem
Procedimentos
Materiais
Variáveis
Tratamento Estatístico

Figura 2.11: Modelo de ordenação dos possíveis elementos da seção de materiais e métodos (Weissberg & Buker, 1990)

A descrição dos procedimentos e dos materiais pode ser feita em conjunto ou separadamente. O autor pode escolher descrever os materiais e equipamentos utilizados juntamente com os passos em que esses materiais foram utilizados, alternando-se métodos e materiais. Por materiais entende-se qualquer item usado para dar andamento ao projeto de pesquisa. Weissberg & Buker apresentam uma categorização de materiais, como apresentado na Figura 2.12.

Materiais
Equipamento de laboratório
Equipamento de campo
Material humano ou animal
Substancias naturais
Materiais fabricados
Pesquisas, questionários e testes
Modelos computacionais
Modelos matemáticos

Figura 2.12: Categorias de materiais (Weissberg & Buker, 1990)

Os materiais usados em um estudo também podem ser descritos de forma separada dos métodos. Um exemplo para o uso desse arranjo é quando várias partes diferentes de equipamento convencional de laboratório são usadas para executar um procedimento rotineiro. Mais comumente, no entanto, os materiais e os métodos são descritos em uma forma integrada, geralmente com ambos os elementos mencionados na mesma sentença.

Se os materiais que foram usados no estudo são conhecidos pelos pesquisadores no campo de pesquisa em questão, o convencional é apenas citar os nomes desses materiais para identificá-los. Entretanto, se no experimento forem usados materiais especialmente projetados para o estudo ou não convencionais, é comum se escrever uma descrição detalhada de cada material. Nesse caso, as seguintes informações devem ser incluídas:

1. Visão geral: este passo constitui-se de uma ou duas sentenças dando uma idéia geral do material e do seu propósito na pesquisa.

2. Descrição das partes principais: neste passo, cada parte ou característica principal do material é descrita em uma seqüência lógica.

3. Descrição funcional: este último passo mostra como as várias características descritas no passo 2 interagem.

2.5 Detalhando os Resultados

Como o próprio nome sugere, na seção Resultados são apresentados os resultados do estudo e um breve comentário sobre eles. Os resultados do trabalho devem ser apresentados numa ordem lógica – que pode ser diversa da ordem em que foi desenvolvida a investigação (Barrass, 1979). O importante é que o leitor consiga acompanhar a exposição dos dados e entender como os resultados foram alcançados.

Alguns autores chamam essa seção de Resultados e Discussão, indicando assim que são apresentados comentários mais extensos sobre os resultados. Outra forma utilizada é deixar tais comentários para próxima seção, chamada de Discussão, e na seção Resultados apresentar os resultados juntamente com alguns comentários breves.

Na apresentação dos resultados é comum o uso de figuras, quadros, tabelas ou outros elementos gráficos seguidos de texto escrito, descrevendo os dados que estão representados. É importante destacar que todos esses elementos gráficos devem sempre estar acompanhados de um texto explicativo, pois enquanto a demonstração dos resultados por meio do uso de figuras, gráficos e outros apresenta esses resultados de forma completa em termos matemáticos, o texto que os acompanha ajuda o leitor a focar sua atenção nos aspectos mais importantes dos resultados e a interpretá-los. Dessa forma, um não pode vir sem o outro. Outro ponto importante é que sempre deve haver no texto uma referência aos elementos gráficos, relacionando-os diretamente ao tópico/resultado específico que o elemento gráfico representa.

O texto que apresenta os resultados pode ser dividido em três elementos básicos de informação: (1) uma sentença localizando o(s) elemento(s) gráfico(s) onde os resultados podem ser vistos; (2) algumas sentenças que apresentam os aspectos mais importantes dos resultados e (3) outras sentenças que comentam esses resultados. Uma outra forma, mais compacta, de se escrever a apresentação dos resultados seria combinar os elementos (1) e (2). Assim, eles passam a ser apresentados como afirmações que apresentam os resultados mais importantes e indicam entre parênteses o elemento gráfico onde esses resultados podem ser vistos.

Quanto ao comentário dos resultados, elemento (3), Weissberg & Buker (1990) sugerem duas formas de ordenação. Numa delas, colocam-se comentários curtos (uma ou duas sentenças) após cada resultado significativo que for mencionado. Outra forma é deixar para fazer todos os comentários no final, após terem sido apresentados todos os resultados. A primeira maneira é chamada "padrão alternado" e a segunda de "padrão seqüencial". O padrão alternado é melhor quando se tem muitos resultados individuais com comentários específicos sobre cada resultado. O padrão seqüencial é usado quando existem muitos resultados individuais aos quais apenas um comentário geral se aplica. Em alguns casos,

nenhum comentário é colocado, deixando para serem todos feitos na seção de discussão. Algumas funções podem ser atribuídas aos comentários dos resultados. As mais comuns são:

- generalizar os resultados;
- explicar possíveis razões para os resultados;
- comparar os resultados com resultados de outros estudos.

2.6 Detalhando a Conclusão

A seção Conclusões é a última do texto e, dessa forma, tem a função de finalizar o assunto. Durante a conclusão, a discussão deve ser uma consideração objetiva dos resultados apresentados na seção anterior e deve conduzir com naturalidade às suas principais conclusões. A conclusão deve relacionar as diversas partes da argumentação e unir as idéias desenvolvidas anteriormente no trabalho. É por isso que se diz que, em certo sentido, a conclusão é uma volta à introdução, ficando no leitor a impressão de estar diante de um sistema harmônico, acabado em si mesmo (Salomon, 1977).

Assim como na Introdução, os pesquisadores usam essa seção para examinar seu trabalho no contexto maior do seu campo de estudo. No entanto, ao contrário da introdução, essa seção guia o leitor da informação específica apresentada nas seções de materiais e métodos e resultados para uma visão mais geral de como os resultados devem ser interpretados. Dessa forma, pode-se dizer que na introdução parte-se do geral para o específico enquanto na conclusão parte-se do específico para o geral. A Figura 2.13 apresenta um esquema do movimento específico para geral seguido na Conclusão.

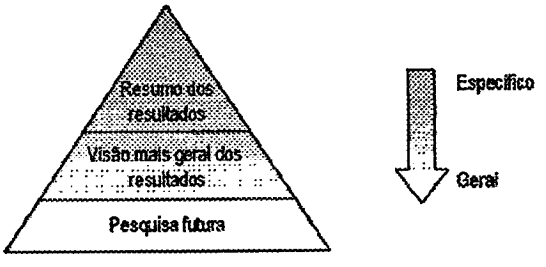


Figura 2.13: Movimento Específico-Geral seguido na Conclusão (Weissberg & Buker, 1990)

A informação que se inclui nessa seção depende em grande parte dos resultados do estudo apresentado, mas o movimento "específico para geral" é uma convenção que a maioria dos escritores segue. Os tipos de informação que se pode incluir nessa seção não são fixos, porém os primeiros elementos são tipicamente aqueles que se referem mais

diretamente ao estudo e aos resultados. Weissberg & Buker (1990) apresentam um modelo descrevendo os possíveis elementos de informação que aparecem no início da seção de conclusões, mostrado na Figura 2.14.

1. Uma referência ao principal propósito e às hipóteses do estudo
2. Uma revisão dos resultados mais importantes, mesmo que esses resultados não suportem a hipótese original do estudo ou não concordem com os resultados de outros pesquisadores
3. Possíveis explicações sobre os resultados (resumidamente)
4. Limitações do estudo que restringem a generalização dos resultados

Figura 2.14: Possíveis passos para o início de uma conclusão (Weissberg & Buker, 1990)

Conforme a seção prossegue, o escritor deve ir distanciando a atenção do leitor dos resultados específicos do estudo e começar a focalizar de forma mais generalizada a importância que o estudo pode ter para outros pesquisadores no campo de estudo, relacionando seu trabalho no contexto maior da sua área de pesquisa. Os últimos elementos de informação na seção de discussão são, portanto, afirmações gerais sobre o estudo, como sugerido por Weissberg & Buker (1990):

1. Implicações do estudo (generalização dos resultados);
2. Recomendações para pesquisas futuras e possíveis aplicações práticas.

É importante ressaltar que essa ordem de elementos da seção Conclusão não é estritamente seguida pelos autores, mas o movimento de elementos mais específicos para elementos mais gerais é convenção.

Nessa seção, mais do que em qualquer outro lugar do trabalho escrito, os escritores tornam explícitas suas próprias visões sobre o estudo e sobre os resultados. Os escritores podem tomar uma posição com respeito às explicações, implicações, limitações ou aplicações dos resultados. Quando isso acontece, é possível perceber partes de sentenças que expressam a opinião do escritor, conforme mostrado no exemplo abaixo. Note que, embora o autor não explicite o sujeito da sentença, fica clara a sua opinião sobre o tópico.

"Ao final desta tese conclui-se que aplicar corretamente o escalonamento de processos não é uma tarefa fácil"

↑

Posição do autor

Além de desempenhar o papel de fecho de uma demonstração, a conclusão deve servir para abrir novos horizontes, para apontar caminhos, para despertar novas questões ou dúvidas, enfim, para inserir o trabalho no fluxo da busca científica que o transcende (Feitosa, 1991). Por isso algumas vezes usa-se o título "Conclusões e Trabalhos Futuros".

2.7 Considerações Finais

Neste capítulo foi feita uma revisão bibliográfica sobre a estrutura de textos acadêmicos. Embora pouco tenha sido encontrado na literatura especificamente sobre a estrutura de textos em português, foram feitas adaptações de referências em inglês, uma vez que outros estudos indicam que a estrutura global dos textos acadêmicos não varia em sua forma, e sim na sua organização na escrita em nível de parágrafo.

Dessa forma, foram apresentadas diretrizes encontradas para a escrita em português e também foi apresentada a estrutura global para textos acadêmicos (Resumo, Introdução, Materiais e Métodos, Resultados, Conclusões). Cada uma das etapas desse plano global foi detalhada, segundo recomendações de Weissberg & Buker (1990), com ênfase nas seções tratadas neste trabalho (Resumo e Introdução).

A realização dessa pesquisa serviu para mostrar, entre outras coisas, que a estrutura esquemática de textos acadêmicos em português é muito parecida com a estrutura usada em inglês. Não há como negar que existe uma variação no conteúdo devido às diferenças culturais e da própria língua, porém essas diferenças não se manifestam no nível da estrutura global quando se trata desse gênero de texto.

A realização desse levantamento bibliográfico serviu também para indicar o pouco desenvolvimento das pesquisas na área de estruturação de textos acadêmicos em português, ao contrário do que acontece com textos em inglês, em que se têm modelos já bem estabelecidos.

No próximo capítulo é apresentado um levantamento bibliográfico sobre ferramentas de auxílio à escrita, em especial escrita acadêmica, que enfocam diferentes aspectos do processo de escrita.

Capítulo

3

Ferramentas de Auxílio à Escrita

Um texto será considerado bom se ele servir bem ao seu propósito e estiver de acordo com as expectativas de sua audiência. Os textos acadêmicos possuem ambos os aspectos bem definidos, ou seja, possuem um propósito bem específico – relatar/informar sobre uma determinada pesquisa – e uma audiência também bem específica/homogênea – a comunidade científica (Sharples & Pemberton, 1992).

Uma das expectativas dos acadêmicos é encontrar textos claros, que tenham um determinado conjunto de componentes tidos como essenciais. Por essa razão, o texto científico acaba por seguir um padrão rigoroso de estruturação. Com o intuito de esclarecer as particularidades desse gênero, vários autores procuram descrever os padrões encontrados e considerados bons, seja por meio de modelos que emergiram de análises (descrição), seja por meio de diretrizes (prescrição), conforme foi apresentado no Capítulo 2.

O fato de o texto acadêmico seguir um padrão de estruturação que pode ser reproduzido abriu caminho para que fossem desenvolvidas ferramentas computacionais que apoiassem tanto o processo de composição do texto quanto o processo de revisão do texto produzido. No entanto, o processo de escrita, seja ela científica ou não, envolve várias etapas, como discutido por Hayes & Flower (1980), de modo que desenvolver um ambiente que suporte o processo como um todo de modo eficiente é bastante complexo. Sendo assim, as ferramentas desenvolvidas até hoje enfocam alguns aspectos desse processo, geralmente atacando um problema específico ou uma categoria de problemas que pode ser tratada em conjunto.

Grande parte das ferramentas desenvolvidas atualmente enfoca o pós-processamento

do texto, isto é, trabalha sobre um texto já construído, geralmente um rascunho, com o objetivo de corrigir erros e melhorar aspectos como clareza e concisão. Nesse grupo incluem-se ferramentas como os corretores ortográficos, gramaticais e estilísticos. Sistemas de pós-processamento são extremamente úteis e de fato ajudam a melhorar a qualidade de um texto em vários aspectos. Entretanto, para se utilizar tais sistemas, o usuário deve ser capaz de produzir um primeiro rascunho. Swales (1990) afirma que mesmo a composição do primeiro rascunho de uma introdução não é fácil até para escritores experientes. Sendo assim, existe uma outra classe de ferramentas que visa apoiar o usuário no processo de composição e não apenas na revisão. Exemplos desses sistemas são o Writer´s Assistant (Sharples & Pemberton, 1992; Sharples, Goodlet, & Clutterbuck, 1994), o WE (Smith & Lansman, 1988; Lansman, Smith, & Weber, 1993), o Composer (Pemberton, Shurville, & Hartley, 1996; Shurville, Hartley, & Pemberton, 1997), o Academic Writer (Broady & Shurville, 2000), o Abstract Helper (Narita, 2000b,a), o Mover (Anthony & Lashkia, 2003) e o AMADEUS (Oliveira Jr., Caldeira, & Fontana, 1992; Fontana *et al.*, 1993; Aluísio, 1995; Aluísio & Oliveira Jr., 1995, 1996; Aluísio & Gantenbein, 1997; Aluísio *et al.*, 2001).

Broady & Shurville (2000) argumentam que sistemas como os citados podem ser pensados como ferramentas cognitivas, uma vez que tentam minimizar a sobrecarga cognitiva imposta ao escritor no momento da escrita de um texto científico. Tal sobrecarga é atribuída ao esforço realizado para a integração dos diferentes níveis de exigências do processo de escrita (Sharples & Pemberton, 1992). Essas exigências podem ser resumidas em quatro classes: exigências "conceituais" (o que escrever e por que), "estruturais" (como as idéias devem ser organizadas), "de expressão" (geração e monitoramento da realização lingüística apropriada) e "estratégicas" (planejar, monitorar e avaliar a realização correta da escrita).

A seguir, são apresentadas descrições das principais características de cada um dos sistemas, de modo a esclarecer os níveis de auxílio ao usuário e como isso é feito. Por fim, é detalhado o ambiente AMADEUS, uma vez que tal ambiente serviu como um dos pontos de partida para o desenvolvimento deste projeto.

3.1 Writing Environment - WE

O WE é um sistema que, baseado no modelo de escrita de Hayes & Flower (1980), tenta simular o processo cognitivo de criação de um texto, fornecendo recursos para que esse processo seja realizado de maneira mais controlada, favorecendo assim o resultado final, ou seja, produzindo um texto melhor.

Sua principal característica é fornecer suporte para que o usuário faça uma espécie

de *brainstorming* no momento do planejamento do texto, atuando como um organizador de idéias. Uma vez que o usuário tem o suporte do sistema para esse planejamento, diminuindo assim a sua sobrecarga cognitiva, espera-se que ele sinta-se mais livre para a geração e a organização das idéias.

O WE baseia-se num processo de composição divido em três estágios: planejamento, escrita e revisão, sendo que o sistema possui módulos específicos para cada etapa do processo. No total, o WE possui quatro módulos, dois deles dedicados ao planejamento do texto, um à edição e um à revisão. Esses módulos são respectivamente chamados de *Network Mode*, *Tree Mode*, *Edit Mode* e *Text Mode*.

O *Network Mode* permite que o usuário mapeie suas idéias em nós e faça as ligações entre elas, relacionando esses nós. Nesse módulo não há preocupação quanto à organização estrutural do texto, mas quanto ao relacionamento existente entre as idéias. Outro módulo do sistema se encarrega de organizar os nós criados no *Network Mode* numa forma hierarquizada, o *Tree Mode*. Nesse módulo é representada a organização estrutural dos nós, definindo os tópicos do texto. O *Tree Mode* funciona como um *outliner*, dando uma visão dos tópicos do texto como um todo. A transição do *Network Mode* para o *Tree Mode* não é automática, ou seja, é o usuário que monta a estrutura de tópicos, arrastando os nós de uma janela para outra.

Estando o texto planejado e organizado hierarquicamente, o usuário então preenche cada nó utilizando o módulo *Edit Mode*. Basicamente, esse módulo aborda o segundo estágio do processo de composição de um texto, a escrita. Nesse módulo o usuário preenche a estrutura de tópicos feita no *Tree Mode*, produzindo assim o primeiro rascunho do texto. Finalmente, no *Text Mode*, o usuário tem uma visão linear do texto, podendo, por exemplo, corrigir erros de coesão entre parágrafos. Com esse último módulo o sistema aborda a última etapa do processo de composição, isto é, a revisão do texto completo.

Em experimento realizado com o WE (Lansman, Smith, & Weber, 1993), os usuários destacaram como principal ponto positivo o fato dos módulos serem apresentados em janelas separadas e paralelas, de modo que a pessoa pode escrever em uma janela enquanto visualiza o outline do documento em outra. Outra vantagem destacada foi a facilidade de se agrupar os nós para montar o *outline* no *Tree Mode*. Como principal desvantagem foi apontada a falta de recursos dos editores do *Edit Mode* e do *Text Mode*. Os usuários sentiram falta das facilidades de edição oferecidas pelos processadores de texto comerciais. Alguns usuários acharam os módulos *Network* e *Tree* redundantes, mas não souberam indicar qual dos dois seria mais relevante e qual poderia ser removido.

3.2 Writer's Assistant

Assim como o WE, o Writer's Assistant também é baseado em modelos de escrita que consideram os aspectos cognitivos do processo. Dessa forma, este ambiente também é dividido em módulos que se especializam em cada etapa do processo, com o objetivo de diminuir a sobrecarga cognitiva imposta ao escritor.

O Writer's Assistant é um protótipo que combina um organizador de idéias, um editor de estrutura e um editor de texto, chamados de *Notes View*, *Structure View* e *Linear View*, respectivamente. No módulo *Notes View* o usuário pode expressar suas idéias na forma de nós e interligá-los através de arcos que especificam o tipo de relacionamento existente, formando assim uma rede de idéias. Trechos de texto podem ser anexados aos nós da rede já nesse módulo. Além dos nós estarem ligados entre si, alguns pontos da rede também são ligados a outros pontos que correspondem à estrutura geral do texto. Dessa forma, a rede de idéias já está diretamente relacionada à estrutura final do texto.

O módulo *Structure View* apresenta uma visão da estrutura hierárquica da rede de idéias editada anteriormente, dando ao usuário a oportunidade de visualizar melhor a estrutura do texto. Uma visão do texto completo é apresentada no módulo *Linear View*.

A característica mais importante do Writer's Assistant é a facilidade de movimentação entre os três módulos do ambiente, isto é, entre as três visões de um texto. O ambiente utiliza um algoritmo que lineariza a rede de idéias construída pelo usuário criando automaticamente um primeiro rascunho do texto. Para a linearização é respeitada a prioridade dos tipos de ligações estabelecidas entre os nós e o posicionamento de partes da rede a uma ou outra posição do texto por meio da ligação feita com a estrutura (Sharples, Goodlet, & Clutterbuck, 1994).

Em Sharples, Goodlet, & Clutterbuck (1994) é relatado um experimento feito para testar o algoritmo de linearização. O algoritmo foi aplicado a doze redes de idéias de tamanho médio (vinte nós). Os textos linearizados foram avaliados quanto à organização e classificados numa escala de 5 pontos. A média obtida para os textos linearizados automaticamente foi de 3.5, ficando entre "organizado" e "razoavelmente organizado". Esse resultado demonstrou que o uso desse algoritmo é aceitável para a criação de um primeiro rascunho, embora melhorias ainda sejam necessárias.

3.3 Composer

O Composer é um projeto de ambiente de escrita técnica para não-nativos da língua inglesa que começou a ser desenvolvido na Universidade de Brighton, UK, mas que não chegou a ser finalizado. Esse projeto pretendia reunir em um único ambiente a abordagem

de diferentes visões implementada no Writer's Assistant com estratégias mais específicas de apoio à composição de textos técnicos, focalizando em especial as necessidades de escritores não-nativos.

Embora o projeto não tenha sido finalizado, referências encontradas na literatura especializada (Pemberton, Shurville, & Hartley, 1996; Shurville, Hartley, & Pemberton, 1997) descrevem os estudos preliminares, a metodologia de desenvolvimento e parte da funcionalidade esperada, o que nos dá uma boa idéia do que o sistema viria a ser.

O Composer foi projetado para a escrita de tipos específicos de textos – relatórios de experimentos, artigos científicos e propostas de projetos. Para especificar o tipo de ajuda que o sistema daria, foi feita uma pesquisa junto aos estudantes não-nativos daquela universidade e aos seus professores. Essa pesquisa foi baseada em questionários e seu objetivo era descobrir os problemas particulares da escrita numa segunda língua e que poderiam ser minimizados com o suporte de um sistema computacional. Os resultados permitiram que fossem elicitados vários requisitos do sistema e também apontaram várias características que deveriam ser evitadas, tanto relacionadas à funcionalidade quanto ao estilo de interação.

A principal característica apontada como desejável na pesquisa quanto à funcionalidade foi que o sistema fornecesse modelos de textos *on-line*, com uma variedade de formas e de propósitos, para que os usuários pudessem reproduzir a estrutura e o estilo. Os usuários apontaram não só a necessidade de modelos estruturais, mas também a necessidade de instâncias de tais modelos. Embora se tenha demonstrado interesse por modelos pré-estabelecidos, foi apontado que não se deveria impor uma seqüência de passos rígida para o processo. Quanto ao estilo de interação, ficou clara a preferência por sistemas que auxiliam o processo e permitem ao usuário aprender a escrever melhor, ao invés de sistemas de pós-processamento.

Baseado nesses requisitos, o Composer foi projetado para ser um ambiente que trabalhasse as três visões do processo de escrita abordadas pelo Writer's Assistant e que também permitisse acesso *on-line* a esquemas estruturais específicos para cada propósito, diretrizes para a boa escrita relacionadas a cada modelo e textos-exemplo. O modo como essa visão estrutural viria a ser apresentada no Composer é exemplificado em (Pemberton, Shurville, & Hartley, 1996) e reproduzido na Figura 3.1.

A Figura 3.1 representa um esquema estrutural para artigos científicos, mostrando as seções típicas desse tipo de texto e a ordem em que as seções aparecem. Além disso, a figura possui informações associadas em dois eixos, vertical e horizontal, que podem ser percebidas pelo formato dos elementos que representam cada seção. A extensão de cada elemento na horizontal representa o grau de generalidade das informações contidas naquela seção. Por exemplo, a Introdução (*Introduction*) deve começar com termos gerais

e então, progressivamente, ir usando termos mais específicos, como é sugerido pelo formato da figura. No eixo vertical, estão indicadas a ordem dos elementos para esse gênero de texto e a extensão relativa que cada seção pode apresentar. O título de cada seção provê acesso a menus que disponibilizam informações relativas à seção, como: informações sobre o papel da seção dentro do texto, textos bem escritos exemplificando a seção, principais erros cometidos quando se escreve tal seção e também conselhos sobre como refinar o texto. Um exemplo de menu é mostrado para a seção Resultados (*Results*).

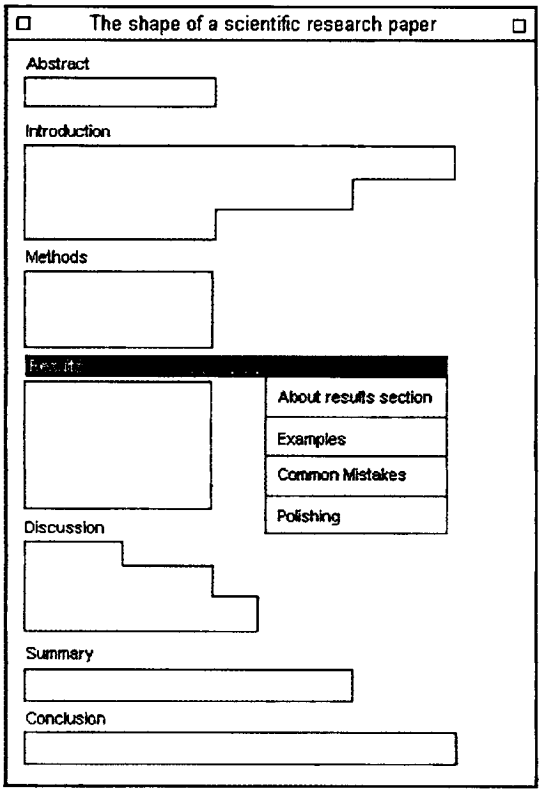


Figura 3.1: Ilustração da visualização de um esquema estrutural no Composer (Pemberton, Shurville, & Hartley, 1996)

3.4 Academic Writer

O Academic Writer é um ambiente de escrita para Web que nasceu do projeto Composer. Seu objetivo é dar suporte a escritores iniciantes e/ou não-nativos em inglês acadêmico. Parte de seus requisitos foi herdada do projeto Composer, porém sua implementação foi simplificada e codificada em páginas Web.

Esse sistema é mais voltado para o ensino do gênero científico do que os sistemas apresentados anteriormente, pois grande parte da sua arquitetura é baseada em hipertextos

explicativos do gênero. O sistema focaliza principalmente três aspectos da escrita científica: (i) a argumentação, encorajando um *brainstorming* estruturado que guia o aluno do tópico para o problema, visando à definição de propósito; (ii) a estruturação, indicando a organização adequada do conteúdo; e (iii) a realização lingüística, descrevendo as características estilísticas e lingüísticas da escrita acadêmica.

O Academic Writer trabalha com sete módulos, chamados de áreas, enfocando diferentes aspectos do processo de escrita: Estratégias, Gênero, Idéias, Foco, Estrutura, Revisão e Língua, sendo que cada área é implementada em uma página. Os recursos oferecidos pelo sistema são classificados como diretrizes, descrições e exemplos, e são organizados de forma a dar suporte ao entendimento de um determinado gênero (p.e. dissertação).

O sistema também mantém um banco de dados onde são armazenadas as idéias de cada usuário (respostas às perguntas feitas pelo sistema) na área *Ideas* de modo que possam ser recuperadas posteriormente. Cada usuário mantém uma senha de acesso para proteger seus dados. A Figura 3.2 apresenta a página que mostra o menu principal do Academic Writer, de onde se pode acessar qualquer uma das sete áreas. Cada área é representada por um ícone seguido de um texto explicativo indicando o tipo de ajuda que o usuário irá encontrar.

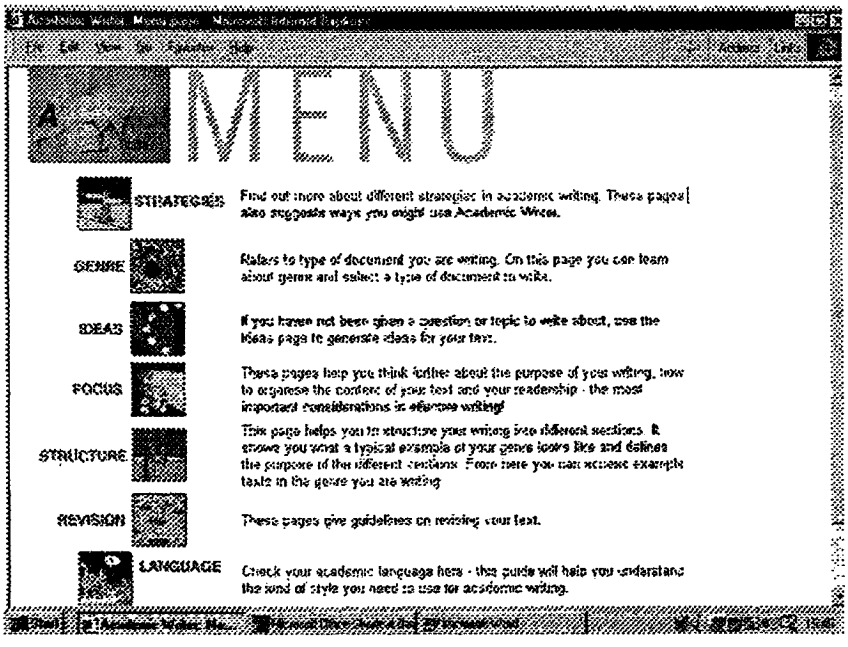


Figura 3.2: Menu principal do Academic Writer (Broady & Shurville, 2000)

Na primeira área, relativa a *Strategies*, o usuário encontra cinco estratégias de escrita que sugerem cinco formas diferentes de se acessar as páginas do Academic Writer. Para cada estratégia, o sistema sugere que as outras áreas sejam acessadas numa determinada ordem, porém essa ordem não é imposta.

Na área *Genre* o usuário seleciona o tipo de gênero e a área nos quais ele está escrevendo. Essa escolha irá determinar as informações que serão mostradas em outras áreas, como por exemplo, na área *Strategies*.

As áreas *Ideas* e *Focus* são independentes de gênero e focalizam a argumentação. Na área *Ideas* o usuário responde a uma lista de perguntas do sistema, chamada de *structured brainstorming*, com o intuito de revelar, por meio de suas respostas, seu tópico de pesquisa e o problema enfocado. A área de *Focus* faz o refinamento das idéias trabalhadas anteriormente, visando focalizar o propósito do trabalho que está sendo escrito e expressá-lo de maneira correta. As informações registradas nessas duas áreas são armazenadas no banco de dados para posterior recuperação.

Na área *Structure* são indicadas as possíveis seções que irão compor o texto do gênero e área escolhidos. Essa escolha deve ter sido feita previamente na área *Genre*. Nessa área, referente à estrutura, o usuário tem acesso a descrições de cada seção, indicando, por exemplo, sua função dentro do texto como um todo. A partir das descrições o usuário pode então acessar exemplos de como cada seção pode ser realizada na prática.

Na área *Language* o usuário encontra um guia de referência, na forma de um hipertexto, para características lingüísticas e estilísticas de textos científicos. Esse guia pode ser navegado de duas formas. Uma delas é por meio de *links* referentes a três máximas da escrita científica ("seja lógico", "seja objetivo" e "seja aberto a dúvidas") e duas características chaves do estilo acadêmico ("use marcadores discursivos" e "escreva de forma clara e simples"). A partir desses *links*, o usuário tem acesso às outras páginas que mostram como esses princípios podem ser realizados lingüisticamente. Outra forma de navegar pela área *Language* é por meio de um índice organizado de A-Z com os pontos relevantes da escrita acadêmica (p.e., o uso de tempos verbais), também ligados a exemplos de textos autênticos.

A área *Revision* tem como objetivo ajudar o usuário a revisar seu texto oferecendo várias *checklists*. Nesse sentido, o sistema provê apenas uma *checklist* originada das respostas dadas nas áreas *Ideas* e *Focus* e das máximas relacionadas ao estilo acadêmico que aparecem na área *Language*.

Numa avaliação inicial, em experimento realizado com quatro estudantes em nível de mestrado na área de *Media-Assisted Language Teaching*, sendo dois nativos e dois não-nativos (um chinês e um espanhol), o Academic Writer se mostrou adequado a seus propósitos, isto é, os recursos oferecidos pelo sistema foram considerados proveitosos pelos avaliadores, entretanto, a avaliação também mostrou que vários aspectos do sistema ainda necessitam de melhorias. Os resultados completos desse experimento são reportados em (Broady & Shurville, 2000).

3.5 Abstract Helper

O Abstract Helper é uma ferramenta para auxiliar a escrita de resumos em inglês (*abstracts*) por escritores não-nativos, em especial os japoneses. Ao contrário das outras ferramentas apresentadas neste capítulo, o Abstract Helper não se preocupa com o processo de geração e organização de idéias, concentrando-se na estruturação e na realização lingüística do resumo. A motivação para construir esse sistema, segundo Narita (2000b), foi tentar solucionar o problema de escrita de textos científicos em inglês enfrentado por escritores japoneses, proveniente não só das diferenças de idioma, mas também da própria organização do texto.

A ferramenta implementa a abordagem de reutilização de textos autênticos como modelos organizacionais e estilísticos para a escrita de textos novos, tendo em vista três objetivos: disponibilizar um córpus paralelo Inglês/Japonês, anotado, de exemplos de resumos bem organizados; permitir ao usuário encontrar um bom modelo para seu resumo ou sua sentença; permitir que o usuário tenha acesso rápido aos recursos lingüísticos para conseguir informação relevante no contexto do seu texto.

Para atender os objetivos, foi coletado um córpus de 539 resumos em inglês, provenientes de publicações relevantes na área da Ciência da Computação e considerados bons exemplos de organização e estilo. Em seguida, foi preparada a tradução para o japonês de cada resumo, visando facilitar o acesso pelos usuários. Os resumo foram traduzidos sentença a sentença e, posteriormente, foram alinhados manualmente.

Esse córpus paralelo foi anotado em SGML utilizando um conjunto de etiquetas-padrão, em dois níveis: resumo como um todo e sentença. No primeiro nível foram marcados o tipo do resumo (p.e. proposta de um novo sistema/modelo/algoritmo) e o esquema organizacional do resumo, que foi classificado em quatro tipos: (i) resumos que iniciam com a sentença tópico; (ii) resumos com a sentença tópico no meio; (iii) resumos que terminam com a sentença tópico; (iv) resumos multiparágrafos. No nível sentencial, foram marcados os papéis de cada sentença, sendo que foram definidos sete papéis: (1) sentença introdutória; (2) sentença tópico; (3) sentença explanatória; (4) sentença de verificação; (5) sentença suplementar; (6) sentença de conclusão; (7) sentença de fechamento. Além dessa anotação em dois níveis, o córpus foi anotado com algumas informações relativas aos verbos utilizados, que foram agrupadas em uma base de dados lexicais, juntamente com as freqüências de ocorrência no córpus. Outro recurso disponível nesse sistema é uma base de colocações em inglês extraídas do córpus e checadas manualmente.

O Abstract Helper pode ser visto como um grande repositório de recursos lingüísticos que podem ser acessados por meio de quatro tipos diferentes de busca: por resumo, por

sentença, por padrão de sentença[1] e por colocação. Não existe uma ordem pré-estabelecida para o acesso aos recursos, porém é esperado que o usuário acesse o sistema num procedimento *top-down*: buscando primeiramente um bom modelo de resumo, em seguida buscando por exemplos de sentenças e, finalmente, buscando informações sintáticas e léxicas.

Narita (2000b) reporta um experimento feito com usuários acadêmicos e o resultado mostrou uma atitude positiva dos avaliadores em relação ao sistema. Entretanto, algumas melhorias foram requisitadas como diversificação do domínio do córpus, maior número de exemplos e produção de uma versão do sistema baseada em janelas.

3.6 Mover

O Mover é uma ferramenta que se propõe a auxiliar a leitura e a escrita de textos científicos apresentando ao usuário a estrutura esquemática do texto. Seu público alvo são estudantes não-nativos e/ou inexperientes em escrita científica. Os autores do sistema partem do princípio de que os modelos de estruturação esquemática (como o CARS de Swales (1990)) constituem uma importante ferramenta no ensino do gênero acadêmico. No entanto, a utilização de tais modelos é limitada pela dificuldade em se identificar os elementos do modelo em um novo texto. Sabe-se que tal tarefa é custosa mesmo para especialistas treinados (Anthony & Lashkia, 2003). É nesse contexto que o Mover atua, auxiliando na identificação dos elementos estruturais do texto, com base num modelo de estruturação pré-definido.

Tal auxílio se dá por meio da identificação automática dos elementos do modelo estrutural, feita por um classificador estatístico. Esse classificador determina, para cada sentença do texto, a categoria à qual a sentença pertence (nesse caso, um elemento do modelo). Dessa forma, a saída do sistema é o texto de entrada com as sentenças etiquetadas de acordo com o modelo estrutural em uso.

O Mover utiliza a abordagem de aprendizado supervisionado, ou seja, o sistema "aprende" as características da estrutura de uma determinada classe de textos com base em um conjunto de exemplos manualmente etiquetados. Essa fase do aprendizado é chamada de *treinamento* e os exemplos utilizados de *conjunto de treinamento*. As etiquetas utilizadas nos textos do conjunto de treinamento determinam o modelo estrutural usado pelo sistema. O conhecimento é extraído do conjunto de treinamento na forma de clusters de palavras. Dessa forma, o Mover estaria apto a funcionar com qualquer tipo de texto e para qualquer modelo estrutural, desde que haja um conjunto de treinamento adequado.

[1] Esse tipo de busca considera, além do papel atribuído à sentença, também características sintáticas e léxicas.

Em Anthony & Lashkia (2003), o uso da ferramenta é demonstrado utilizando-se como conjunto de treinamento *abstracts* de artigos da área de Tecnologia da Informação (TI). Os *abstracts* foram etiquetados manualmente com base em uma versão modificada do modelo CARS de Swales (Anthony, 1999). Detalhes sobre o processo de treinamento do classificador utilizado pelo Mover serão discutidos adiante, no Capítulo 4. A Figura 3.3 mostra a tela principal do sistema com um *abstract* já classificado.

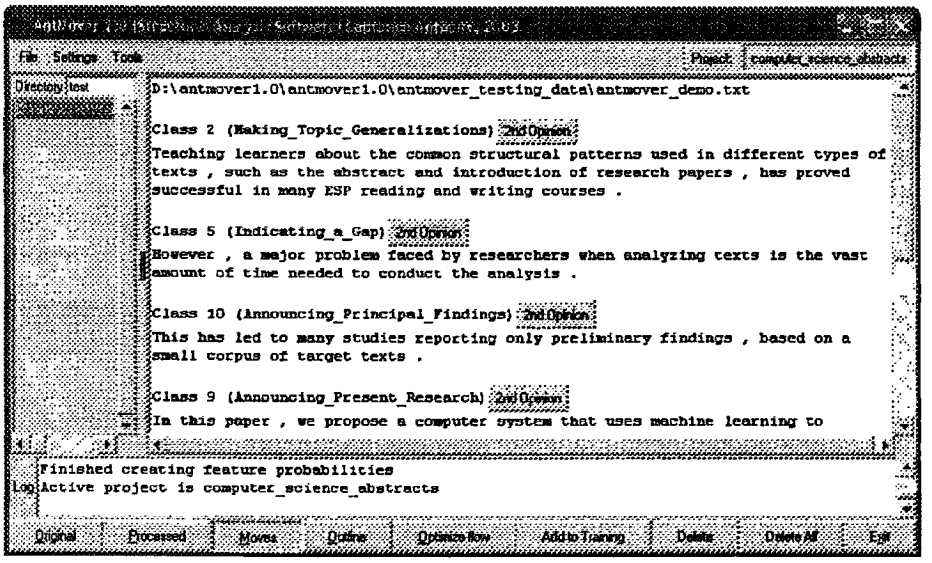


Figura 3.3: Tela principal do Mover (Anthony & Lashkia, 2003)

Para avaliar o Mover em seu contexto de uso, isto é, como ferramenta de auxílio à leitura e escrita, foram feitos dois experimentos com um grupo de seis estudantes de pós-graduação em Ciência da Computação, falantes não-nativos da língua inglesa.

No primeiro experimento, foi pedido aos estudantes que analisassem manualmente 10 *abstracts* de artigos de TI e comentassem sobre sua estrutura retórica em termos do mesmo modelo utilizado pelo sistema (versão simplificada do modelo CARS citada anteriormente). O tempo de duração e os comentários dos estudantes foram gravados. Em uma segunda fase, foi pedido que realizassem a mesma tarefa com o auxílio do Mover. Os resultados mostraram que, com o auxílio do Mover, os estudantes puderam completar a tarefa de forma mais rápida e muito mais eficiente, já que as análises feitas com o auxílio do Mover se aproximavam bastante da análise feita por um leitor experiente, enquanto que, sem o auxílio do sistema, muitos estudantes nem conseguiram completar a tarefa.

No segundo experimento, planejado para medir a utilidade do Mover no processo de escrita, foi pedido aos estudantes que avaliassem manualmente um de seus próprios *abstracts* e fizessem correções à estrutura, se julgassem necessárias. Em seguida, foi pedido que realizassem a mesma tarefa com o auxílio do Mover. Novamente, o tempo de duração foi gravado. Os resultados mostraram que o uso do Mover diminuiu drasticamente o tempo

gasto para a realização da tarefa (de uma média de 90 minutos para 4 minutos). Ainda, verificou-se que os estudantes foram muito mais críticos quando realizaram a avaliação manual, muitas vezes "críticos demais", fazendo correções desnecessárias. Com o auxílio do Mover, essa "crítica excessiva" não foi observada.

3.7 AMADEUS

O AMADEUS[2] é um ambiente de auxílio e ensino da escrita técnica em inglês, voltado para escritores não-nativos que sofrem influência da língua materna quando escrevem em inglês. É composto por quatro ferramentas inter-relacionadas: ferramenta de referência, ferramenta de suporte, ferramenta de crítica e ferramenta tutorial, sendo que as três primeiras já foram implementadas e a quarta está em fase de projeto.

O AMADEUS foi fundamentado nos estudos de alunos brasileiros em pós-graduação no exterior feito por Fontana *et al.* (1993) cujos resultados mostraram que algumas deficiências na escrita dos alunos estão relacionadas ao mau uso ou omissão de expressões mais ou menos convencionais que desempenham funções específicas no texto acadêmico. Uma solução para esse problema, e que foi implementada no AMADEUS, é a reutilização de expressões padrão utilizadas em textos reconhecidos como bons para que, no caso de dúvida no momento da escrita, o escritor tenha uma base de bons exemplos a qual ele possa recorrer.

As três ferramentas atualmente implementadas no ambiente AMADEUS trabalham com essa abordagem, embora cada uma focalize um nível diferente de auxílio. A Tabela 3.1 apresenta o relacionamento entre as características do usuário relativas ao conhecimento de inglês e de escrita técnica e as quatro ferramentas previstas para o ambiente.

Tabela 3.1: Adequação do tipo de ferramenta quanto ao conhecimento do usuário (Barros, 2000)

Escrita Técnica / Inglês	Boa Experiência	Alguma Experiência	Nenhuma Experiência
Bom domínio	*Ferramenta de Referência*		
Problemas de coesão em nível de parágrafo		*Ferramenta de Suporte*	
Problemas na escrita para um propósito e audiência específicos		*Ferramenta de Crítica*	
Problemas de coesão em vários níveis			*Ferramenta Tutorial*

As ferramentas de Referência e Suporte foram desenvolvidas como parte do trabalho de doutorado de Aluísio (1995) e foram implementadas para o domínio da Física Experi-

[2] *Amiable Article Development for User Support*

mental. A ferramenta de Crítica foi desenvolvida durante o trabalho de mestrado de Silva (1999) e trabalha com o domínio específico da comunidade de HCI[3].

As seções seguintes detalham cada uma das três ferramentas do ambiente, indicando contexto de uso, formalismos computacionais utilizados, recursos lingüísticos acessados e outros detalhes de implementação.

3.7.1 Ferramenta de Referência

Como o próprio nome sugere, essa ferramenta tem como objetivo servir de referência para o escritor não-nativo, provendo uma base de expressões padrão que podem ser utilizadas no contexto da escrita científica, nesse caso específico, artigos científicos. Estudos como os descritos no Capítulo 2 desta tese indicam a existência de padrões que podem ser reproduzidos, mesmo no nível de sentenças, de forma que um número relativamente pequeno de padrões pode ser suficiente para cobrir uma grande proporção das expressões usadas em uma área de conhecimento específica. Para tal, foi compilada uma base de expressões autênticas de artigos científicos da física, através da análise de córpus, categorizando essas expressões de acordo com os componentes e subcomponentes da estrutura esquemática de um artigo. As expressões padrão são apresentadas juntamente com lacunas, que devem ser preenchidas com o material factual da pesquisa do escritor.

Note que as expressões apresentadas na Figura 3.4 variam em tamanho, de períodos simples (entrada *a*) e compostos (entrada *d*) a vários períodos (entrada *i*). A idéia subjacente é fornecer, para alguns casos, os conectivos lógicos necessários para a conexão entre períodos. Além de apresentar a realização adequada de certas funções retóricas, as expressões fornecem também substantivos, adjetivos e verbos adequados à descrição de processos, comparação de dados e figuras e outros. Essa característica sugere que a base de expressões seja específica a cada área de pesquisa/conhecimento.

A Figura 3.5 ilustra a arquitetura da ferramenta de Referência, enfatizando dois processos principais, composição e aquisição, e os recursos disponibilizados pela ferramenta. A ferramenta utiliza quatro recursos/bases:

1. a base de expressões padrão, chamada de Base de Padrões;
2. uma lista de componentes e subcomponentes da estrutura esquemática para artigos proposta por Deyes (1982)[4], com ponteiros para a Base de Padrões;
3. uma lista de funções retóricas (p.e. definições, exemplificações, etc) que também apontam para as entradas correspondentes na Base de Padrões;

[3] *Human-Computer Interaction*

[4] Deyes, T. (1982). Discourse, Science and Scientific Discourse (The Raw Material of Comprehension in ESP). Working Paper 6, Brazilian ESP Project, PUC/SP, São Paulo *apud* (Aluísio, 1995).

a) *Importance of the field, general interests, etc.*
There has been substantial interest in the fabrication of ...

b) *Description of an effect, phenomenon, etc.*
The phenomenon of ... induced by ... has not only provided a sensitive and convenient probe for monitoring ... (membrane breakdown) but has also revealedthe irreversible changes that can occur during ...

c) *Previous reports on related work.*
Several papers have reported measurements aimed at obtaining eidence for, and insight into, ... processes in ...

d) *What is lacking in the field.*
Although significant advances have been made in the understandig of how ... (something) influences ... (another), very little further attention appears to have been given to the ...

e) *Difficulties faced in a particular analysis*
Further difficulties arise from the limited ... availableand the requirement for a ...

f) *Improvements on previous works, models, etc.*
This is an improvement over the existing practice of simply inserting (calculating, determining, etc) ...

g) *What the present work does.*
The purpose of the work reported here was to study the influencsof ... on the ...

h) *Relevance of this work to the field or other areas*
The surface properties of ... apart from the pure physical chemical interest will help to elucidate the role of ...in many ... phenomena.

i) *Layout or Outline of the paper*
The organisation (outline) of the (this) paper is as follows. In Section II wedescribe ... The ... is presented in section III. In Sec. II we solve the ... equation giving expressions for ... This is necessary for the work of Sec. III, In which the extended ... equation is derived. Numerical results of the theory are given in Sec. IV, together with a comparison with ... and ... calculations.

Figura 3.4: Exemplos de expressões padrão para introduções (Aluísio, 1995)

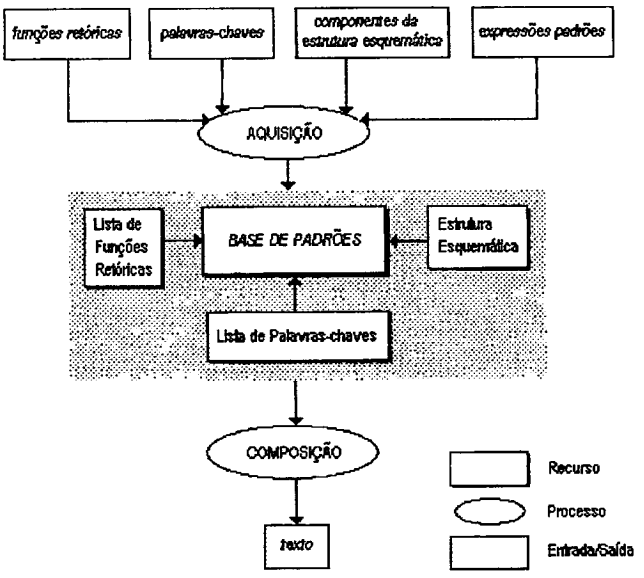


Figura 3.5: Arquitetura da ferramenta de Referência (Aluísio, 1995)

4. uma lista de palavras-chave características do gênero (p.e. *report, paper, model, technique,* etc) que permite acesso rápido às entradas da Base de Padrões por meio de uma busca simples que retornará todas as expressões contendo a palavra-chave

pesquisada.

Sendo assim, durante o processo de composição o usuário pode acessar a Base de Padrões de três formas: a partir da estrutura esquemática, da lista de funções retóricas e da lista de palavras-chave. O processo de aquisição diz respeito à inserção de novo material lingüístico à ferramenta, isto é, inserção de novas expressões padrão, funções retóricas, componentes da estrutura esquemática e palavras-chaves. Assim o usuário pode personalizar a ferramenta, inserindo o material lingüístico que desejar.

Testes realizados com a ferramenta de Referência junto a alunos de pós-graduação mostraram uma avaliação positiva. Algumas características favoráveis apontadas foram a familiarização com a estrutura retórica de artigos científicos e melhorias na organização do texto. A ferramenta também se mostrou útil para superar o problema do bloqueio inicial com relação a escrita em uma língua não materna.

É importante destacar que a ferramenta de Referência é voltada para escritores não-nativos que possuem um bom conhecimento tanto da língua inglesa quanto do gênero científico (conforme apontado na Tabela 3.1), uma vez que a ferramenta trabalha com construções gramaticalmente sofisticadas e não oferece nenhum auxílio direto para a organização no nível de parágrafo ou de seção. Esse tipo de auxílio é oferecido pela ferramenta de Suporte, que é apresentada a seguir.

3.7.2 Ferramenta de Suporte

A ferramenta de Suporte é voltada para usuários que, embora tenham um conhecimento razoável da língua inglesa, não possuem muita experiência em escrita científica, encontrando dificuldades em organizar seu texto de forma adequada ao gênero, produzindo textos sem coesão. Essa ferramenta é destinada à escrita de introduções de artigos curtos (*letters*) da área da Física experimental e utiliza o método de resolução de problemas conhecido como Raciocínio Baseado em Casos (RBC) (Mantaras & Plaza, 1997).

Visando diminuir o problema da falta de coesão dos textos escritos em inglês por não-nativos, Aluísio (1995) propôs uma abordagem baseada em casos para a construção de uma ferramenta para a escrita de introduções que utiliza dois mecanismos principais: instanciação de um esquema detalhado de introdução com material lingüístico de textos autênticos e aplicação de adaptações interativas que flexibilizam as opções do esquema instanciado. O primeiro mecanismo fornece a estrutura retórica do texto real para que o escritor veja como os mecanismos de coesão são expressos na língua. O segundo adapta essa estrutura às necessidades atuais do novo escritor.

A abordagem proposta pode ser resumida em três princípios:

1. pré-compilação do conhecimento do gênero em um esquema detalhado que é utili-

zado para mapear introduções em casos e como fonte de conhecimento na fase de adaptação dos casos;

2. reutilização do material lingüístico não factual dos textos reais, ajudando a aumentar a fluência dos textos e servindo de fonte de conhecimento na fase de adaptação; e
3. utilização de raciocínio baseado em casos, RBC, como modelo. Esse princípio dita as fases da interação e os recursos utilizados, que são:

    - Uma Base de Casos, gerada através da instanciação do esquema detalhado do texto com material lingüístico de textos autênticos, sendo que o índice de cada caso é a sua própria estrutura retórica;
    - Métricas de Similaridade, utilizadas para recuperar os melhores casos dada a estrutura retórica do texto a ser escrito;
    - Regras de Revisão, utilizadas na adaptação interativa de um caso autêntico para outras necessidades.

A ferramenta de Suporte é um protótipo que implementa tal abordagem. Ela foi originalmente implementada em Sicstus Prolog com a interface em Xview ToolKit e Netscape Browser, mas posteriormente foi migrada para o ambiente Windows. A Figura 3.6 apresenta a arquitetura da ferramenta, destacando os processos e recursos utilizados.

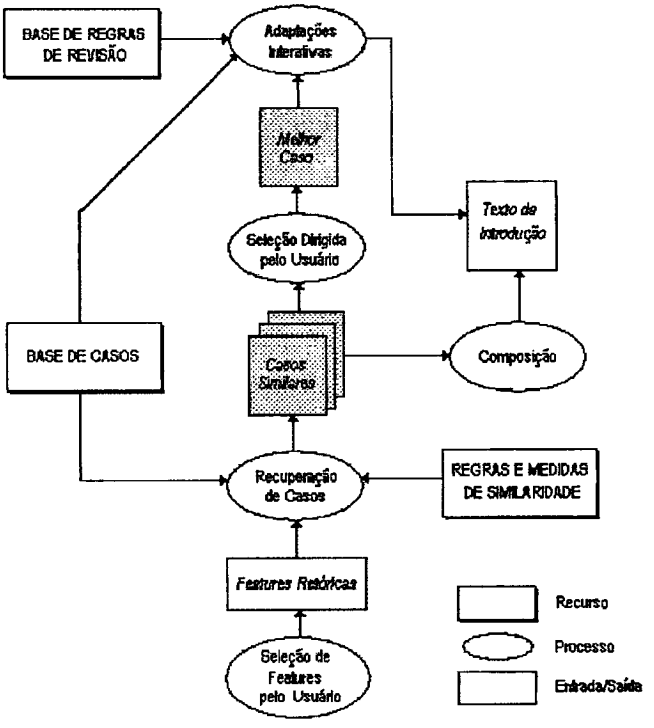


Figura 3.6: Arquitetura da ferramenta de Suporte (Aluísio, 1995)

Como pode ser observado na figura, existem três fontes de conhecimento na ferramenta:

- a **Base de Casos**,
- a **Base de Regras de Revisão** e
- as **Regras e Medidas de Similaridade**.

A **Base de Casos** foi obtida do processo de análise de córpus de introduções de artigos da área da Física Experimental. Essa base contém cinqüenta e quatro instâncias de estruturas retóricas de introduções autênticas, descrevendo os componentes do caso e as estratégias dos componentes. Por sua vez, cada estratégia contém uma lista de mensagens associadas. Cada predicado Prolog (entrada de caso) tem associado um arquivo com o texto da introdução anotado com suas características retóricas (componentes e estratégias). É a partir desse repositório que os casos são recuperados.

A **Base de Regras de Revisão** também foi obtida em parte (estrutura esquemática) da análise de córpus e é composta por regras de adaptações genéricas, pois servem para adaptar qualquer uma das trinta estratégias e suas quarenta e cinco mensagens associadas. São quatro os tipos de adaptações possíveis: inserção de mensagens, remoção de mensagens, mudança de material lingüístico da mensagem e troca de mensagem/estratégia por uma similar.

As **Regras e Medidas de Similaridade** são baseadas em regras de similaridade entre listas (casamento de padrões) e na medida de similaridade conceitual chamada *nearest neighbors matching* (Kriegsman & Barletta, 1993). Essas regras são utilizadas no processo de recuperação dos casos, realizando a busca de acordo com a requisição (estrutura retórica) feita pelo usuário.

Os processos representados na Figura 3.6 mostram como o usuário interage com a ferramenta. Através do processo *Seleção de Features pelo Usuário* o usuário monta a sua requisição, que alimentará o processo *Recuperação de Casos* que é realizado pela ferramenta. Esse processo toma as *Features* Retóricas fornecidas (requisição do usuário), as estruturas retóricas de cada caso da Base de Casos e as Regras e Medidas de Similaridade e devolve ao usuário os casos mais próximos a sua requisição. Tendo os melhores casos recuperados, o usuário escolhe o melhor caso através do processo *Seleção Dirigida pelo Usuário* e assim pode iniciar as *Adaptações Interativas*. Através das adaptações e da entrada de material lingüístico relativo a sua pesquisa o usuário compõe sua introdução. É importante ressaltar que o usuário é livre para escolher se quer fazer as adaptações interativas ou não, podendo escrever a sua introdução somente a partir dos casos recuperados ou do melhor caso escolhido.

Quanto à avaliação da ferramenta de Suporte, ela não foi testada por usuários reais, no entanto, uma limitação evidente é o tamanho reduzido da base de casos. Para que a ferramenta apresente um desempenho próximo do ideal um número maior de casos deve ser incorporado à base.

3.7.3 Ferramenta de Crítica

Um sistema de crítica é composto por dois agentes, um computador e um usuário, trabalhando em colaboração em um processo cíclico. Ambos contribuem com o que conhecem sobre o domínio do problema a ser solucionado. O usuário inicia a interação gerando um produto que será criticado e, depois, faz as mudanças que julgar adequadas, participando juntamente com o sistema do processo de resolução do problema. A tarefa básica do sistema é o reconhecimento e a indicação de deficiências do produto, produzindo uma crítica que permita ao usuário iniciar a próxima iteração deste processo. Com as sugestões dadas pelo sistema, o usuário pode consertar o problema ou procurar obter explicações adicionais. A Figura 3.7 ilustra a abordagem de crítica descrita.

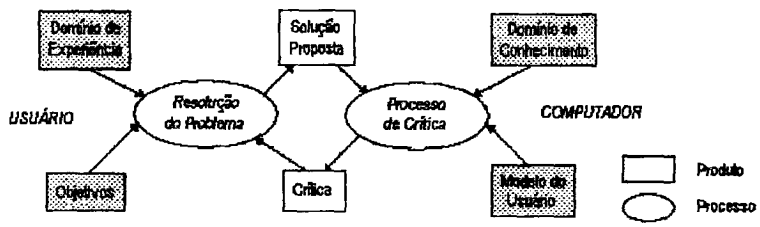


Figura 3.7: A abordagem de críticas (Fischer *et al.*, 1991)

A ferramenta de Crítica incorporada ao ambiente AMADEUS implementa a abordagem descrita focalizada em dois agentes, um usuário humano e um crítico artificial, e foi baseada no modelo proposto por Fischer *et al.* (1991). Essa ferramenta foi implementada para auxiliar na escrita de introduções de artigos da área de HCI e, para tal, foi compilado um córpus de cinqüenta e uma introduções de artigos publicados na CHI'96[5]. Esses textos foram analisados e suas estruturas retóricas foram anotadas, segundo o modelo de Aluísio (Aluísio, 1995; Aluísio & Oliveira Jr., 1996). Também foram anotadas outras informações retóricas particulares dos artigos dessa área. A junção do modelo de Aluísio com as informações específicas foi chamada de Estrutura Dual (Silva, Pelizzoni, & Aluísio, 1998; Silva, 1999). A Figura 3.8 apresenta a arquitetura da ferramenta, destacando os processos (realizados pelo crítico e pelo usuário) e os recursos utilizados.

Como está ilustrado na figura, o processo de crítica inicia-se assim que o usuário apresenta um produto para a ferramenta. Para criticá-lo, a ferramenta deve obter o(s) objetivo(s) do usuário. Isto pode ser feito através de dois processos mutuamente exclusivos: *Reconhecimento de Objetivos*, que reconhece o objetivo do usuário através das escolhas utilizadas no produto, ou *Aquisição de Objetivos* que utiliza informações explícitas fornecidas pelo usuário.

Tendo os objetivos definidos, o modelo utiliza dois tipos de analisadores para avaliar

[5] *Conference on Human Factors in Computing Systems* realizada em 1996.

o produto: *Analítico* e *Diferencial*. O *Analisador Analítico* faz uma análise do produto baseado em um conjunto de *guidelines* (regras heurísticas que ajudam o escritor a adequar seu artigo à comunidade em questão, nesse caso a comunidade da HCI) enquanto que o *Analisador Diferencial* utiliza um caso recuperado da Base de Casos através da abordagem de Raciocínio Baseado em Casos (RBC) como modelo para ser comparado com o produto do usuário. O módulo que faz a recuperação dos casos segue as especificações do trabalho de Aluísio (1995) e recupera o caso mais similar ao produto do usuário. O *Analisador Analítico* é responsável por checar o conteúdo (quais componentes devem estar presentes na introdução) e o *Diferencial* a organização (a ordem mais provável desses componentes, se alguns dos componentes podem ser opcionais, ou aparecer mais de uma vez).

A saída final dos analisadores são *fatos* a respeito das diferenças entre o produto do usuário e o caso recuperado, e da análise do produto em relação às *guidelines*. Esses fatos são reportados ao usuário da forma mais adequada possível, tendo como base as Estratégias de Críticas (p.e. crítica ativa, crítica passiva) e as Formas de Crítica (elogios, críticas diretas, indiretas, sugestões diretas e instruções e sugestões impessoais para melhorias). Com essas informações, o usuário pode então gerar uma nova versão do produto e iniciar novamente o processo.

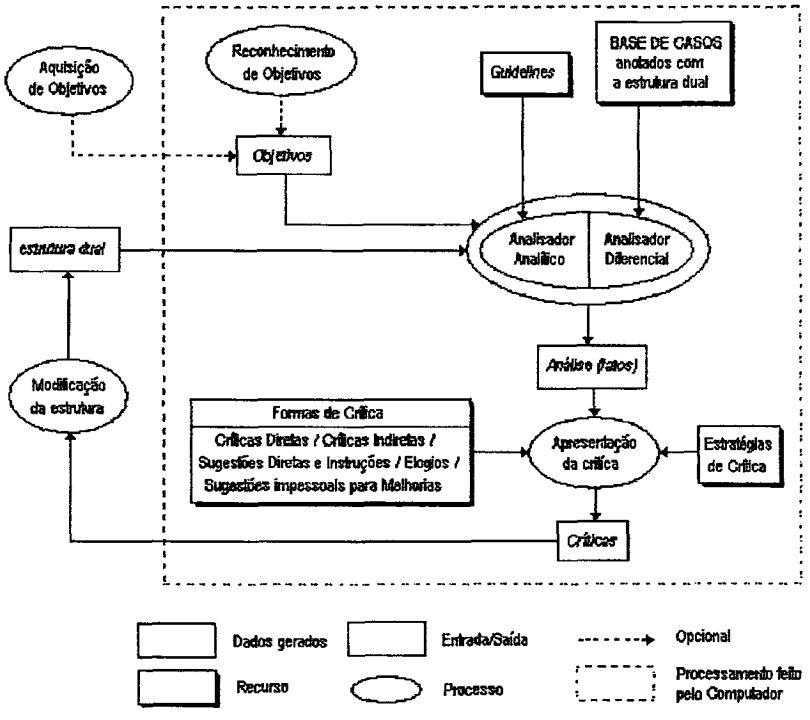


Figura 3.8: A arquitetura da Ferramenta de Crítica (Silva, 1999)

A ferramenta de Crítica foi avaliada em dois aspectos: operacionalidade e praticidade. Para testar a operacionalidade do sistema foram realizadas simulações com três classes

de usuários: principiante, intermediário e especialista. Os usuários das três classes se beneficiaram da ferramenta, pois através da interação com as críticas fornecidas pelo sistema produziram estruturas melhores do que as produzidas inicialmente. Para a análise da praticidade foram observados quatro aspectos: (i) Extensibilidade de *guidelines*, de Casos e Estruturas de novas seções ou de novos tipos de artigos; (ii) Personalização; (iii) Portabilidade para uma nova comunidade de pesquisa; e (iv) Custo de implementação. Como resultado concluiu-se que os três primeiros aspectos são possíveis e podem ser implementados, porém a um alto custo. Detalhes dessa avaliação são encontrados em (Silva, 1999).

3.7.4 Ferramenta Tutorial

A ferramenta tutorial é um módulo que foi idealizado para o ambiente AMADEUS, mas que ainda se encontra em fase de projeto. Ela difere das outras ferramentas do ambiente por ser adequada quando o usuário é completamente novato ou quando existe algum tópico que ele desconheça inteiramente. Nesse caso, é desejável uma interação tutor/aprendiz, pois o sistema tem mais conhecimento que o usuário. Uma tecnologia que poderá ser empregada na ferramenta tutorial é a de agentes inteligentes, de modo que a ferramenta possa gerenciar um agente colaborador e um agente tutor. Essa abordagem colaborador/tutor já foi utilizada por Chan & Baskin (1990)[6].

3.8 Comentários sobre as Ferramentas

As ferramentas descritas neste capítulo foram apresentadas de acordo com a abordagem que utilizam para auxiliar escritores acadêmicos, indo da mais geral para a mais específica, isto é, enquanto as primeiras ferramentas abordam o processo desde a geração de idéias, as últimas focalizam o produto (texto).

As primeiras ferramentas descritas, o WE e o Writer's Assistant, partem do princípio de que, diminuindo-se a sobrecarga cognitiva do escritor durante o processo de escrita, por meio do suporte à organização e hierarquização de idéias, ele se sentirá mais livre para gerenciar todas as restrições embutidas no processo (Sharples & Pemberton, 1992), sendo maiores as chances de produzir bons textos. Essa abordagem focaliza o processo, visando melhorar o produto. No entanto, essas ferramentas não oferecem auxílio para dificuldades como o bloqueio na geração do primeiro rascunho ou a falta de conhecimento da retórica do gênero acadêmico ou mesmo o seu uso inadequado.

[6]Chan, T.W., Baskin, A.B. (1990). Learning Companion Systems. In C. Frasson and G. Gauthier, eds., *Intelligent Tutoring Systems: At the Crossroads of Artificial Intelligence and Education*, Norwood: Ablex *apud* (Silva, 1999)

Preocupações com essas dificuldades foram demonstradas no projeto Composer, que era direcionado para o auxílio de escritores não-nativos que tinham que escrever em inglês. Uma abordagem adotada para resolver ambas as dificuldades foi a utilização de textos autênticos como modelos de boa escrita, de modo que o usuário pudesse se familiarizar com a retórica observando sua realização e, ao mesmo tempo, quebrar o bloqueio inicial da escrita reutilizando padrões observados nos textos-exemplo. O Academic Writer, ferramenta derivada do projeto Composer, implementa essa abordagem e também utiliza outro recurso para auxiliar na geração das idéias iniciais: um *brainstorming* dirigido, isto é, uma seção de perguntas que tenta fazer com que o usuário identifique as idéias principais que devem ser esclarecidas em seu texto (p.e., qual é sua "*research question*"). Uma vez que essa ferramenta é dirigida para usuários que escrevem em uma segunda língua, também oferece apoio lingüístico, como padrões de escrita (*collocations*) e listas de falsos cognatos.

Outras ferramentas voltadas para escritores não-nativos focaram diretamente o produto e assim se aprofundaram nos aspectos mais superficiais do texto, como a estruturação esquemática e suas estratégias de realização. Esse é o caso do Abstract Helper e do ambiente AMADEUS. Essas ferramentas também utilizam a abordagem de reutilização de exemplos autênticos, como o Academic Writer, porém oferecem mais detalhes quanto à função retórica dos elementos. O AMADEUS, em especial, oferece ao usuário um conhecimento bastante detalhado, uma vez que, além de determinar componentes estruturais, identifica também um grupo de estratégias retóricas para cada componente que pode ser empregado para a sua realização.

O Mover, também voltado para estudantes não-nativos, foca apenas a estrutura esquemática do texto, fornecendo ao estudante um *feedback* imediato sobre a estrutura. Embora tenha se mostrado uma ferramenta bastante útil na leitura de textos acadêmicos, seu auxílio durante o processo de escrita é limitado.

É importante destacar que à medida que o conhecimento utilizado pela ferramenta vai sendo mais aprofundado, um recorte no seu domínio de aplicação também é feito. Enquanto ferramentas que abordam aspectos mais gerais, como o WE, o Writer's Assistant e o Academic Writer, abrangem o texto como um todo, ferramentas que oferecem uma ajuda mais especializada e direcionada à estrutura esquemática, como o Abstract Helper, o Mover e o AMADEUS, são específicas para uma seção ou outra do texto (o Abstract Helper e o Mover abordam o Resumo e o AMADEUS a Introdução).

3.9 Considerações Finais

Neste capítulo foi apresentada uma revisão das ferramentas de auxílio à escrita encontradas na literatura que se ocupam de duas fases do processo de escrita: planejamento e escrita. Uma das ferramentas apresentadas também auxilia à leitura de textos acadêmicos (Seção 3.6). As ferramentas que tratam da fase de revisão, também chamadas de ferramentas de pós-processamento (p.e., corretores ortográficos, gramaticais e estilísticos), não foram abordadas por não serem o foco deste trabalho. Outra classe de ferramentas que não foi abordada nessa revisão bibliográfica são os hipertextos explicativos, que funcionam como sistemas de referência e auto-instrução, devido a sua funcionalidade óbvia.

Essa revisão bibliográfica teve por objetivo verificar quais os problemas de escrita acadêmica atualmente tratados por ferramentas computacionais e quais as abordagens utilizadas por essas ferramentas para solucionar esses problemas. Os principais problemas atacados pelas ferramentas estudadas podem ser resumidos em quatro classes, a saber:

a. Sobrecarga cognitiva no momento da geração e organização das idéias;
b. Bloqueio para escrita do primeiro rascunho;
c. Falta de conhecimento da retórica do gênero acadêmico ou uso inadequado da mesma;
d. Falta de familiaridade com padrões de escrita utilizados para o gênero acadêmico.

Na tentativa de minimizar esses tipo de problemas, as seguintes abordagens são adotadas por essas ferramentas:

1. Apoio à organização das idéias através do uso de redes de nós e elos e de *outliners*;
2. Uso de brainstorm dirigido;
3. Uso de hipertextos explicativos do gênero acadêmico;
4. Uso de textos autênticos como exemplos de boa escrita;
5. Uso de possíveis modelos de estruturas esquemáticas;
6. Identificação automática da estrutura esquemática;
7. Crítica da estrutura esquemática;
8. Indicação de estratégias retóricas para a realização de componentes da estrutura esquemática;
9. Uso de bases de padrões de escrita (p.e. *collocations*, texto reutilizável).

A Tabela 3.2 relaciona esse conjunto de abordagens com as ferramenta descritas neste capítulo.

Uma vez que era necessário delimitar os problemas a serem tratados e considerando o foco deste trabalho, o auxílio à escrita de Resumos e Introduções de teses e dissertações,

Tabela 3.2: Resumo das ferramentas de auxílio à escrita investigadas

	WE	Writer's Assistant	Academic Writer	Abstract Helper	Mover	AMADEUS
Apoio à organização de idéias	√	√	√			
Brainstorm dirigido			√			
Hipertextos explicativos			√			
Base de exemplos autênticos			√	√		√
Modelos de estrutura esquemática				√		√
Identificação automática da estrutura					√	
Crítica da estrutura						√
Indicação de estratégias retóricas						√
Base de padrões de escrita			√	√		√

decidimos tratar as classes de problemas b, c e d. Tais problemas interferem na escrita de textos acadêmicos mesmo por escritores nativos, especialmente quando inexperientes. Tal observação pode ser comprovada pelos resultados da análise dos resumos do CorpusDT, apresentada na Subseção 5.2.3.

A metodologia do ambiente AMADEUS foi utilizada como ponto de partida para a exploração das abordagens 4, 5, 7, 8 e 9. A abordagem 6, identificação automática da estrutura esquemática, também foi adotada, mas de uma forma diferenciada da utilizada pelo Mover, já que nossa proposta é empregá-la como parte de um processo de crítica. Essa nova abordagem será esclarecida no Capítulo 6.

No próximo capítulo é apresentada uma revisão dos sistemas de detecção automática da estrutura esquemática de textos acadêmicos disponíveis na literatura, incluindo o Mover, já discutido neste capítulo. Todos esses sistemas utilizam técnicas supervisionadas de aprendizado de máquina e foram desenvolvidos para atender a diferentes propósitos. Essa revisão permitiu que avaliássemos qual seria a abordagem mais adequada aos nossos objetivos, considerando o tipo de modelo esquemático adotado por cada sistema e a relação custo-benefício entre o esforço de implementação e a precisão alcançada.

Capítulo

4

Detecção Automática da Estrutura Esquemática de Textos Acadêmicos

Nesta tese, já foi discutida a natureza padronizada dos textos acadêmicos e como vários autores procuram descrever os padrões encontrados e considerados bons, seja por meio de modelos que emergiram de análises (descrição) ou de diretrizes (prescrição) (Capítulo 2). Entretanto, conforme já comentado no Capítulo 3, interpretar tais modelos e identificar seus elementos em textos reais continua sendo um problema. De fato, problemas relacionados à construção de modelos, que sejam adequados e que possam ser reproduzidos de maneira consistente, têm sido discutidos por vários autores (Anthony, 1999; Teufel, Carletta, & Moens, 1999; Mizuta & Collier, 2004).

Outro aspecto dessa questão, que tem sido discutido na literatura, é a automatização do processo de identificação de elementos estruturais em textos reais. Uma vez que os textos acadêmicos podem ser modelados em padrões que podem ser identificados por agentes humanos (embora com certa dificuldade), esses padrões também poderiam ser "aprendidos" e reproduzidos por agentes artificiais. Esses agentes artificiais, por sua vez, poderiam atuar isoladamente, auxiliando tarefas como a leitura e interpretação de textos, anotação de córpus, entre outras, ou ainda como parte de um sistema mais complexo e com finalidade específica, como é o caso do ambiente proposto nesta tese.

Neste capítulo, são apresentados três sistemas que utilizam a detecção automática de elementos da estrutura esquemática de textos acadêmicos por meio de métodos de Inteligência Artificial: o Critique (Burstein, Marcu, & Knight, 2003; Burstein, Chodorow,

& Leacock, 2003), o Mover (Anthony & Lashkia, 2003) e o Argumentative Zoning (Teufel & Moens, 2002). Cada um dos sistemas foi projetado com objetivos e para domínios específicos, mas todos fazem uma classificação automática tal que, dada uma sentença de um texto, é feito o mapeamento dessa sentença em um elemento da estrutura esquemática. Tais elementos serão chamados aqui de categorias retóricas.

A seguir, apresentamos a descrição das principais características dos três classificadores, de modo a esclarecer os aspectos envolvidos no processo de aprendizado, e também resultados das avaliações intrínsecas de cada classificador.

4.1 Critique

O sistema de análise textual Critique é parte de um projeto maior chamado *Criterion* (Burstein, Chodorow, & Leacock, 2003), que tem por objetivo o desenvolvimento de um sistema educacional comercial que forneça uma pontuação (*scoring*) e também avaliação (*feedback*) automática de textos dissertativos acadêmicos (*essays*). O projeto *Criterion* vem sendo desenvolvido ao longo de vários anos pelo ETS (*Educational Testing Service*), USA, juntamente com pesquisadores de várias universidades norte-americanas, e conta com um investimento de mais de um milhão de dólares.

Para avaliar e pontuar o texto, o Critique faz análises em vários níveis, detectando erros de gramática e uso, estilos indesejáveis e também considerando informações sobre a estrutura esquemática do texto. Esse último aspecto será discutido nesta seção.

O modelo de estrutura esquemática utilizado pelo Critique considera que um texto dissertativo bem escrito deve conter certos elementos, que incluem: (i) material introdutório (*introductory material*), (ii) uma declaração da tese (*thesis statement*), (iii) as idéias principais (*main ideas*), (iv) idéias de suporte (*supporting ideas*) e (v) uma conclusão (*conclusion*). Tal modelo reflete bem a estrutura de textos "persuasivos", isto é, em que o escritor tenta convencer o leitor de que sua tese ou posição (com relação a um determinado tópico) é válida. Esse modelo gerou o esquema de anotação apresentado na Figura 4.1, que reflete as categorias retóricas (i.e., elementos da estrutura esquemática) procuradas nos textos[1].

O componente que faz a análise da estrutura esquemática, detalhado em Burstein, Marcu, & Knight (2003), utiliza um algoritmo de votação baseado em árvore de decisão, que considera as saídas de três sistemas independentes de análise estrutural, chamados de analisadores discursivos. Dois desses analisadores usam métodos probabilísticos e o outro segue a abordagem simbólica. A anotação do córpus de treinamento foi feita manualmente

[1] A categoria *Other* inclui títulos e segmentos considerados "irrelevantes", isto é, que não se encaixam em nenhuma das categorias consideradas.

Introductory material
Thesis
Main idea
Supporting idea
Conclusion
Other

Figura 4.1: Esquema de anotação utilizado no sistema Critique

com base no modelo estrutural (Figura 4.1) e, para tal, dois anotadores humanos foram treinados. O córpus de treinamento resultante contém 1.462 textos sobre seis tópicos (assuntos) diferentes.

Para medir a reprodutibilidade do modelo utilizado na anotação, isto é, quão semelhante será o resultado da anotação feita por diferentes anotadores usando o mesmo conjunto de categorias, foram realizados experimentos com os dois anotadores humanos, utilizando-se a estatística *Kappa* (Siegel & Castellan, 1988) para medir a concordância entre k anotadores sobre N itens. Nesse caso, itens são sentenças e o número de categorias é igual a seis (Figura 4.1). A fórmula para o cálculo do *Kappa* é:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

onde P(A) estima a concordância para cada par de juízes e P(E) estima a concordância aleatória entre eles. O valor do *Kappa* varia entre -1 e 1, sendo que -1 indica máxima discordância, 0 indica que houve apenas concordância ao acaso e 1 indica concordância perfeita entre os juízes. Os resultados demonstraram um alto índice de concordância entre eles, com *Kappa* superior a 0.8.

4.1.1 O Analisador Simbólico

O analisador simbólico utiliza o C5.0, um algoritmo de aprendizado de máquina baseado em árvore de decisão. Os vetores de atributos utilizados pelo algoritmo são determinados automaticamente para todas as sentenças do conjunto de treinamento. O conjunto de atributos é composto por elementos que enfocam relações retóricas com base na RST (*Rhetorical Structure Theory*) (Mann & Thompson, 1988), marcadores discursivos (incluindo termos e estruturas sintáticas que funcionam como marcadores), itens lexicais (categorizados como gerais ou específicos a certas categorias), elementos da estrutura sintática (orações subordinantes, orações subordinadas substantivas, orações subordinadas reduzidas de infinitivo, orações relativas e verbos auxiliares), posição da sentença e do parágrafo (e da sentença dentro do parágrafo) e sinais de pontuação de fim de sentença.

A análise RST dos textos é feita automaticamente por meio do parser discursivo de

Marcu (2000). A partir da saída do parser (árvore retórica), dois atributos são extraídos: (i) o *status* da sentença, isto é, se ela é **núcleo** ou **satélite** da relação (em relação ao seu nó pai) e (ii) qual é a relação retórica (segundo a RST) existente.

A análise dos marcadores discursivos é feita por meio de um léxico contendo palavras, termos e certas construções sintáticas que podem funcionar como marcadores discursivos. Esse léxico também contém informações sobre o contexto de ocorrência em que cada item (palavra, termo, estrutura sintática) pode ser considerado um marcador. A análise de itens lexicais também é feita com base em um léxico que contém itens classificados como sendo indicadores de determinadas categorias.

As características sintáticas, de posição e de pontuação também são detectadas automaticamente, porém não são fornecidos detalhes sobre a implementação.

4.1.2 Os Analisadores Probabilísticos

Os dois analisadores discursivos probabilísticos foram modelados em um *framework Noisy-Channel*. Tal *framework* é composto por um modelo lingüístico e por um modelo *noisy-channel*, que será chamado aqui de "canal ruidoso", ou simplesmente "canal". O modelo lingüístico estima as probabilidades de seqüências de etiquetas retóricas, enquanto que o canal ruidoso estima as probabilidades de seqüências de etiquetas retóricas dada uma seqüência de sentenças. Dessa forma, dada uma seqüência W de sentenças não-etiquetadas (*essay*), a saída do *framework* será uma seqüência de etiquetas retóricas E que maximiza o produto das probabilidades do modelo lingüístico ($P(E)$) e do canal ruidoso ($P(W \mid E)$):

$$E = \text{argmax}_E P(E)P(W \mid E)$$

Tanto o modelo lingüístico quanto o canal são representados como máquinas de estados-finitos probabilísticas (MEFP), ou seja, MEFs em que cada transição possui uma probabilidade associada. Para estimar as probabilidades das transições, as MEFPs são treinadas usando-se técnicas simples de máxima verossimilhança (*maximum-likelihood*) (Manning & Schütze, 2001).

O canal ruidoso é um transdutor (*transducer*) probabilístico de oito estados, que gera pares do tipo <*seqüência de etiquetas retóricas, seqüência de sentenças*>.

O modelo lingüístico tem como objetivo determinar a seqüência mais provável de etiquetas retóricas. Esse modelo considera apenas dependências locais, isto é, aproxima a probabilidade da seqüência considerando que cada nova etiqueta depende apenas das duas etiquetas anteriores (trigramas). Os trigramas e suas probabilidades ($P(e_3 \mid e_1 e_2)$) são extraídos diretamente a partir do córpus de treinamento. Esse modelo baseado em

trigramas também é representado como uma MEFP.

Sendo assim, dois analisadores discursivos probabilísticos diferentes foram treinados a partir do *framework* proposto: (i) um utilizando apenas o canal ruidoso e (ii) outro utilizando o canal e o modelo lingüístico. Em ambos os casos, o objetivo foi buscar na MEFP (ou na composição de MEFPs, no caso (ii)) o caminho de maximiza a probabilidade da saída.

4.1.3 Avaliação Intrínseca

Na avaliação dos analisadores discursivos foi considerado cada analisador individualmente e também os três analisadores em conjunto, como parte do sistema de votação. Ainda, foi utilizado como *baseline* um analisador "posicional", que implementa um algoritmo que considera a posição das sentenças no *essay*, da seguinte forma:

> Classifique a 1ª sentença do texto como *Introductory material.*
>
> Classifique todas as sentenças do 1º parágrafo, exceto a 1ª, como *Thesis statement.*
>
> Classifique a 1ª sentença de todos os parágrafos do corpo do texto como *Main ideas.*
>
> Classifique todas as sentenças de todos os parágrafos do corpo do texto, exceto a 1ª, como *Supporting ideas.*
>
> Classifique todas as sentenças do último parágrafo como *Conclusion.*
>
> Classifique todo texto de cabeçalho ou rodapé, ou sentenças sem pontuação final como *Other.*

Para as avaliações dos analisadores individuais, os resultados da classificação foram computados aplicando-se 10-*fold cross-validation* ao córpus de 1.462 *essays*, ou seja, usando-se nove-décimos do córpus para treinamento e um-décimo para teste, em 10 rodadas, redistribuindo os dados aleatoriamente em cada rodada. O sistema de votação também foi avaliado usando-se 10-*fold cross-validation* sobre a saída dos analisadores individuais. A Tabela 4.1 apresenta os valores de *Precision*, *Recall* e *F-Measure* para cada sistema individual. *Precision* é a proporção de instâncias corretamente reconhecidas, dado o total de instâncias reconhecidas e *Recall* é a proporção de instâncias corretamente reconhecidas, dado o total de instâncias da amostra. A medida *F-Measure* é uma uma combinação das medidas *Precision* e *Recall*, sendo forma conveniente de expressá-las como um único valor. A Tabela 4.2 apresenta os valores dessas medidas comparando-se a *baseline*, o melhor sistema individual e o sistema de votação.

Conforme pode ser observado na Tabela 4.1, os três analisadores discursivos apresentam desempenho melhor que a *baseline*. O analisador individual de melhor desempenho foi o simbólico(*F-Measure* =0.81), porém seus resultados não foram tão bons quanto os

Tabela 4.1: Performance de cada sistemas individual (*Precision*, *Recall* e *F-Measure*) (Burstein, Marcu, & Knight, 2003)

Baseline			Simbólico			Probabilísticos					
						Canal			Canal + Trigrama		
P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
0.71	0.70	0.70	0.81	0.81	0.81	0.69	0.70	0.69	0.78	0.80	0.79

Tabela 4.2: Performance da *baseline*, do melhor sistema individual e do sistema de votação (*Precision*, *Recall* e *F-Measure*) (Burstein, Marcu, & Knight, 2003)

Baseline			Melhor sistema individual			Sistema de votação		
P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
0.71	0.70	0.70	0.81	0.81	0.81	0.85	0.85	0.85

do sistema de votação, que obteve o melhor desempenho entre todos os sistemas avaliados (*F-Measure* =0.85).

4.2 Mover

O Mover (Anthony & Lashkia, 2003), como ferramenta de auxílio à escrita, já foi comentado no capítulo anterior (Seção 3.6). Nesta seção, o foco é a abordagem utilizada pela ferramenta para a detecção automática da estrutura esquemática de textos acadêmicos.

Conforme comentado anteriormente, o modelo de estrutura esquemática utilizado pelo Mover é baseado no modelo CARS de Swales (1990), modificado para refletir a estrutura de *abstracts* de artigos científicos da área de Tecnologia da Informação. Tal modelo originou o esquema utilizado na anotação do córpus de treinamento do Mover. Esse esquema reflete as categorias de sentenças presentes no córpus analisado e é apresentado na Figura 4.2.

Claim centrality
Generalize topics
Indicate a gap
Announce research
Announce findings
Evaluate research

Figura 4.2: Esquema de anotação utilizado pelo Mover

Para detectar a estrutura retórica, o Mover utiliza um método estatístico, implementando um classificador Naive Bayes simples (Lewis, 1998). Experimentos com outros algoritmos de aprendizado de máquina mais complexos, como Árvores de Decisão e Redes Neurais, foram realizados (Anthony & Lashkia, 2003), porém, o classificador bayesiano apresentou melhor desempenho para a tarefa especificada.

Como atributos de classificação, foi implementado um modelo do tipo *bag of words*, porém aplicado a clusters de palavras (expressões), em vez de palavras isoladas. O modelo adaptado foi chamado de *bag of clusters*. Dessa forma, cada sentença do conjunto de treinamento é divida em clusters de tamanho que variam de uma a cinco palavras e, em seguida, cada cluster é mapeado na categoria da sentença a qual ele pertence. Entretanto, tal procedimento acaba inserindo, no modelo, clusters de pouca utilidade para a classificação, que constituem "ruídos". Para remover os ruídos, foi utilizada uma medida estatística chamada *Information Gain (IG)* (Manning & Schütze, 2001), que mede a importância relativa de cada cluster na distinção das diferentes categorias presentes no conjunto de treinamento. Dessa forma, ranqueando os clusters de acordo com sua pontuação *IG*, é possível estabelecer um limiar de corte, que, em princípio, elimina o ruído.

Além do *bag of clusters*, foram consideradas a posição das sentenças dentro do *abstract* e as probabilidades de determinados agrupamentos de categorias (p.e., a probabilidade de sentenças classificadas como "*Claim centrality*" serem seguidas de sentenças classificadas como "*Generalize topics*").

4.2.1 Avaliação Intrínseca

Para a avaliação do classificador, os resultados foram computados aplicando-se 5-*fold cross-validation* a um córpus de 100 *abstracts* (692 sentenças), ou seja, usando-se quatro-quintos do córpus para treinamento e um-quinto para teste, em 5 rodadas, redistribuindo os dados aleatoriamente em cada rodada. Como *baselines*, foram utilizados métodos simples de classificação: (i) atribuição aleatória de categorias e (ii) atribuição da categoria mais freqüente a todas as sentenças.

A precisão média obtida pelo classificador bayesiano foi de 68%. Esse resultado é bem melhor do que os resultados das duas *baselines* (16.6% e 26%, respectivamente). Na Tabela 4.3, é apresentada a matriz de confusão demonstrando os resultados de uma das rodadas de *cross-validation*. Essa matriz permite observar os tipos de erros cometidos pelo classificador, comparando os resultados da anotação humana (representados nas colunas da matriz) com os resultados do classificador (representados nas linhas da matriz). Na matriz de confusão também é apresentada a precisão do classificador por categoria.

Uma característica do classificador Naive Bayes, que é utilizada pelo Mover, é a obtenção de um ranque de categorias para cada sentença, de acordo com as probabilidades calculadas. O Mover utiliza esse ranque de probabilidades para oferecer ao usuário opções de correção da classificação. A utilização da segunda opção do classificador na correção das classificações incorretas aumenta a precisão geral para 86%.

Tabela 4.3: Matriz de Confusão: anotação manual vs. anotação Mover (Anthony & Lashkia, 2003)

		Humano					
		Claim centrality	*Generalize topic*	*Indicate a gap*	*Announce research*	*Announce findings*	*Evaluate research*
Mover	*Claim centrality*	2 (28%)	4	0	0	1	0
	Generalize topic	0	17 (82%)	0	0	3	1
	Indicate a gap	0	2	1 (17%)	0	2	1
	Announce research	0	0	0	34 (92%)	3	0
	Announce findings	0	2	0	2	25 (66%)	9
	Evaluate research	0	2	0	2	8	16 (57%)

4.3 Argumentative Zoning (AZ)

Argumentative Zoning é uma técnica proposta por Teufel & Moens (2002) para segmentar automaticamente um artigo científico em "zonas argumentativas", isto é, partes do texto que possuem certos papéis retóricos dentro do texto como um todo. O objetivo principal dessa segmentação é determinar a atribuição intelectual (autoria), ou seja, determinar "quem fez o que". Segundo Teufel & Moens (2000), são três os possíveis agentes a quem tal atribuição pode ser feita:

- à área de pesquisa em geral.
- a outro pesquisador;
- ao próprio autor;

A motivação para o desenvolvimento do *Argumentative Zoning* foi usar esse tipo de informação retórica na sumarização automática de artigos científicos. Dessa forma, seria possível construir um sumário que destacasse não só a contribuição proposta pelo artigo sumarizado, mas também seu posicionamento com relação aos trabalhos citados.

A detecção automática das zonas argumentativas é realizada por meio de classificador estatístico – chamado AZ – implementado como um classificador Naive Bayes, como em Kupiec, Pedersen, & Chen (1995). Para treinar o classificador, foi criado um córpus manualmente anotado segundo um esquema pré-definido de categorias, contendo 80 artigos publicados em conferências da área de Lingüística Computacional. O esquema de anotação, que modela as sentenças de um artigo científico em sete categorias retóricas, é apresentado na Figura 4.3. Vale notar que as três categorias iniciais (*Background, Other*

e *Own*) modelam exatamente os três possíveis agentes citados anteriormente.

Background
Other
Own
Aim
Contrast
Basis
Textual

Figura 4.3: Esquema de anotação utilizado pelo AZ

Para medir a reprodutibilidade do esquema de anotação, foram realizados experimentos com três anotadores humanos treinados, utilizando um conjunto de 22 artigos provenientes do córpus de 80 artigos. A concordância entre os anotadores, medida por meio da estatística *Kappa* (K), foi considerada boa, dada a natureza subjetiva da tarefa (K=0.71).

O *Argumentative Zoning* assume que o papel retórico de uma sentença pode ser "lido" da superfície do texto. Dessa forma, para a determinação dos atributos de classificação, foram utilizadas heurísticas que medem o quanto é apropriado classificar uma sentença em determinada categoria (zona argumentativa). Tais heurísticas podem ser resumidas da seguinte forma:

- As zonas argumentativas tendem a aparecer em posições típicas num artigo;
- Características lingüísticas como tempo verbal e voz têm relação com as zonas;
- As zonas tendem a seguir outras zonas particulares;
- Inícios de zonas são lingüisticamente marcados por meta-marcadores como *"Other researchers claim that"*;
- Afirmações sem atribuição intelectual formal são interpretadas como tendo a mesma atribuição das sentenças prévias, dentro do mesmo limite de atribuição.

Com base nessas heurísticas, foram definidos atributos que podem ser determinados automaticamente a partir do texto de entrada. Em suma, são 16 atributos que mapeiam aspectos estruturais, sintáticos[2] e lexicais. Uma breve descrição dos atributos é apresentada na Tabela 4.4.

A determinação dos atributos de cada sentença é feita por meio de um *pipeline* de operações que envolvem os seguintes recursos previamente construídos: (i) um conjunto de *templates* de títulos/subtítulos; (ii) uma gramática utilizada para fazer o *parsing* das

[2]Os atributos sintáticos (*Voice, Tense* e *Modal*) são estimados com base nas características do primeiro verbo *finito* da sentença. Diz-se *finita* qualquer forma verbal dos modos indicativo, imperativo ou subjuntivo.

Tabela 4.4: Atributos utilizados pelo AZ (Teufel & Moens, 2002)

Nome	Descrição
Location	Posição da sentença com relação a 10 segmentos pré-definidos que dividem o texto em regiões
Section structure	Posição da sentença dentro de uma seção
Paragraph structure	Posição da sentença dentro do parágrafo
Headline	Tipo do título da seção (mapeado em uma lista de 15 tipos)
Length	A sentença (em palavras) é longa ou curta?
Title	A sentença contém palavras que também ocorrem no título ou subtítulos?
*TF*IDF*	A sentença contém "termos significativos" de acordo com a medida TF*IDF[3]?
Voice	Voz do primeiro verbo finito da sentença
Tense	Tempo do primeiro verbo finito da sentença
Modal	O primeiro verbo finito da sentença é modificado por um auxiliar modal?
Citation	A sentença contém uma citação? Se sim, é para *self-citation*? Em que posição da sentença a citação ocorre?
History	Categoria mais provável da sentença anterior
Formulaic	Tipo de expressão padrão que a sentença contém
Agent	Tipo de agente que a sentença contém
SegAgent	Tipo de agente contido da sentença anterior
Action	Tipo de ação que a sentença contém, com ou sem negação

referências bibliográficas; (iii) um conjunto de *templates* de expressões padrão (*formulaic patterns*); (iv) um léxico de ações; e (v) um léxico de agentes.

4.3.1 Avaliação Intrínseca

Para avaliar o desempenho do AZ, os resultados da classificação foram computados aplicando-se 10-*fold cross-validation* ao córpus de 80 artigos, utilizando o classificador bayesiano treinado com todo o conjunto de atributos. Na avaliação dos resultados, foram consideradas três *baselines*. Como *Baseline* 1, foi usada a saída do classificador textual *Rainbow* (McCallum, 1996), que classifica as sentenças de acordo com a medida TF*IDF. Como *Baseline* 2, foi utilizada a escolha aleatória de categorias, porém baseada na distribuição do córpus, de forma que categorias mais freqüentes têm maior probabilidade de serem escolhidas do que categorias menos freqüentes. Como *Baseline* 3, foi utilizada a atribuição da categoria mais freqüente a todas as sentenças. A Tabela 4.5 apresenta os resultados do AZ em comparação com as *baselines*, em termos do *Kappa* entre o sistema e um anotador humano, acurácia[4] e *Macro-F*[5]. Conforme pode ser observado, o AZ apresenta um desempenho melhor do que as três *baselines* utilizadas.

[3] *Term Frequency * Inverse Document Frequency*

[4] Entende-se por acurácia a porcentagem geral de acerto do classificador.

[5] *Macro-F* é a média das *F-measures* das sete categorias.

Tabela 4.5: Desempenho do AZ e das *baselines*, calculado em comparação com um anotador humano (Teufel & Moens, 2002)

	AZ	*Baseline* 1	*Baseline* 2	*Baseline* 3
Kappa	0.45	0.30	0	-0.13
Acurácia	0.73	0.72	0.48	0.67
Macro-F	0.50	0.30	0.14	0.11

Também foram realizados experimentos utilizando algoritmos mais complexos de aprendizado de máquina (o indutor de classificadores baseados em regras RIPPER (Cohen, 1995, 1996) e um classificador baseado em Máxima Entropia (Mikheev, 1998)), porém os resultados alcançados foram similares ao do classificador bayesiano (Teufel, 1999).

4.4 Comentários sobre os Sistemas

Embora os sistemas apresentados neste capítulo tenham objetivos finais distintos (avaliação, auxílio à leitura/escrita e sumarização) e, dessa forma, trabalham com tipos de textos e esquemas de estruturação retórica diferentes, todos têm em comum um componente de classificação automática das sentenças de um texto acadêmico em um conjunto pré-determinado de categorias retóricas. Existe uma variação considerável de desempenho entre os classificadores, mas a sua comparação direta é complexa, já que os textos e categorias utilizadas diferem e, como conseqüência, a complexidade da tarefa também difere. Entretanto, com a análise apresentada, é possível ter-se uma idéia do nível de desempenho alcançado pelos classificadores retóricos atuais.

Todos os autores pesquisados realizaram experimentos utilizando mais de um algoritmo de aprendizado de máquina e foi constatado que, para esse tipo de tarefa de classificação, classificadores simples como o Naive Bayes apresentam uma performance similar à de métodos mais complexos, proporcionando, dessa forma, uma boa relação de custo-benefício (custo do processo de treinamento do classificador em relação ao benefício da precisão da classificação).

Com relação aos atributos de classificação, existe variação entre os sistemas apresentados, contudo, todos são fortemente baseados em duas características principais: (i) a presença de marcadores discursivos e/ou expressões que funcionam como marcadores nas sentenças e (ii) o contexto de ocorrência da sentença (ou fluxo de categorias). Experimentos reportados tanto por Anthony & Lashkia (2003) (Mover) quanto por Teufel & Moens (2002) (AZ), apontam os atributos relacionados às duas características citadas como sendo os atributos com maior poder de distinção entre as categorias desejadas. Burstein, Marcu, & Knight (2003); Burstein, Chodorow, & Leacock (2003) (Critique) não fazem observações sobre o poder de cada atributo individualmente.

Dois dos classificadores abordados (Critique e AZ) trabalham exclusivamente com textos em inglês, já que utilizam recursos e ferramentas específicas para essa língua. O Mover, em princípio, seria capaz de classificar textos acadêmicos em qualquer língua, já que as expressões utilizadas como atributos são determinadas automaticamente durante o treinamento e o sistema não faz uso de recursos construídos previamente.

Para testar a aplicabilidade do Mover a textos em outra língua, realizamos um experimento de classificação utilizando um córpus de resumos de dissertações/teses em português, provenientes do CorpusDT (descrito no Capítulo 5). Os resultados alcançados, acurácia em torno de 45% e e *Kappa* em torno de 0.2, ficaram bem abaixo dos relatados por Anthony & Lashkia (2003) para o inglês. Duas hipóteses principais foram levantadas na tentativa de justificar a queda no desempenho. A primeira hipótese está relacionada com as diferenças dos córpus de treinamento. Embora ambos os córpus, tanto em inglês (I) quanto em português (P), sejam compostos por resumos acadêmicos, I é composto apenas por sentenças que foram classificadas por anotadores humanos sem nenhum grau de dúvida, enquanto P é composto por resumos completos, incluindo sentenças com problemas de ambigüidade na anotação manual. A ambigüidade presente no córpus afeta diretamente os resultados, já que o classificador irá reproduzir os padrões aprendidos durante o treinamento. Outra hipótese está relacionada com a diferença entre as língua (inglês vs. português). Foi observado que a variação de expressões padrão presentes no córpus P é maior que no córpus I, talvez devido à maior variação morfológica da língua portuguesa em relação à língua inglesa (em especial, no contexto de textos científicos). Como a seleção das expressões que serão usadas como atributos é baseada nas suas freqüências, o algoritmo encontra problemas em distinguir expressões que sejam significativas para a classificação quando treinado com o córpus P, uma vez que, devido à grande variação de expressões, as freqüências são relativamente baixas e uniformes. Embora essas sejam hipóteses que parecem válidas, ambas necessitam de uma investigação mais aprofundada para que possam ser confirmadas.

4.5 Considerações Finais

Neste capítulo foi feita uma revisão das ferramentas encontradas na literatura que fazem a detecção automática da estrutura retórica de textos acadêmicos. Todas as ferramentas utilizam aprendizado de máquina, predominantemente, modelos estatísticos, e baseiam-se em características superficiais do texto.

Com base nesta revisão, pode-se concluir que, em princípio, qualquer uma das abordagens utilizadas pelos sistemas descritos poderia ser adaptada de forma direta para a construção de um classificador retórico para textos acadêmicos em português. Entretanto, algumas dificuldades fazem dessa uma tarefa não trivial:

- A construção de um bom córpus de treinamento não é trivial, devido a dois fatores principais: (i) a dificuldade em se conseguir textos acadêmicos bem escritos em português (como os publicados em periódicos relevantes, que em geral são em inglês) e (ii) ao custo de se fazer a anotação retórica manual de um córpus volumoso;
- Falta de ferramentas robustas para o processamento do português;
- A própria complexidade da língua portuguesa (grande variação morfológica e sintática).

Considerando-se tais limitações e os objetivos deste trabalho relacionados ao uso de tal classificador, decidimos seguir a abordagem apresentada por Teufel & Moens (2002) (AZ), descrita na Seção 4.3. Embora o AZ tenha sido projetado para ser parte de um sistema de sumarização, as categorias utilizadas podem ser facilmente mapeadas para um esquema de anotação que sirva aos nossos propósitos. Ainda, a determinação dos atributos utilizados pelo AZ não é tão dependente de ferramentas especializadas (como o Critique, que utiliza um parser retórico), o que facilita sua adaptação para outras línguas, neste caso, o português.

O sistema de detecção automática da estrutura esquemática de Resumos de teses e dissertações, chamado *AZPort– Argumentative Zoning for Portuguese* –, incluindo as várias etapas de sua construção (desde a criação do córpus de treinamento até o modelo estatístico utilizado na classificação), será apresentado no Capítulo 6, como parte do ambiente de auxílio à escrita acadêmica proposto nesta tese.

Capítulo

5

Levantamento de Requisitos e Análise de Córpus

Como foi apresentado no Capítulo 3, existem várias ferramentas de auxílio à escrita acadêmica já desenvolvidas ou ainda em desenvolvimento. No entanto, todas são voltadas para a escrita de textos acadêmicos em inglês, seja para autores nativos ou não. Não foram encontradas referências a ferramentas dessa natureza voltadas para a escrita acadêmica em português. Dessa forma, o projeto de uma ferramenta computacional de auxílio à escrita de textos acadêmicos em português é uma iniciativa inédita nessa área.

Para nos guiar em tal projeto, utilizamos a seguinte abordagem: primeiramente investigamos a literatura especializada sobre sistemas computacionais dessa natureza, mesmo que para outra língua, e levantamos os problemas que esses sistemas atacam e como são abordados. Numa segunda etapa, analisamos um córpus de textos acadêmicos escritos em português (o CorpusDT, apresentado na Seção 5.2), visando identificar problemas de escrita, como estruturação inadequada, uso de léxico inadequado, uso de tempo verbal inadequado, entre outros, que poderiam ser minimizados com o auxílio de uma ferramenta computacional especializada. Essa análise nos deu uma visão dos principais problemas de escrita e das possíveis soluções que poderiam ser implementadas para minimizá-los. Ainda, nos interessava saber o que os usuários potenciais esperam encontrar em um sistema computacional como esse, uma vez que não existe semelhante disponível. Para esse fim, realizamos uma pesquisa de opinião por meio de um questionário, aplicado a uma amostra da comunidade local de estudantes de pós-graduação.

O conjunto das atividades mencionadas acima, incluindo a revisão bibliográfica, pode ser entendido como a fase de levantamento de requisitos no desenvolvimento do modelo computacional objetivado neste trabalho. Neste capítulo, completamos o relato dos resultados desse levantamento, reportando a pesquisa de opinião, na seção seguinte, bem como a análise do corpus, na Seção 5.2.

5.1 Pesquisa com Potenciais-Usuários

Na pesquisa de opinião integrante do levantamento dos requisitos do modelo, aplicamos um questionário a um grupo de 28 estudantes de pós-graduação do ICMC[1], sendo 22 da área da Ciência da Computação, 2 da Matemática Computacional, 2 da Matemática e 2 da Estatística. Os estudantes da área da Ciência da Computação foram selecionados de modo que fossem entrevistadas pelo menos duas pessoas de cada uma das áreas de concentração da pós-graduação do ICMC. Dessa forma, evitamos uma amostra viciada em sujeitos que escrevam influenciados pelo mesmo orientador ou pelos mesmos modelos. Esses estudantes foram divididos em dois grupos: 4 iniciantes e 24 experientes. Chamamos de iniciantes os sujeitos sem experiência em escrita acadêmica (estudantes que haviam acabado ingressar no programa de mestrado e ainda não haviam apresentado a monografia de qualificação), e de experientes os que já tenham produzido algum texto desse tipo (estudantes que já estavam finalizando a dissertação de mestrado ou que faziam parte do programa de doutorado).

5.1.1 Descrição do Questionário

O questionário, iniciado com um pequeno texto explicativo, abordou três aspectos: (1) o conhecimento pessoal sobre a escrita de textos acadêmicos, (2) as funcionalidades de uma possível ferramenta e (3) o modo de interação da ferramenta com o usuário, totalizando 21 questões, que foram elaboradas como questões de múltipla escolha ou como questões em que os itens deveriam ser classificados de acordo com uma dada escala. Também foram elaboradas duas questões abertas para que os usuários pudessem fazer sugestões. O questionário completo é apresentado no Apêndice A.

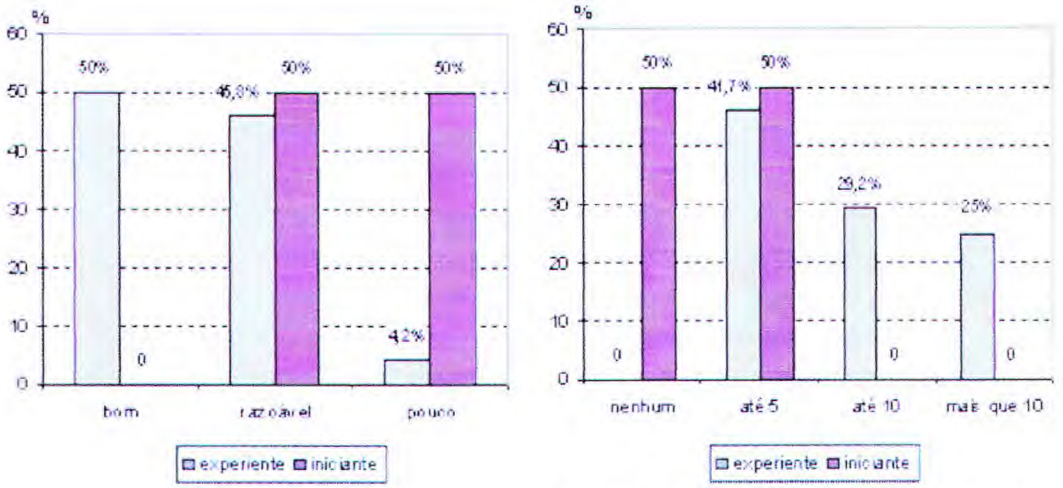
5.1.2 Resultados da Pesquisa

As figuras a seguir demonstram os resultados obtidos para cada grupo, em percentual. As figuras 5.1(a) a 5.4(b) enfocam a escrita acadêmica de modo geral, sem a interferência de aspectos computacionais. As figuras 5.5 e 5.6 enfocam as características funcionais de

[1]Instituto de Ciências Matemáticas e de Computação – USP/São Carlos

uma possível ferramenta computacional e as figuras 5.7(a) e 5.7(b) enfocam preferências quanto ao modo de interação com essa ferramenta.

Pela Figura 5.1(a) podemos notar que a maioria dos estudantes experientes considera seu conhecimento sobre a escrita acadêmica "bom" ou "razoável", ainda que apenas 25% já tenha escrito mais de 10 textos acadêmicos (Figura 5.1(b)). Mesmo entre os estudantes iniciantes, 50% consideram ter um conhecimento "razoável", embora parte dessa porcentagem nunca tenha escrito um texto acadêmico individualmente. Em contrapartida, por meio da análise do córpus, foi possível identificar problemas como estruturação inadequada, uso de léxico inadequado, uso de tempo verbal inadequado e outros problemas menos específicos, como erros ortográficos e gramaticais. Isso nos leva a crer que, embora os estudantes assumam ter um conhecimento bom ou razoável do gênero acadêmico, existem dificuldades na realização textual desse gênero.



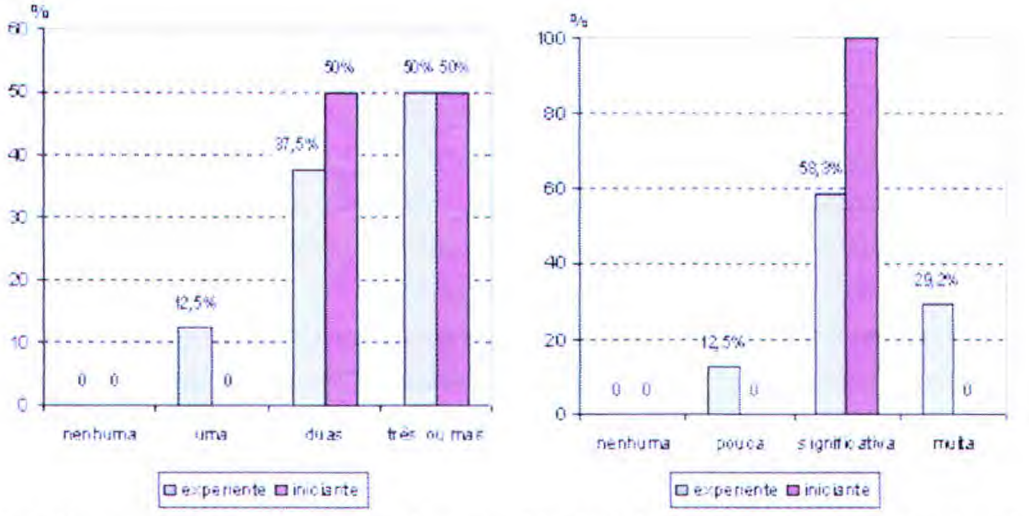
(a) Nível de conhecimento

(b) Número de textos acadêmicos produzidos

Figura 5.1: Respostas de estudantes iniciantes e experientes com relação à escrita acadêmica

A Figura 5.2(a) mostra que tanto estudantes experientes como iniciantes fazem várias revisões em seus textos antes de considerá-los prontos. É sabido que grande parte das revisões é feita pelo próprio orientador, daí sua grande influência na produção do texto final (Figura 5.2(b)). Essa influência também foi observada no córpus. Textos produzidos por estudantes sob a mesma orientação tendem a ser semelhantes tanto na estrutura como nas escolhas lexicais. Essa semelhança também advém dos textos utilizados pelos estudantes como modelos no momento da escrita do seu próprio texto, conforme apresentado na Figura 5.3.

Como pode ser visualizado na Figura 5.3, 50% dos estudantes experientes utilizam como modelo um texto anterior produzido por seu orientador ou por outro estudante



(a) Número médio de revisões em cada texto

(b) Nível de influência do orientador (ou alguém mais experiente) na produção de um texto

Figura 5.2: Respostas de estudantes iniciantes e experientes com relação à revisão dos textos produzidos

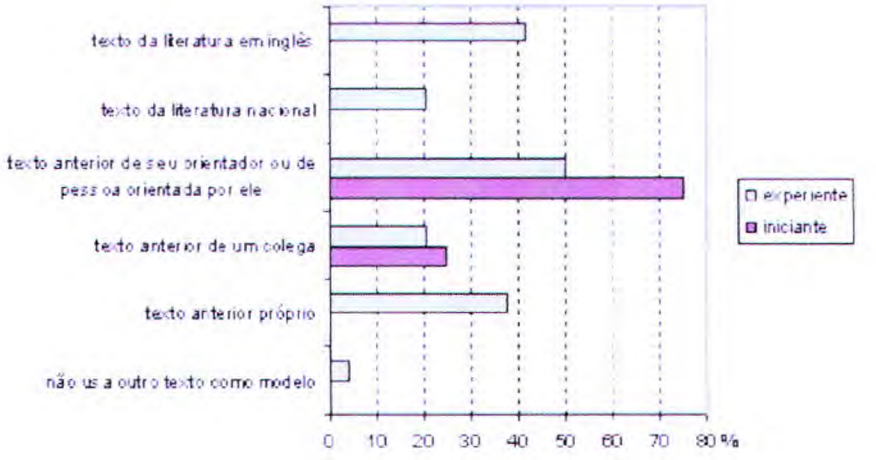
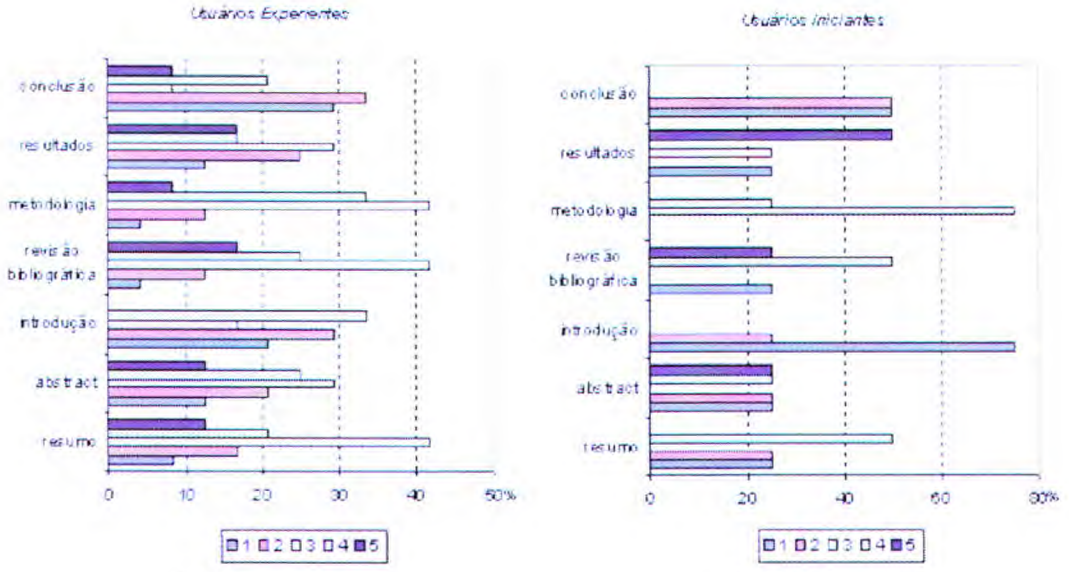


Figura 5.3: Respostas de estudantes iniciantes e experientes quanto ao tipo de texto utilizado como modelo para a escrita do seu próprio texto

orientado por ele. Entre os estudantes iniciantes essa porcentagem é ainda maior. Esse é outro fator que contribui para a semelhança dos textos do córpus. Ainda sobre os modelos, vale destacar que 41,7% dos estudantes experientes utilizam textos da literatura em inglês. De fato, conforme discutido no Capítulo 2, a estrutura esquemática do gênero acadêmico não apresenta diferenças em suas realizações nas línguas inglesa e portuguesa. Portanto, textos em inglês podem ser utilizados como bons modelos, e de fato isso ocorre.

Em uma das questões do questionário foi pedido que os estudantes indicassem o grau de dificuldade encontrada na escrita de cada seção padrão de um texto acadêmico, utilizando uma escala de 1 a 5, em que o 1 indicaria maior grau de dificuldade. A Figura 5.4(a) mostra a porcentagem das respostas dos estudantes experientes. Podemos observar que a Conclusão foi considerada a seção mais problemática (29,2%), seguida da Introdução (20,8%). A seção Revisão Bibliográfica foi considerada o menos problemática (25% grau 4 e 16,7% grau 5). A Figura 5.4(b) mostra a porcentagem das respostas dos alunos iniciantes, em que a Introdução foi considerada a seção mais problemática (75%), seguida da Conclusão (50%). A seção Resultados foi considerada o menos problemática para os estudantes iniciantes (50%).



(a) Experientes (b) Iniciantes

Figura 5.4: Respostas de estudantes experientes e iniciantes quanto ao grau de dificuldade encontrado na escrita de cada seção

Numa segunda etapa do questionário, foi apresentada uma lista de funções que poderiam ser implementadas em uma ferramenta computacional de auxílio à escrita acadêmica. Foi pedido aos estudantes que classificassem essas funções utilizando uma escala com quatro níveis: "indispensável", "desejável", "interessante" e "dispensável". A descrição das funções e da escala, tal qual foi apresentada aos estudantes, pode ser vista no Apêndice A.

Conforme mostrado na Figura 5.5, duas funções foram destacadas como indispensáveis pelos estudantes experientes: "correção ortográfica" (95,8%) e "correção gramatical e de estilo" (91,7%). Acreditamos que o fato dessas funções já serem conhecidas e utilizadas pelos estudantes provocou um resultado tão expressivo em relação às outras funções. A função "revisão do texto final" também se destacou como indispensável (50%). Com

exceção dos dois tipos de *brainstorming* sugeridos (ambos considerados indispensáveis por apenas 16,7%), o restante das funções obteve classificações semelhantes. Ao contrário do que aconteceu com as funções de correção ortográfica, gramatical e de estilo, acreditamos que o fato de os estudantes nunca terem utilizado os outros tipos de funções dificultou a classificação, de modo que não houve resultados mais expressivos.

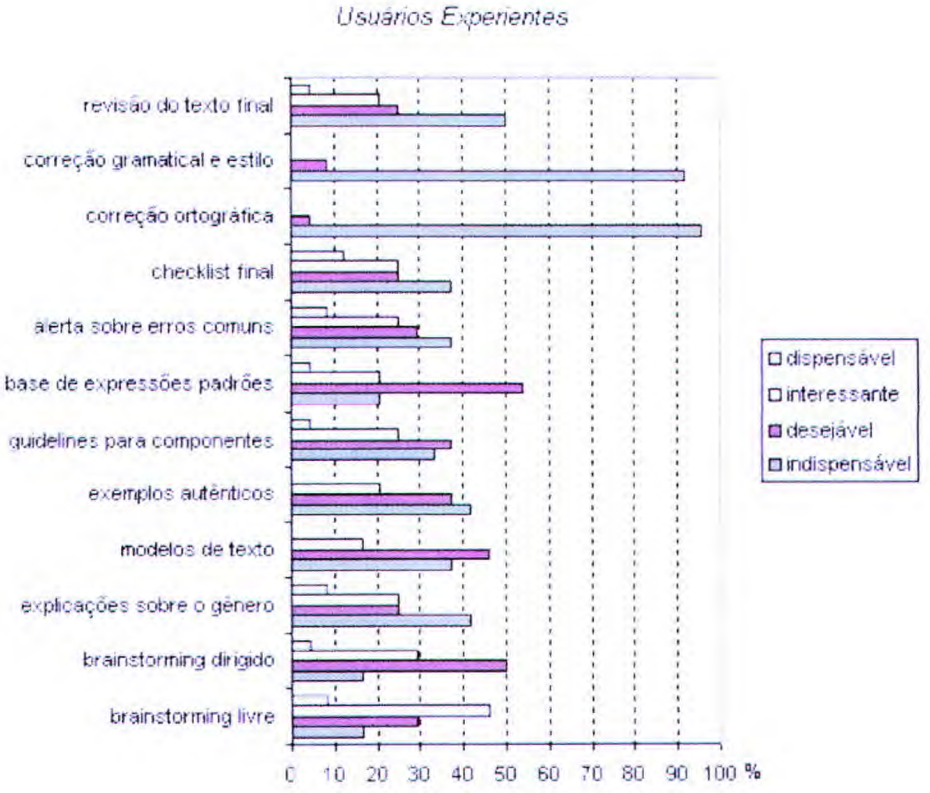


Figura 5.5: Respostas dos estudantes experientes quanto a relevância das funções de uma ferramenta computacional de auxílio à escrita

Os estudantes iniciantes também destacaram as funções "correção ortográfica" e "correção gramatical e de estilo" como indispensáveis (ambas com 75%). Diferentemente dos estudantes experientes, 75% dos iniciantes julgaram a função "*guidelines* para a escrita de seções" indispensável. A função "uso de exemplos autênticos" não foi julgada indispensável, no entanto, 100% dos estudantes iniciantes a classificaram com desejável. Uma vez que esses estudantes têm pouca ou nenhuma experiência na escrita acadêmica, é natural que tenham se mostrado mais interessados em *guidelines* e modelos autênticos do que os estudantes experientes. Como pode ser visto na Figura 5.6, nenhuma função foi julgada como dispensável pelos iniciantes.

É importante recordar que a amostra de estudantes iniciantes (4) é bem menor do que a amostra de estudantes experientes (24), por isso os dados apresentados na Figura 5.6

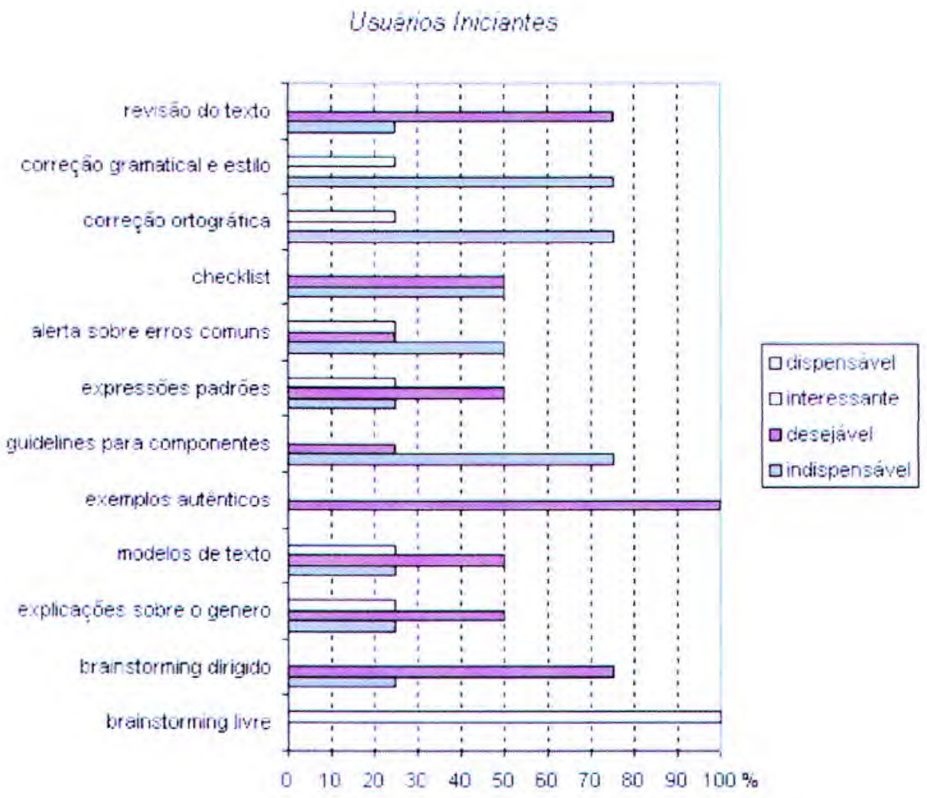


Figura 5.6: Respostas dos estudantes iniciantes quanto a relevância das funções de uma ferramenta computacional de auxílio à escrita

podem parecer mais expressivos que os da Figura 5.5, mas, na realidade, os estudantes iniciantes apresentaram as mesmas dificuldades sentidas pelos estudantes experientes em classificar as funções dentro da escala sugerida.

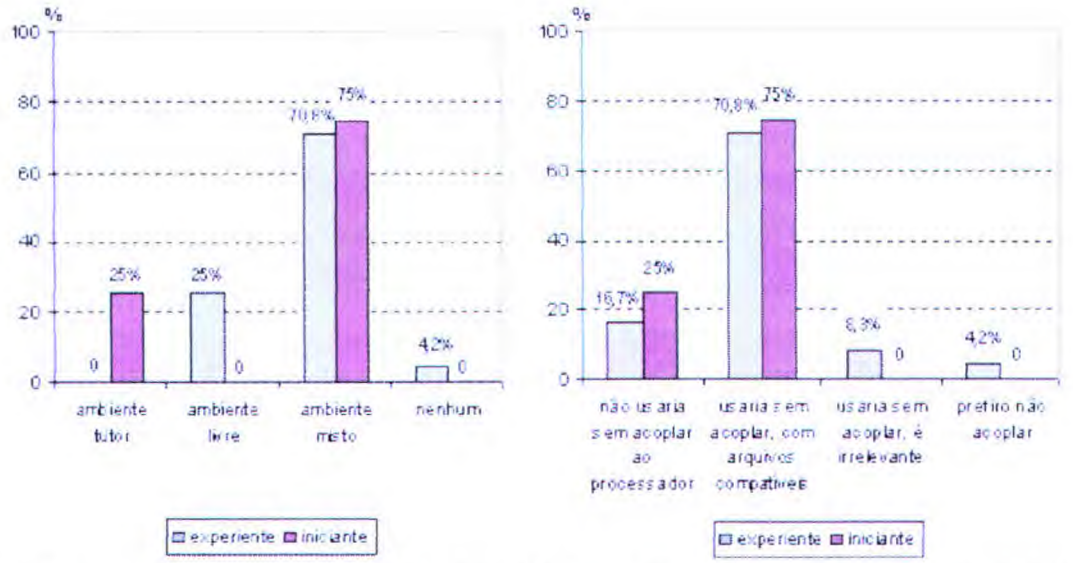
Foi dado aos estudantes um espaço para que sugerissem outras funções que julgassem interessantes, diferentes das listadas no questionário. Alguns estudantes (2 iniciantes e 8 experientes) fizeram as seguintes sugestões de recursos ou funções desejáveis:

- lista de palavras comuns ao gênero acadêmico com os respectivos significados e sinônimos;
- lista de palavras/expressões que devem ser evitadas no gênero acadêmico, com sugestões de substituição;
- manutenção de um histórico dos textos feitos por um usuário na ferramenta e inserção desses textos numa base para posterior recuperação e reutilização;
- dicionário português/inglês on-line;
- geração automática de abstract a partir do resumo;
- apoio de uma ferramenta de tradução automática;
- apoio de uma ferramenta de sumarização automática;

- montagem de uma *checklist* personalizada pelo usuário;
- apoio ao controle de versões;
- criação de um espaço para anotações, de forma que essas anotações fossem ligadas ao texto original por *links*;
- verificação da coerência entre Introdução e Conclusões;
- verificação da coerência entre as Considerações Iniciais e Finais de cada capítulo.

Quanto ao modo de interação com o usuário (Figura 5.7(a)), ficou clara a preferência dos estudantes, tanto experientes quanto iniciantes, por um ambiente misto, isto é, em que o usuário seja livre para acionar as funções que desejar e que o sistema também interfira quando achar conveniente, porém sem a intensidade de um tutor. Foi sugerido por dois estudantes experientes que o nível de interferência do sistema possa ser personalizado, ou seja, que o usuário possa configurar itens em que o sistema pode interferir ou não, e que esse perfil seja armazenado para posteriores utilizações.

A Figura 5.7(b) evidenciou que o não acoplamento de uma ferramenta de auxílio à escrita acadêmica a um processador de textos não altera a motivação da maioria dos estudantes, tanto experientes (70,8%) como iniciantes (75%), em usá-la, desde que os arquivos manipulados pela ferramenta sejam compatíveis com os manipulados pelos processadores de texto.



(a) Estilo de interação de interação com o usuário

(b) Relevância do acoplamento de uma ferramenta de auxílio à escrita acadêmica a um processador de textos

Figura 5.7: Respostas de estudantes iniciantes e experientes quanto aos aspectos operacionais de uma ferramenta de auxílio à escrita acadêmica

5.1.3 Considerações sobre os Resultados

Como foi citado no início desta seção, o intuito dessa pesquisa foi investigar o que potenciais usuários esperam de uma ferramenta de auxílio à escrita acadêmica. Essa experiência foi muito importante, pois nos deu um cenário mais completo das dificuldades desses usuários e do tipo de auxílio que esperam encontrar. Mais especificamente, esses resultados serviram de complemento ao levantamento dos problemas de escrita e às possíveis soluções investigadas (apresentadas no Capítulo 3), no sentido de adequar as abordagens de auxílio à escrita ao cenário dos problemas tratados nesta tese.

Dessa maneira, as seguintes observações (feitas a partir dos resultados apresentados na subseção anterior) nortearam nossa proposta:

- é comum o uso de outros textos como modelos, o que evidencia o reuso de estruturas esquemáticas e de material lingüístico;
- a escrita da Introdução é considerada problemática, tanto por estudantes experientes quanto por iniciantes;
- a escrita do Resumo é considerada como tendo um grau médio de dificuldade, embora a análise do córpus tenha evidenciado problemas de escrita nessa seção;
- além das funções tradicionais de auxílio à escrita (corretores ortográficos, gramaticais e estilísticos), foi demonstrado especial interesse pelas seguintes funções: revisão do texto final, acesso a *guidelines* para escrita das seções e uso de exemplos autênticos;
- considerando o modo de interação, é dada preferência a um modelo misto, em que o usuário é livre para acionar as funções que desejar, mas em que o sistema também possa interferir quando for necessário;
- o fato de o ambiente não estar acoplado a um processador de texto não limita seu uso.

5.2 Coleta, Anotação e Análise do Córpus

Para a investigação da estrutura esquemática e dos padrões de escrita utilizados por escritores brasileiros em teses e dissertações, fez-se necessária a formação de um córpus específico para esse propósito. No entanto, ainda era preciso restringir o córpus a uma área de conhecimento específica, pois sabe-se que existem diferenças estruturais entre diferentes áreas e que, quanto mais específico for o propósito de um córpus, maior a chance de esse córpus ser representativo dentro do seu propósito. Sendo assim, decidimos direcionar a pesquisa a dissertações e teses em Ciência da Computação. A escolha por essa área se deveu principalmente ao fato deste trabalho estar sendo desenvolvido em um departamento de Ciência da Computação. O córpus resultante foi chamado de CorpusDT.

5.2.1 Os Textos Coletados

Devido ao fato de trabalharmos em um departamento de Ciência da Computação, a coleta das teses e dissertações foi de certa forma facilitada. Todos os trabalhos foram coletados já em formato eletrônico. Alguns arquivos foram encontrados no repositório público de teses e dissertações da USP[2], outros nos sites dos respectivos grupos de pesquisa. O restante dos arquivos foi coletado diretamente com seus autores ou com seus respectivos orientadores.

Uma vez que os arquivos estavam em vários formatos, sendo alguns deles não tratáveis por ferramentas para manipulação de córpus (p.e., arquivos *postscritp*), foi necessário um grande esforço para a conversão desses arquivos. Além da conversão em si, foi necessário extrair as figuras, gráficos e tabelas dos textos, substituindo cada um desses elementos por uma etiqueta que os identificasse. O conteúdo textual dos trabalhos foi mantido inalterado, incluindo possíveis erros ortográficos/gramaticais. Os trabalhos estão hoje armazenados em arquivos do tipo texto sem formatação (TXT).

Foram coletados trabalhos de várias áreas de pesquisa da Ciência da Computação, num total de 52 arquivos, sendo 49 dissertações de mestrado e 3 teses de doutorado. Os trabalhos datam de 1994 a 2001, sendo que a grande parte data de 1999 a 2001. A Tabela 5.1 mostra a distribuição dos trabalhos entre as áreas de pesquisa identificadas no córpus.

Tabela 5.1: Número de trabalhos em cada área de pesquisa

Área de Pesquisa	Dissertação	Tese
Banco de Dados	3	-
Inteligência Computacional	7	-
Engenharia de Software	15	1
Hipermídia	12	-
Sistemas Digitais	1	-
Sistemas Distribuídos e Programação Concorrente	10	2
Computação Gráfica e Processamento de Imagens	1	-
Total	49	3
	52	

Cada dissertação ou tese está armazenada em um único arquivo texto, sendo que foram mantidas as identificações iniciais de cada trabalho como primeiro elemento dentro do texto. Os arquivos estão identificados de acordo com o seu tipo (dissertação ou tese) e de acordo com a área de pesquisa a que pertencem. Os nomes de arquivos iniciados em "t_" identificam teses de doutorado, os restantes identificam dissertações de mestrado. A sigla que segue "t_" ou que nomeia o arquivo identifica a área de pesquisa do trabalho. Cada arquivo é identificado dentro da sua área por um número. Desse modo, o arquivo

[2]www.teses.usp.br

t_es1.txt identifica uma tese de doutorado da área de Engenharia de Software.

As Tabelas 5.2 e 5.3 apresentam alguns números que dão uma idéia da distribuição do córpus com relação ao seu tamanho. A Tabela 5.2 mostra uma comparação entre os tipos de texto, dissertação e tese. São apresentados os totais de palavras no córpus de cada tipo (dissertações e teses), a média de palavras em dissertações e em teses e também o número de palavras da maior e da menor dissertação e tese presente no córpus.

Tabela 5.2: Totais de palavras em dissertações e teses

	Total	Média	Maior	Menor
Teses	173.579	57.860	64.473	54.497
Dissertações	1.429.323	29.170	47.071	13.653

A Tabela 5.2 mostra os totais de palavras distribuídos pelas áreas da Ciência da Computação contempladas no córpus. Para a soma desses totais foram consideradas tanto as dissertações quanto as teses. É apresentado para cada área: o número de textos da área, o total de palavras nos textos daquela área e a média de palavras por texto. Também é mostrado o total de palavras global do córpus (1.602.902 palavras) e a média geral de palavras, isto é, se tomados todos os textos de todas as áreas.

Tabela 5.3: Totais de palavras divididos por área de conhecimento

Área de Pesquisa	No. textos	No. palavras	Média
Banco de Dados	3	79.561	26.520
Inteligência Computacional	7	181.757	25.965
Engenharia de Software	16	556.441	34.778
Hipermídia	12	312.117	26.010
Sistemas Digitais	1	26.894	26.894
Sistemas Distribuídos e Programação Concorrente	12	425.139	35.428
Computação Gráfica e Processamento de Imagens	1	20.993	20.993
Total	52	1.602.902	30.825

5.2.2 Anotação Eletrônica do CorpuDT

Para que o córpus possa ser manipulado e "compreendido" por ferramentas computacionais é necessário que ele esteja eletronicamente anotado com um conjunto de etiquetas apropriadas. Para essa anotação, optamos por utilizar o XML (*Extensible Markup Language*), um padrão para a anotação de documentos, aprovado pelo W3C[3], que vem sendo mundialmente utilizado para a anotação de córpus. Optamos pelo XML principalmente pela facilidade de manipulação, visto que um documento XML pode ser visualizado em *browsers* comuns, como o Microsoft Internet Explorer.

[3]World Wide Web Consortium - http://www.w3.org [11-10-2004]

O XML é subconjunto do SGML e foi criado para a manipulação de documentos na Internet. Esse padrão incorpora características do SGML e amplia o poder de estruturação de documentos, facilitando o intercâmbio de dados entre aplicações e possibilitando que o formato de apresentação dos dados seja independente da anotação, uma vez que o XML não especifica como os seus elementos devem ser apresentados. Na verdade, não especifica nem mesmo se eles devem ser apresentados. Uma forma de apresentação de documentos XML consiste em transformá-los em documentos HTML[4], possibilitando assim que sejam visualizados em um *browser* comum (W3C, 2001).

A definição de um documento XML pode ser feita por meio de um *XML Schema*, que define as etiquetas que serão utilizadas para a estruturação dos dados do documento. Considerando o tipo informação manipulada neste trabalho, propomos um conjunto de etiquetas para a anotação do CorpusDT em quatro níveis:

- anotação do layout da dissertação (identificação, capítulos, etc);
- anotação de componentes das seções Resumo e Introdução;
- anotação de estratégias retóricas de cada componente;
- anotação de padrões de textos que possam ser reutilizados nos Resumos e Introduções;
- anotação de marcadores discursivos.

O *XML schema* que define esse conjunto de etiquetas é apresentado no Apêndice B. As etiquetas foram propostas com base na literatura pesquisada sobre a estruturação esquemática de textos acadêmicos e também na anotação utilizada nas bases de casos do projeto AMADEUS. É importante destacar que o *schema* definido é bastante flexível no nível dos componentes e das estratégias retóricas, pois os nomes dos componentes e estratégias não foram especificados. Sendo assim, haverá sempre um modelo de componentes e estratégias que definirá esses nomes, como o exemplo apresentado na Figura 5.8 da seção anterior. Isso proporciona liberdade no caso de serem necessárias modificações no modelo e permite que diferentes modelos sejam utilizados em conjunto com o *schema* definido.

5.2.3 Análise da Estrutura do CorpusDT

Com o objetivo de levantar certos padrões esquemáticos presentes no CorpusDT, foi realizada uma análise da estruturação dos textos, visando estabelecer um recorte das partes que seriam mais aprofundadas no momento da anotação.

Feita uma primeira análise, notou-se que os componentes de mais alto nível relatados na bibliografia pesquisada de fato encontravam-se nos textos do córpus. Os macrocompo-

[4] *Hypertext Markup Language*

nentes presentes nas dissertações e teses analisadas são: Resumo, *Abstract*[5], Introdução, Revisão da Literatura, Metodologia Utilizada, Resultados Obtidos, Discussão dos Resultados e Conclusões.

Desses macrocomponentes, os que aparecem nos textos de forma bem definida são Resumo, Introdução e Conclusão. Esse fato foi constatado na análise do córpus e já era esperado, pois essas seções exercem papéis extremamente importantes em um texto acadêmico. O macrocomponente Revisão da Literatura aparece tanto de forma definida, em uma única seção, como "misturada" a outros macrocomponentes como, por exemplo, a Metodologia. Em geral, isso depende do assunto abordado na dissertação ou tese. Quando a dissertação (ou tese) trata de um assunto específico, é comum que a Revisão da Literatura seja escrita como uma seção única. Quando a dissertação (ou tese) trata vários assuntos (por exemplo, trabalhos que envolvem mais de uma subárea de pesquisa da Ciência da Computação e trabalhos multidisciplinares), esse macrocomponente tende a ser dividido em mais de uma seção, geralmente acompanhando as partes da Metodologia.

Os outros macrocomponentes observados, Metodologia Utilizada, Resultados Obtidos e Discussão dos Resultados, aparecem no córpus seguindo diferentes padrões de combinação. Em geral, os três são combinados em diferentes seções de texto, seguindo a ordem metodologia-resultado-discussão dentro de cada seção. No entanto, esse padrão pode ser cíclico, isto é, pode ficar se repetindo dentro da seção. Isso faz com que os macrocomponentes fiquem "embaralhados", dificultando a identificação de cada um. Outro padrão comum é a combinação dos Resultados Obtidos com a Discussão dos Resultados em uma ou mais seções. Nesse caso, também pode ocorrer a repetição cíclica do padrão dentro de uma mesma seção. Em alguns casos, cada um desses macrocomponentes aparece isoladamente, mas isso não se mostrou comum no córpus.

5.2.4 Anotação e Análise da Estrutura Esquemática dos Resumos e Introduções

Uma vez que as seções Resumo e Introdução são os focos desta tese, coletamos alguns dados numéricos que demonstram de forma geral elas ocorrem no CorpusDT. A Tabela 5.4 apresenta a média de palavras dos resumos, mostrando também o tamanho, em palavras, do maior e do menor resumo encontrados no córpus.

Tabela 5.4: Dados sobre o tamanho em palavras dos Resumos

Média de palavras	Maior	Menor
188	*425*	*65*

[5]Neste trabalho, não citaremos o *abstract* como um componente, por se tratar de uma parte de texto escrita em inglês que, normalmente, é uma tradução do resumo.

A Tabela 5.5 apresenta a média de palavras das introduções em geral, incluindo dissertações e teses, a média de palavras das introduções de dissertações e a média de palavras das introduções de teses. Nessa tabela também consta o tamanho, em palavras, da maior e da menor introdução encontradas no córpus.

Tabela 5.5: Dados sobre o tamanho em palavras das Introduções

Média Dissertação	Média Tese	Média Total	Maior	Menor
1369	1981	1404	2652	359

Para que pudéssemos analisar a estrutura esquemática (ou retórica) dos Resumos e Introduções, foi realizada a anotação manual dessas seções, com base em modelos estruturais definidos especialmente para esta tarefa. Esse processo de anotação manual considerou três níveis de análise: identificação de componentes estruturais, identificação de estratégias retóricas utilizadas para a realização de cada componente e identificação de expressões padrão que servem como sinalizadores do papel argumentativo sentença. Esses três níveis de análise foram empregados por refletirem os níveis de auxílio pretendidos no ambiente *SciPo*: como estruturar o texto (componentes), como realizar cada componente da estrutura (estratégias retóricas) e quais são as expressões comuns utilizadas em cada componente do texto. A seguir são apresentados os resultados da análise dos Resumos e Introduções individualmente.

Anotação e Análise dos Resumos

A Figura 5.8 mostra o modelo utilizado para a anotação dos resumos. O ponto de partida para a criação deste modelo foram três outros modelos já abordados nesta tese: (i) o CARS de Swales (1990), o de Weissberg & Buker (1990) e o de Aluísio & Oliveira Jr. (1996). Apesar dos modelos de Swales (1990) e Aluísio & Oliveira Jr. (1996) descreverem originalmente a estrutura de introduções de artigos científicos, uma análise inicial dos resumos mostrou que suas estruturas básicas podem ser aplicadas também a resumos, desde que observadas as particularidades de cada seção.

Durante o processo de anotação manual dos resumos, foi observada uma dificuldade na distinção dos componentes Propósito e Resultado. Acreditamos que tal dificuldade se deve ao fato do córpus ser de uma área científico-tecnológica, em que muitas vezes o resultado do trabalho vem embutido no componente Propósito. Dificuldades de distinção também foram observadas envolvendo os componentes Metodologia e Resultado. É importante destacar que, embora o córpus seja composto por dissertações e teses defendidas, elas não passaram por uma revisão rigorosa, como é o caso, por exemplo, de artigos publicados em revistas e em conferências de prestígio. Esse fato também acabou por dificultar a anotação, devido a vários problemas de escrita, como estruturação confusa e erros gramaticais. De

Contexto
- C1. Declarar proeminência do tópico
- C2. Familiarizar termos e conceitos
- C3. Introduzir a pesquisa a partir da grande área

Lacuna
- L1. Citar problemas/dificuldades
- L2. Citar necessidades/requisitos
- L3. Citar a ausência ou pouca pesquisa anterior

Propósito
- P1. Indicar o propósito principal
- P2. Detalhar/Especificar o propósito
- P3. Introduzir mais propósitos

Metodologia
- M1. Listar critérios ou condições
- M2. Citar/Descrever materiais e métodos
- M3. Justificar a escolha pelos materiais e métodos

Resultado
- R1. Descrever o artefato
- R2. Apresentar resultados
- R3. Comentar/Discutir resultados

Conclusão
- Co1. Apresentar conclusões
- Co2. Apresentar contribuições/valor do trabalho
- Co3. Apresentar recomendação

Figura 5.8: Modelo utilizado na anotação dos Resumos do CorpusDT

qualquer maneira, sabe-se que a anotação da estrutura retórica de textos acadêmicos é uma tarefa bastante subjetiva e custosa, mesmo em se tratando de textos bem escritos.

A anotação dos Resumos permitiu o reconhecimento de padrões de estruturação recorrentes, seja na forma de padrões previstos no modelo ou de padrões que ocorrem dentro de um limite aceitável de desvio. Com relação à ocorrência dos componentes do modelo nos resumos do córpus, Propósito, Resultado e Metodologia foram os componentes mais freqüentes. A freqüência de ocorrência de todos os elementos do modelo é apresentada na Tabela 5.6. Esses resultados estão de acordo com outros trabalhos da literatura sobre a análise estrutural de resumos (Santos, 1996; Motta-Roth & Hendges, 1998). Esses autores também relatam o caráter opcional do elementos iniciais e finais de seus modelos e uma tendência ao uso mais freqüente dos elementos centrais, em especial do elemento Propósito.

Também foram analisados os desvios do modelo padrão que podem ser prejudiciais ao resumo. São considerados desvios prejudiciais aqueles que podem comprometer a interpretação das informações contidas no resumo ou que comprometem o entendimento do trabalho ao qual o resumo pertence. Os principais desvios observados foram: (i) o uso incorreto de expressões padrão e tempos verbais característicos de certos componentes, (ii) organização estrutural ineficiente e (iii) ênfase inapropriada em componentes específicos.

Tabela 5.6: Freqüência de componentes observada no córpus de Resumos

Componente	Freqüência
Contexto	55.7%
Lacuna	42.3%
Propósito	100%
Metodologia	63.4%
Resultado	67.3%
Conclusão	30.7%

Por (i), entendemos aqueles casos em que o escritor utiliza num componente, expressões padrão características de outros. Por exemplo, o uso do marcador discursivo "no entanto", que usualmente sinaliza uma sentença de Lacuna, em uma sentença de Contexto e sem exercer o seu papel de marcador de contraste.

Por (ii), entendemos os casos em que (a) a estruturação do Resumo, embora contendo elementos do modelo padrão, é feita de tal forma que a interpretação do texto pode ser prejudicada ou, no mínimo, dificultada. Um exemplo de tal problema seria a estrutura [P M C L P]. O escritor iniciou o resumo indicando o propósito principal (primeiro P) e então descreveu a metodologia utilizada (M). Nesse ponto do Resumo, um movimento mais natural seria apresentar os resultados, entretanto, o escritor fez uma nova contextualização com indicação de lacuna (C L), como um meio de introduzir ao leitor um detalhamento do propósito principal (último P). A presença dos componentes Contexto (C) e Lacuna (L) separando o propósito principal do seu detalhamento pode prejudicar a interpretação do texto, levando o leitor a ter dificuldades em identificar o real propósito da pesquisa relatada. Outro exemplo desse problema seria a estrutura [L C L P Co], em que faltam componentes considerados essenciais ao resumo, como Metodologia e Resultado.

Por (iii), entendemos os casos em que é dada ênfase a determinados componentes em detrimento de outros. O caso mais comum desse tipo de problema é a ênfase na contextualização (Contexto + Lacuna) em detrimento de outras informações, como Resultados. Um exemplo de tal problema seria um resumo com estrutura [C L C L P], em que o componente Propósito (P) corresponde a 16.9% do resumo. O restante do texto (83.1%) é dedicado à contextualização do problema. Considerando que se trata de um resumo de dissertação, em que não há um limite rigoroso de tamanho, seria desejável que o resumo apresentasse também informações diretamente relacionadas à pesquisa relatada (Metodologia e Resultado).

Três outros desvios também foram observados, embora com menor freqüência: (iv) trechos que não puderam ser classificados com nenhuma das etiquetas do modelo; (v) trechos indicando informações óbvias e, dessa forma, irrelevantes para o resumo e (vi) a presença da estrutura do trabalho, informação típica de introdução.

Os desvios do tipo (ii) e (iii) são os mais freqüentes no CorpusDT. Dos 52 resumos do

CorpusDT, 39 apresentam algum desvio relativo ao tipo (ii) e 12 resumos apresentam o desvio (iii).

Quanto as ocorrências do desvio (ii), em 30 resumos foi observada a ausência de componentes essenciais; em 5 foram observados problemas de organização e em 4 resumos foi observada tanto a ausência de componentes quanto problemas de organização. Com relação à ausência de componentes essenciais, a maior ocorrência envolve a falta dos componentes Resultado e Metodologia. Com relação a problemas de organização, a maior ocorrência envolve o uso do componente Conclusão precedendo os componentes Resultado ou Metodologia.

Quanto as ocorrências do desvio (iii), a ênfase na contextualização (considerados aqui os componentes Contexto e Lacuna) em detrimento de outros componentes, como Metodologia e Resultado, foi observada em 12 resumos. Nesses 12 casos, mais de 50% do resumo é dedicado à contextualização.

Anotação e Análise das Introduções

A Figura 5.9 mostra o modelo utilizado para a anotação das Introduções. Este modelo foi adaptado do modelo proposto por Aluísio & Oliveira Jr. (1996), a fim de acomodar as particularidades desse novo córpus. O processo de anotação manual foi similar ao realizado para os Resumos, contemplando os mesmos níveis de análise.

Durante o processo de anotação das Introduções, foram observadas as mesmas dificuldades de distinção entre componentes relatadas para os Resumos. Além disso, surgiu uma nova dificuldade de distinção entre os componentes Contexto e Revisão da Literatura, devido ao uso freqüente de citações em sentenças de contextualização. Notou-se que sentenças apresentando conhecimento aceito na área de pesquisa, ou seja, que atuam como contextualização, freqüentemente contêm citações (em geral ao orientador do trabalho ou obras importantes na área), embora muitas vezes desnecessárias. Devido à alta ocorrência desse fato no córpus, concluímos que isso é um procedimento comum nessa área. Ainda com relação às citações, o oposto também foi observado no córpus, isto é, introduções sem nenhuma citação, embora com menor freqüência. Nesses casos, a distinção entre contextualização e revisão da literatura também é problemática, sendo totalmente dependente do conhecimento do anotador com relação à área de pesquisa em questão.

Quanto aos desvios com relação ao modelo padrão de Introduções, novamente foram observados problemas semelhantes aos dos Resumos: (i) uso incorreto de expressões padrão e tempos verbais característicos de certos componentes, (ii) organização estrutural ineficiente e (iii) ênfase inapropriada em componentes específicos em detrimento de outros. Um novo problema relacionado à organização estrutural ineficiente encontrado nas introduções é (iv) a repetição desnecessária de informações. Isso foi observado especialmente

Contexto
- C1. Declarar proeminência do tópico
- C2. Familiarizar termos e conceitos
- C3. Introduzir a pesquisa a partir da grande área
- C4. Introduzir fatos da área

Revisão da Literatura
- Rv1. Apresentar revisão histórica da área
- Rv2. Citar tendências atuais na área
- Rv3. Organizar citações da área geral para o tópico
- Rv4. Indicar progresso na área
- Rv5. Citar/Descrever o estado da arte
- Rv6. Agrupar citações por abordagem
- Rv7. Citar/Descrever uma pesquisa particular

Lacuna
- L1. Citar problemas/dificuldades
- L2. Citar necessidades/requisitos
- L3. Citar a ausência ou pouca pesquisa anterior
- L4. Levantar questões
- L5. Continuar pesquisa anterior

Propósito
- P1. Fazer referência ao propósito
- P2. Indicar o propósito principal
- P3. Detalhar/Especificar o propósito
- P4. Introduzir mais propósitos

Metodologia
- M1. Listar critérios ou condições
- M2. Citar/Descrever materiais e métodos
- M3. Justificar a escolha pelos materiais e métodos

Resultado
- R1. Descrever o artefato
- R2. Apresentar resultados
- R3. Comentar/Discutir resultados

Contribuições/Valor do trabalho
- V1. Apresentar contribuições
- V2. Apresentar o valor do trabalho
- V3. Justificar o trabalho

Estrutura do trabalho
- E1. Indicar os capítulos/seções do trabalho
- E2. Indicar estrutura do capítulo/seção
- E3. Apresentar lista dos tópicos abordados

Figura 5.9: Modelo utilizado na anotação das Introduções do CorpusDT

em introduções que fazem uso de subseções. Muitas vezes, essa estratégia de organização da introdução em tópicos separados atrapalha a estruturação, pois o escritor passa a repetir o mesmo tipo de informação de formas diferentes em cada uma das subseções. Isso gera introduções excessivamente longas e difíceis de interpretar.

Assim como nos resumos, o desvio mais observado no CorpusDT foi o (ii) – organização estrutural ineficiente –, que envolve tanto a ausência de componentes essenciais quanto

problemas de organização, isto é, a ordem em que os componentes aparecem no texto. Foram analisadas 49 introduções do CorpusDT, sendo que 41 apresentaram algum desvio relacionado ao tipo (ii). Em 26 introduções foi observada a ausência de componentes essenciais; em 4 foram observados problemas de organização e em 11 introduções foi observada tanto a ausência de componentes quanto problemas de organização. Com relação à ausência de componentes essenciais, a maior ocorrência envolve a falta dos componentes Resultado e Metodologia. Com relação a problemas de organização, a maior ocorrência envolve o uso dos componentes Resultado e/ou Metodologia precedendo a definição do Propósito principal.

5.2.5 Considerações sobre os Resumos e Introduções

A anotação manual do córpus de Resumos e Introduções e a análise dos resultados dessa anotação nos permitiu confirmar que os modelos adotados nesta tese de fato refletem o padrão de organização estrutural da comunidade acadêmica focalizada. Além disso, possibilitou que fossem identificados diversos problemas de escrita específicos dessa comunidade, servindo de subsídios para a proposta do ambiente *SciPo*.

Além de servir à análise estrutural, a anotação manual dos Resumos e Introduções também visou (i) fornecer material de treinamento para o sistema *AZPort* (no caso dos Resumos) e (ii) compor as bases de exemplos autênticos do ambiente *SciPo*. Os papéis do córpus em (i) e (ii) serão detalhados no próximo capítulo (Subseções 6.2.2 e 6.1.1, respectivamente).

5.3 Considerações Finais

A partir dos resultados apresentados neste capítulo, foi possível eleger diversas características de interesse para o modelo de aplicação de auxílio à escrita desenvolvido. O cruzamento do que os usuários em potencial imaginam ser desejável com as observações do córpus, especialmente no que se refere a problemas de escrita, corrobora a hipótese de que terá utilidade um modelo com as seguintes características:

i. baseado em córpus, promovendo a reutilização de exemplos;
ii. baseado em crítica, predispondo a um desenvolvimento colaborativo;
iii. que permita liberdade quanto ao ponto de partida da interação (revisando um texto pronto ou ainda auxiliando no esboço do primeiro rascunho); e, portanto,
iv. que possibilite um refinamento cíclico do texto.

No capítulo seguinte será descrito um protótipo operacional que implementa tal modelo.

Dessa forma, podemos atribuir ao ambiente *SciPo* três macro-funções:

1. Apoio à estruturação do texto, que inclui o uso de modelos estruturais tidos como adequados e crítica da estrutura construída pelo usuário, permitindo também a visualização de exemplos de estruturas esquemáticas reais, extraídas de uma base de exemplos;
2. Apoio à estruturação a partir do texto pronto, que inclui a detecção automática da estrutura esquemática do texto submetido pelo usuário, bem como a crítica dessa estrutura, da mesma forma que ocorre em 1;
3. Apoio à escrita do texto, que inclui funções de suporte à escrita do texto de acordo com a estrutura refinada pelo usuário, seja essa estrutura proveniente de 1 ou de 2.

Na verdade, 1 e 2 são funções com pontos de partida distintos para um mesmo processo de crítica da estrutura, que conduz o usuário à produção de uma estrutura considerada adequada. Ao final desse processo, o usuário pode se beneficiar da função 3.

A Figura 6.1 mostra a relação entre essas macro-funções no ambiente *SciPo*, destacando os principais sistemas envolvidos na implementação de cada macro-função. Tais sistemas, representados na figura como retângulos, são: (i) o sistema de suporte ao reuso do córpus de exemplos, (ii) o sistema de crítica e (iii) o sistema de detecção automática da estrutura esquemática (classificador).

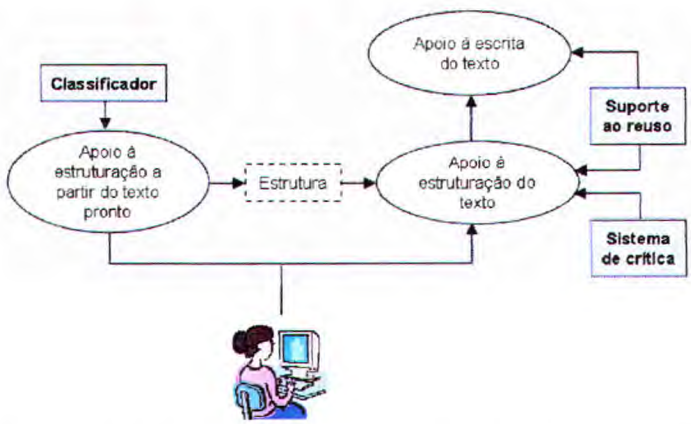


Figura 6.1: Relação entre as três macro-funções do ambiente *SciPo*

Nas próximas seções serão detalhadas as três macro-funções do ambiente *SciPo*, incluindo os sistemas envolvidos em cada uma delas.

Capítulo

6

Proposta de um Ambiente Computacional de Auxílio à Escrita Acadêmica

Neste capítulo apresentamos a proposta de um ambiente computacional de auxílio à escrita acadêmica, denominado *SciPo– SCIentific POrtuguese*, que implementa um modelo baseado em córpus e em sistemas de crítica, visando auxiliar, especialmente escritores iniciantes, na tarefa de estruturação e escrita de textos acadêmicos. Na versão atual, o *SciPo* trata as seções Resumo e Introdução. O uso da abordagem baseada em córpus foi herdada do projeto AMADEUS e parte do princípio de que os textos acadêmicos são passíveis de reuso, seja do ponto de vista da sua estrutura retórica ou do seu material lingüístico (expressões padronizadas). Quanto à abordagem de crítica utilizada, foi adotado o modelo proposto por Fischer *et al.* (1991) e também utilizado por Silva, Pelizzoni, & Aluísio (1998), em que dois agentes (um usuário e um crítico artificial) interagem em colaboração em um processo cíclico, já ilustrado na Figura 3.7 (Subseção 3.7.3, Cap. 3).

Na proposta desta nova abordagem, foram introduzidas duas inovações: (i) o uso integrado dos conceitos de reutilização de estruturas e material lingüístico e de sistemas de crítica em um único ambiente Web, de forma que tanto a estruturação quanto a escrita do texto possam ser apoiadas pelo ambiente computacional; e (ii) o uso de um sistema de detecção automática da estrutura esquemática como parte do processo de crítica, permitindo assim que o usuário inicie a interação com o ambiente a partir de um texto já produzido.

6.1 Apoio à Estruturação do Texto

O apoio à estruturação do texto se dá por meio de um conjunto específico de funções e recursos, cujo objetivo é auxiliar o usuário a compor estruturas adequadas para cada seção de um texto acadêmico, com base em modelos pré-estabelecidos e no uso de exemplos autênticos de cada seção. Na versão atual do ambiente *SciPo*, essa macro-função está implementada para as seções Resumo e Introdução. Uma vez que o usuário opte por compor uma estrutura de Resumo ou Introdução utilizando o *SciPo*, as seguintes funções de suporte ao reuso de córpus e de crítica podem ser acessadas:

a. navegação das bases de exemplos;
b. pesquisa das ocorrências de determinado componente ou estratégia retórica;
c. apoio à composição de estruturas esquemáticas;
d. crítica da estrutura construída;
e. recuperação dos exemplos com estruturas similares à estrutura construída.

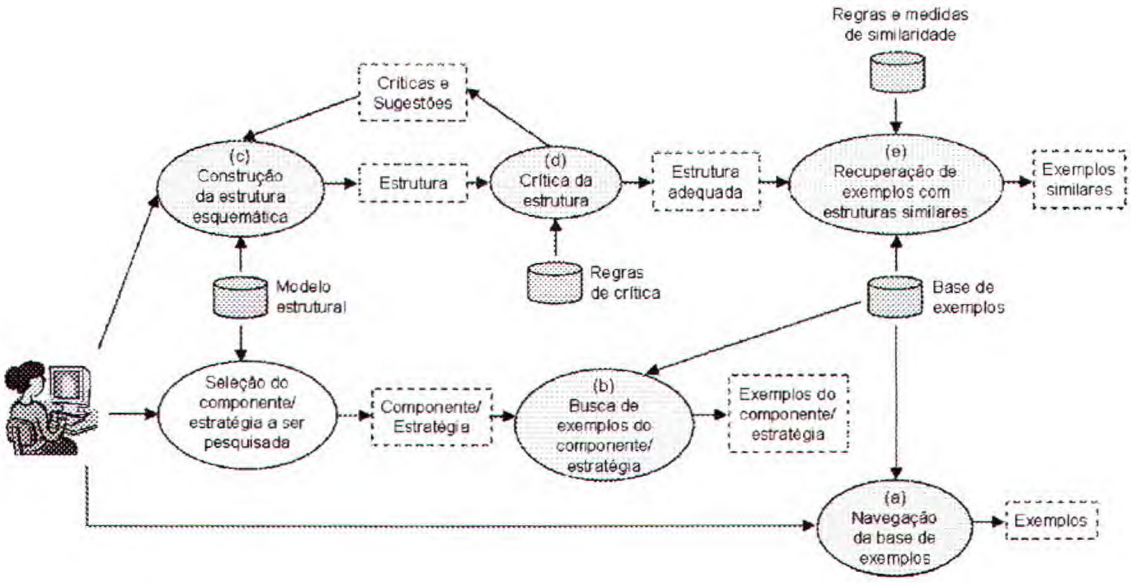


Figura 6.2: Visão geral do processo de apoio à estruturação do texto

A Figura 6.2 mostra como o usuário pode interagir com essas funções para a criação da estrutura esquemática desejada, bem como as fontes de conhecimento empregadas em cada processo. Essas funções, assim como as fontes de conhecimento, foram adaptadas do ambiente AMADEUS. Embora não tenha sido explicitado na figura, o primeiro passo da interação do usuário com o ambiente é a escolha da seção a ser tratada, neste caso, Resumo ou Introdução. Tendo escolhido a seção desejada, o usuário pode: (a) navegar

na Base de exemplos, visualizando todos os exemplos anotados de acordo com suas respectivas estruturas esquemáticas, podendo também reutilizar as estruturas, se desejado; (b) selecionar um determinado componente ou estratégia retórica do Modelo estrutural, para que sejam recuperadas todas as ocorrências desse componente/estratégia na Base de exemplos; ou (c) compor uma estrutura esquemática, com base no Modelo estrutural, que será submetida ao (d) processo de crítica da estrutura e, posteriormente, alimentará a (e) recuperação de exemplos similares. No processo de crítica, a estrutura construída pelo usuário é analisada com base nas Regras de crítica e, então, são retornadas ao usuário as críticas e sugestões cabíveis. O usuário poderá modificar sua estrutura de acordo com as sugestões do sistema e submetê-la novamente à crítica. Esse ciclo se repete até que seja alcançada uma estrutura satisfatória, isto é, até que nenhuma regra de crítica seja acionada. A recuperação de exemplos similares usa como entrada a estrutura construída pelo usuário (requisição), as estruturas esquemáticas da Base de exemplos e as Regras e medidas de similaridade, para devolver ao usuário os exemplos mais próximos a sua requisição. Vale lembrar que toda vez que um exemplo da base é mostrado ao usuário, sua estrutura esquemática, expressões padrão (trechos considerados reutilizáveis) e marcadores discursivos são apresentados de forma destacada, conforme exemplificado na Figura 6.3.

Resumo r_es11

Propósito: *Indicar o propósito principal*

Um estudo empírico visando avaliar a eficácia em revelar erros, a dificuldade de satisfação e o custo de aplicação do critério Análise de Mutantes **é apresentado neste trabalho**.

Propósito: *Introduzir mais propósitos*

Eficácia e custo *também* **foram avaliados para** os critérios Potenciais-Usos, **comparando-se** *assim* **os resultados obtidos** para esses critérios. **A especificação e implementação** de uma estratégia de minimização de conjuntos de casos de teste adequados ao critério Análise de Mutantes *também* **é apresentada**.

Resultado: *Apresentar resultados*

Através dos resultados obtidos observou-se que os critérios Potenciais-Usos (baseado em fluxo dados) e o critério Análise de Mutantes (baseado em erros) **são promissores** e apresentam características complementares **que merecem ser investigadas** em um experimento de maior porte.

Conclusao: *Apresentar conclusões*

A utilização de mutação restrita e de minimização de conjunto de casos de teste **constituem mecanismos que viabilizam** a aplicação desses critérios em ambientes de produção de software.

Figura 6.3: Apresentação de um exemplo de Resumo

Ainda sobre a Figura 6.2, vale notar que as fontes de conhecimento estão representadas de forma genérica, embora o ambiente utilize conhecimentos específicos para o tratamento de cada seção — Resumo ou Introdução. Esses conhecimentos são acessados pelo ambiente de acordo com a seção que está sendo tratada. São quatro as fontes de conhecimento utilizadas:

i. um modelo estrutural;
ii. uma base de exemplos;
iii. um conjunto de regras de crítica;
iv. um conjunto de regras e medidas de similaridade.

Dessa forma, quando tratando de Resumos, o ambiente utiliza o Modelo estrutural de Resumos (apresentado previamente na Figura 5.8, Cap. 5), a Base de exemplos de Resumos (composta pelos 52 resumos manualmente anotados de acordo com os quatro níveis de análise também discutidos na Seção 5.2, Cap. 5), as Regras de crítica de Resumos e as Regras e medidas de similaridade (que se aplicam a ambas as seções). De maneira análoga, quando tratando de Introduções, o ambiente utiliza o Modelo estrutural de Introduções (Figura 5.9, Cap. 5), a Base de exemplos de Introduções (52 introduções manualmente anotadas), as Regras de crítica de Introduções e as Regras e medidas de similaridade.

Para que a macro-função de apoio à estruturação do texto seja compreendida mais claramente, serão descritas nas seções seguintes as fontes de conhecimento citadas, com exceção dos modelos estruturais de Resumo e Introdução, uma vez que já foram abordados no capítulo anterior.

6.1.1 As Bases de Exemplos

Tanto a Base de exemplos de Resumos quanto a de Introduções foram obtidas a partir do processo de anotação manual do córpus, apresentado na Subseção 5.2.4, Cap. 5. A Base de Resumos contém 52 instâncias de estruturas esquemáticas de resumos autênticos. A estrutura de cada resumo é representada por meio de um predicado Prolog. Cada predicado é associado ao arquivo XML correspondente, contendo o texto do resumo anotado com suas características retóricas (componentes, estratégias, expressões padrão e marcadores discursivos) e, ainda, observações críticas sobre sua estrutura. As observações críticas foram geradas aplicando-se aos exemplos as mesmas regras de crítica utilizadas no processo de crítica da estrutura citado anteriormente. Dessa forma, a Base de exemplos é composta pelos exemplos propriamente ditos e pelo conjunto dos predicados correspondentes. A Base de Introduções contém 48 instâncias de estruturas esquemáticas de introduções autênticas, representadas de forma similar à descrita para os Resumos. Quatro introduções do córpus original (que contém 52 introduções) não foram incluídas na Base de

exemplos, por terem sido consideradas excessivamente problemáticas.

A Figura 6.4 apresenta um exemplo de Resumo codificado em XML. Neste exemplo é possível observar as anotações dos componentes (etiqueta <Componente>), das estratégias retóricas (etiqueta <Estrategia>), das expressões padrão (etiqueta <Reutilizavel>) e dos marcadores discursivos (etiqueta <Marcador>). O Schema XML que define todas as etiquetas utilizadas é apresentado no Apêndice B.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<Resumo id="ES11" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="D:/Resumos/schema_resumo.xsd">
    <Componente><Nome>Propósito</Nome>
        <Estrategia><Nome>Indicar o propósito principal</Nome>
            <Reutilizavel>Um estudo empírico visando avaliar</Reutilizavel> a eficácia em revelar erros, a dificuldade
            de satisfação e o custo de aplicação do critério Análise de Mutantes<Reutilizavel> é apresentado neste
            trabalho</Reutilizavel>.
        </Estrategia>
        <Estrategia><Nome>Introduzir mais propósitos</Nome>
            Eficácia e custo <Reutilizavel><Marcador>também</Marcador> foram avaliados para</Reutilizavel> os
            critérios Potenciais-Usos, <Reutilizavel>comparando-se <Marcador>assim</Marcador> os resultados
            obtidos</Reutilizavel> para esses critérios. <Reutilizavel>A especificação e implementação</Reutilizavel>
            de uma estratégia de minimização de conjuntos de casos de teste adequados ao critério Análise de Mutantes
            <Reutilizavel><Marcador>também</Marcador> é apresentada</Reutilizavel>.
        </Estrategia>
    </Componente>
    <Componente><Nome>Resultado</Nome>
        <Estrategia><Nome>Apresentar resultados</Nome>
            <Reutilizavel>Através dos resultados obtidos observou-se que</Reutilizavel> os critérios Potenciais-Usos
            (baseado em fluxo dados) e o critério Análise de Mutantes (baseado em erros) <Reutilizavel>são promis-
            sores</Reutilizavel> e apresentam características complementares <Reutilizavel>que merecem ser
            investigadas</Reutilizavel> em um experimento de maior porte.
        </Estrategia>
    </Componente>
    <Componente><Nome>Conclusão</Nome>
        <Estrategia><Nome>Apresentar conclusões</Nome>
            <Reutilizavel>A utilização de</Reutilizavel> mutação restrita e de minimização de conjunto de casos de teste
            <Reutilizavel>constituem mecanismos que viabilizam</Reutilizavel> a aplicação desses critérios em ambientes
            de produção de software.
        </Estrategia>
    </Componente>
</Resumo>
```

Figura 6.4: Exemplo de resumo codificado em XML

A representação da estrutura esquemática dos exemplos como predicados Prolog possui a seguinte forma:

case(Nome_Caso, Lista_de_componentes).

O parâmetro **Nome_caso** armazena o identificador do exemplo (nome do arquivo) representado pela estrutura **case**. A **Lista_de_componentes** segue a forma **c(Nome_componente, Estratégia)** e **Estratégia** tem a forma **s(Nome_estratégia, Lista_de_mensagens)**. O parâmetro **Lista_de_mensagens**, proveniente do projeto AMADEUS, não é utilizado pelo *SciPo*, de modo que foi substituído pela variável anônima do Prolog (_). A Figura 6.5 apresenta o predicado Prolog correspondente ao Resumo apresentado na Figura 6.4.

O conjunto de predicados que compõem cada uma das bases é utilizado como um índice de acesso aos exemplos, permitindo (a) a navegação pela base de exemplos e (b) a pes-

```
case(r_es11,
[c(proposito,s(indicar_o_proposito_principal,_)),
c(proposito,s(introduzir_mais_propositos,_)),
c(resultado,s(apresentar_resultados,_)),
c(conclusao,s(apresentar_conclusoes,_))]).
```

Figura 6.5: Exemplo de estrutura codificada em Prolog

quisa de exemplos de determinados componentes/estratégias. Além disso, o conjunto de predicados também é a base utilizada para (d) a recuperação de exemplos estruturalmente similares à estrutura requisitada pelo usuário.

6.1.2 As Regras e o Processo de Crítica

Tanto as Regras de crítica de Resumos quanto as Regras de crítica de Introduções foram formuladas com base nos modelos estruturais utilizados pelo ambiente *SciPo* e tentam corrigir padrões de estruturação considerados problemáticos. Foram consideradas sugestões prescritivas de boas estruturas de Resumos e Introduções encontradas na literatura (discutidas no Capítulo 2) e também os principais desvios observados no córpus, com relação aos modelos estruturais, que possam prejudicar de alguma forma o entendimento do texto. As regras consideram desvios de conteúdo (falta de componentes/estratégias tidos como essenciais) e desvios de ordem (ordem de ocorrência dos componentes/estratégias na estrutura), além de diferenciar desvios graves (apresentados ao usuário como "críticas") de desvios leves (apresentados ao usuário como "sugestões"). Dessa forma, são quatro as classes de regras:

i. críticas de conteúdo, que tratam da falta de componentes/estratégias essenciais;
ii. críticas de ordem, que tratam de problemas estruturais graves com relação à ordem em que os componentes/estratégias aparecem na estrutura;
iii. sugestões de conteúdo, que tratam da falta de componentes/estratégias tidos como opcionais; e
iv. sugestões de ordem, que tratam de problemas estruturais leves com relação à ordem em que os componentes/estratégias aparecem na estrutura;

Embora os modelos de Resumos e Introduções utilizados pelo ambiente *SciPo* reflitam os elementos estruturais esperados para essas seções, não é necessário que todos os seus componentes estejam presentes, nem que a ordem desses componentes seja rigorosamente obedecida, para que uma estrutura seja considerada boa. Dessa forma, as Regras de críticas consideram alguns componentes como obrigatórios enquanto outros são tratados como opcionais. No caso dos Resumos, os componentes Propósito, Metodologia e Resultado são considerados obrigatórios e os componentes Contexto, Lacuna e Conclusão, opcionais.

No caso das Introduções, são considerados obrigatórios os componentes **Contexto**, **Revisão da Literatura**, **Lacuna**, **Propósito**, **Metodologia** e **Resultado** enquanto os componentes **Contribuições/Valor do trabalho** e **Estrutura do trabalho** são considerados opcionais.

O processo de crítica estrutural do *SciPo* é ilustrado na Figura 6.6. Nessa figura estão representadas tanto as Regras de crítica aplicáveis a Resumos quanto as aplicáveis a Introduções. Vale lembrar que esse processo é cíclico e só será finalizado quando nenhuma regra das classes "críticas" for violada.

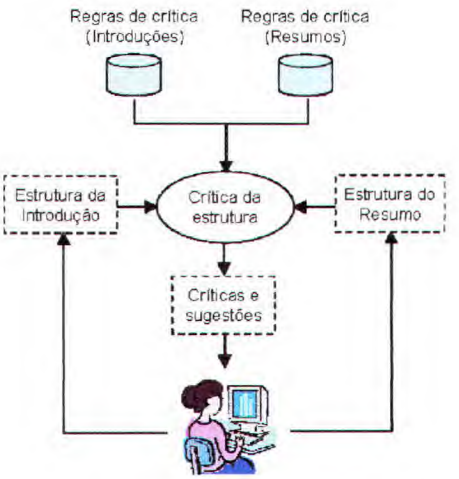


Figura 6.6: Processo de crítica implementado no ambiente *SciPo*

A Figura 6.7 exemplifica as críticas e sugestões emitidas pelo sistema de crítica considerando-se a estrutura de Resumo apresentada na Figura 6.5. Vale notar que juntamente com a crítica/sugestão, é dada uma justificativa, informando o usuário sobre a relevância da crítica/sugestão. O conjunto completo das Regras de crítica (para Resumos e para Introduções), incluindo as mensagens de justificativa, é apresentado no Apêndice C.

6.1.3 As Regras e Medidas de Similaridade e a Recuperação de Exemplos Similares

As Regras e Medidas de Similaridade são baseadas em regras de similaridade entre listas (casamento de padrões) e na medida de similaridade conceitual *nearest neighbors matching* (Kriegsman & Barletta, 1993). Essas regras são utilizadas no processo de recuperação de exemplos similares, que é baseado na abordagem RBC (Raciocínio Baseado em Casos).

Na abordagem RBC, inspirada no reuso de experiências empregado pelos seres humanos, o conhecimento é especificado como um conjunto de problemas previamente resolvidos (casos) em um dado domínio. O método de resolução de problemas consiste em recuperar

Críticas e Sugestões

Crítica: Faltam componentes essenciais

Falta *Metodologia*

O resumo deve indicar objetivamente ao leitor a metodologia empregada para a realização do seu trabalho. Acrescente pelo menos uma das 3 estratégias de metodologia. Escolha a que for mais adequada ao seu resumo.

- *Mostrar exemplos do componente Metodologia*

Sugestão: Tente enriquecer sua estrutura

Adicionando *Contexto* e *Lacuna*

Que tal acrescentar *Contexto* e *Lacuna*? As estratégias de contextualização ajudam o leitor a entender melhor o escopo do seu trabalho. Com essas estratégias você pode convencer o leitor de que sua área de pesquisa é relevante e merece ser explorada. Vale a pena acrescentar alguma estratégia de *Contexto*! A contextualização também prepara o leitor para compreender melhor o problema que você pretende atacar em sua pesquisa. Esse problema é enunciado através das estratégias de *Lacuna*. A declaração de uma lacuna em sua área de pesquisa complementa a contextualização e motiva o propósito do seu trabalho. Escolha as estratégias mais adequadas ao seu resumo.

- *Mostrar exemplos do componente Contexto*
- *Mostrar exemplos do componente Lacuna*

Figura 6.7: Resultado da aplicação das Regras de Crítica de Resumo à estrutura apresentada na Figura 6.5

o caso que resolve um problema similar ao problema atual do usuário e adaptá-lo à nova situação. O caso modificado, que atende agora à nova situação, pode então ser guardado para uso futuro. Sistemas que empregam RBC podem ser puros, isto é, baseados somente em casos e em nenhum outro tipo de conhecimento, ou podem integrar RBC com outras formas de raciocínio, como o baseado em regras. No *SciPo*, é empregado o tipo mais simples de raciocínio baseado em casos, isto é, um sistema de recuperação que mantém uma base de casos, podendo recuperar aqueles que satisfazem total ou parcialmente uma dada requisição. Dessa maneira, no processo de recuperação de exemplos similares, é feita uma busca na base de exemplos (Subseção 6.1.1) pelas estruturas mais próximas à requisição do usuário.

A entrada para esse processo de busca é a estrutura esquemática especificada pelo usuário, convertida em uma lista Prolog similar à apresentada na Figura 6.5 (chamada de "requisição"), que será comparada às estruturas da Base de exemplos (chamadas de "casos"), utilizando três tipos de comparações entre listas: igualdade, inclusão e intersecção

não-nula. A partir desses tipos de comparações, foram derivadas buscas que tratam das seguintes situações:

1. **Requisição totalmente contida no caso**: são recuperados os casos que contêm totalmente a requisição, considerando a ordem dos elementos da lista. Por exemplo, considerando a requisição [a,b,c,d,e], ela não está contida no caso [a,1,b,c,2,d,e], mas sim no caso [1,2,a,b,c,d,e]. O valor de similaridade, um número entre 0 e 1, é dado pela divisão do tamanho da requisição pelo tamanho do caso. Os casos iguais à requisição são desconsiderados nessa situação, pois são tratados pela comparação de igualdade.
2. **Requisição parcialmente contida no caso**: são consideradas as intersecções não-nulas entre caso e requisição, desde que o tamanho da requisição seja menor ou igual ao tamanho do caso. Casos idênticos à requisição, ou que contêm toda a requisição, também são desconsiderados. Vale ressaltar que a intersecção retornada, por exemplo, entre as listas [h,i,j,l,m,n,o] e [i,h,j,l,m,o,n], é a lista [j,l,m]. O valor de similaridade é dado pela divisão do tamanho da intersecção pela soma entre o tamanho da requisição, o tamanho do caso e o valor negativo do tamanho da intersecção.
3. **Caso totalmente contido na requisição**: similar à situação 1, são recuperados os casos que estão totalmente contidos na requisição, considerando a ordem dos elementos da lista. O valor de similaridade é dado pela divisão do tamanho do caso pelo tamanho da requisição.
4. **Caso parcialmente contido na requisição**: análoga à situação 2, com a ressalva de que o tamanho do caso deve ser menor que o tamanho da requisição. O valor de similaridade é computado da mesma maneira que em 2.
5. **Requisição totalmente contida no caso, sem considerar a ordem dos componentes**: similar a situação 1, porém desconsiderando a ordem dos elementos;
6. **Requisição parcialmente contida no caso sem considerar a ordem dos componentes**: similar a situação 2, porém desconsiderando a ordem dos elementos;
7. **Caso totalmente contido na requisição, sem considerar a ordem dos componentes**: similar a situação 3, porém desconsiderando a ordem dos elementos;
8. **Caso parcialmente contido na requisição sem considerar a ordem dos componentes**: similar a situação 4, porém desconsiderando a ordem dos elementos;
9. **Requisição totalmente contida no caso, permitindo a repetição de componentes**: esse é um caso especial da busca 1. A ordem da requisição é considerada, porém é permitido que componentes semelhantes se repitam dentro da mesma vizinhança. Em outras palavras, um determinado caso é recuperado nessa situação se ele puder ser construído a partir da requisição, da seguinte maneira: para cada componente da requisição, insere-se qualquer número de componentes, desde que

esses componentes sejam semelhantes.

Para que o processo de recuperação de exemplos similares fosse simplificado do ponto de vista do usuário, as nove buscas apresentadas acima, mais dois testes de igualdade entre listas, foram agrupados em quatro níveis de recuperação, a saber:

- **Nível 1**: Casos que contêm toda a requisição, incluindo a situação de requisição igual ao caso, levando em consideração a ordem dos componentes/estratégias (situações 1 e 9);
- **Nível 2**: Casos que contêm parte da requisição, levando em consideração a ordem dos componentes/estratégias (situações 2, 3 e 4);
- **Nível 3**: Casos que contêm toda a requisição, incluindo a situação de requisição igual ao caso, desconsiderando a ordem dos componentes/estratégias (situação 5);
- **Nível 4**: Casos que contêm parte da requisição, desconsiderando a ordem dos componentes/estratégias (situações 6, 7 e 8).

Os resultados obtidos em cada nível são mutuamente exclusivos, ou seja, um caso sempre será apresentado em apenas um dos níveis, considerando-se a prioridade entre os níveis. O grau de prioridade vai do 1 (maior prioridade) ao 4 (menor prioridade).

A Figura 6.8 exemplifica o processo de recuperação de exemplos similares. A estrutura pesquisada (requisição) é mostrada na parte superior da figura e a seguir é apresentado o resultado, considerando os quatro níveis de busca. As porcentagens exibidas ao lado do identificador dos exemplos indica o grau de similaridade do caso em relação à requisição.

Na Figura 6.8, a estrutura esquemática é composta por pares do tipo componente::estratégia. Vale ressaltar que a recuperação de exemplos similares também pode ser feita por meio de uma estrutura composta por pares do tipo componente::*, isto é, sem especificar a estratégia do componente. Esses pares com estratégias genéricas são tratados de forma semelhante aos pares em que tanto o componente quanto a estratégia estão especificados. A única diferença é que, quando utilizados pares do tipo componente::*, as estratégias são desconsideradas, tornando a busca menos específica.

6.2 Apoio à Estruturação a partir do Texto Pronto

O apoio à estruturação a partir do texto pronto se dá de forma semelhante à descrita na seção anterior. A diferença está no ponto de partida da interação do usuário com o ambiente. Essa macro-função permite que o usuário inicie a interação submetendo à análise estrutural um texto já redigido. Isso é possível devido ao uso de um sistema capaz de detectar automaticamente a estrutura esquemática do texto. Conforme discutido no

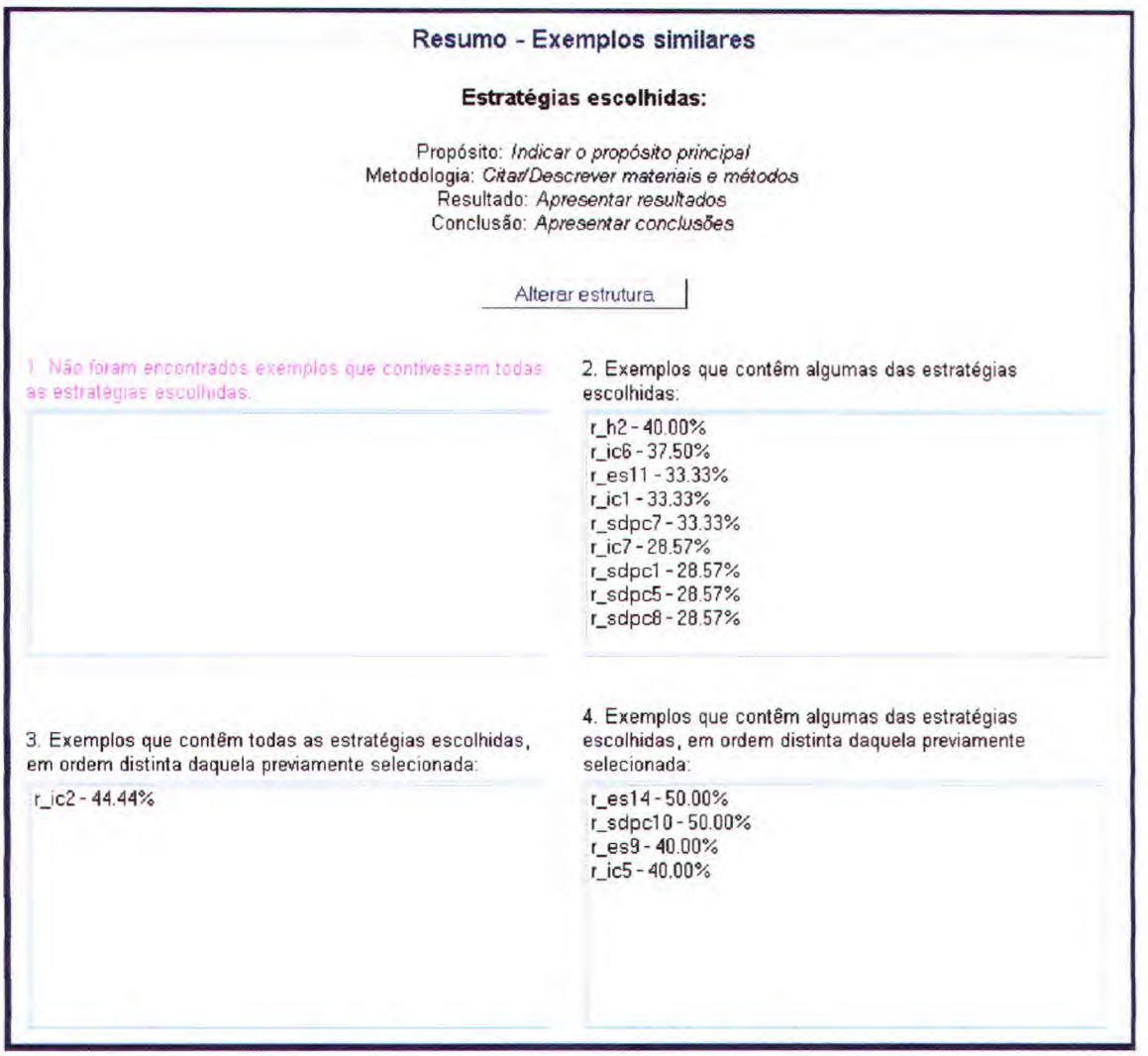


Figura 6.8: Exemplo de resultado da recuperação de exemplos similares

Capítulo 4, esse sistema pode ser implementado por meio de um classificador textual. A Figura 6.9 demonstra como esse sistema se integra no processo de crítica já discutido.

6.2.1 Detecção Automática da Estrutura Esquemática de Textos Acadêmicos em Português

Para testar a viabilidade dessa nova abordagem de crítica, foi proposto um sistema de detecção automática da estrutura esquemática de resumos acadêmicos em português, chamado *AZPort— Argumentative Zoning for Portuguese*. A seção Resumo foi escolhida por constituir um estudo de caso suficiente para a avaliação inicial da viabilidade do uso desse tipo de classificador em um ambiente como *SciPo*, além do fato de o Resumo conter

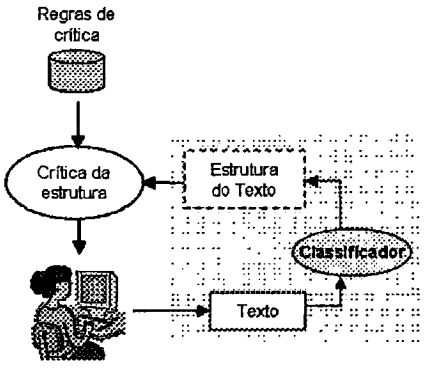


Figura 6.9: Processo de apoio à estruturação a partir do texto redigido

componentes semelhantes aos da Introdução, possibilitando, dessa forma, que a extensão do classificador para introduções possa ocorrer de forma escalável. Conforme já comentado no Capítulo 4, o *AZPort* segue a abordagem apresentada por Teufel & Moens (2002), que propõem a segmentação automática de um texto acadêmico em "zonas argumentativas".

A seguir são apresentadas as várias etapas envolvidas na construção do *AZPort*, iniciando pela criação do córpus usado como material de treinamento.

6.2.2 Material de Treinamento

Uma vez que o método de aprendizado empregado no *AZPort* é supervisionado, foi necessário criar um conjunto de exemplos que servisse de material de treinamento. Decidimos por utilizar os resumos do CorpusDT, por já estarem anotados com as categorias retóricas de nosso interesse. Entretanto, para que o material de treinamento fosse confiável, era necessário comprovar a reprodutibilidade da anotação, isto é, se a anotação poderia ser reproduzida de forma semelhante por diferentes anotadores humanos. O modelo de componentes e estratégias utilizado no processo de anotação manual dos Resumos, descrito no Capítulo 5, não era suficientemente reprodutível, pois gerava muitas discordâncias, principalmente no nível das estratégias. De fato, quanto mais refinado é o modelo de anotação, mais difícil é conseguir que diferentes anotadores concordem entre si. Dessa forma, para aumentar a reprodutibilidade do modelo (apresentado na Figura 5.8, Cap. 5), foram excluídas as estratégias, resultando em um esquema de anotação similar ao utilizado por Teufel, Carletta, & Moens (1999). Esse esquema é apresentado na Figura 6.2.2.

Considerando esse novo esquema, os resumos foram submetidos a um novo processo de anotação. Para medir a reprodutibilidade desse esquema, foi feito um experimento com três anotadores humanos que já tinham conhecimento do domínio do córpus e também sobre escrita acadêmica. Para medir a estabilidade do esquema, isto é, quão semelhante será o resultado da anotação feita por um mesmo anotador em diferentes intervalos de tempo,

CONTEXTO (C)
LACUNA (L)
PROPÓSITO (P)
METODOLOGIA (M)
RESULTADO (R)
CONCLUSÃO (Co)
ESTRUTURA (E)

Figura 6.10: Esquema de anotação utilizado no AZPort

foi repetido o experimento de anotação com um dos anotadores, usando um intervalo de tempo (entre um experimento e outro) de três meses.

Durante o processo de anotação, uma das dificuldades encontradas pelos anotadores foi o alto número de sentenças com papéis argumentativos sobrepostos, o que levou a dúvidas sobre a categoria correta a ser associada à sentença. Anthony (1999) também relata esse tipo de conflito durante o processo anotação de um córpus de introduções de artigos da área de Engenharia de Software. Na tentativa de minimizar essa dificuldade, foi escrito um documento contendo *guidelines* que especificam estratégias para lidar com os principais tipos de conflitos que os anotadores poderiam encontrar, tal como PROPÓSITO vs. RESULTADO. Esse documento foi inspirado nas guidelines do AZ original (Teufel, 1999) e é apresentado no Apêndice F.

Ambos os experimentos foram realizados com os 52 resumos do CorpusDT, contendo 366 sentenças (10.936 palavras). Para o experimento de reprodutibilidade, foram usados seis resumos na fase de treinamento, que foi realizada em três rodadas, cada uma consistindo de etapas de explicação, anotação e discussão. Após a fase de treinamento, foi pedido aos anotadores que anotassem 46 resumos, sentença a sentença, classificando cada sentença com uma única categoria. Como dito anteriormente, para o experimento de estabilidade, foi pedido a um dos anotadores participantes do experimento anterior que repetisse a anotação dos 46 resumos tendo se passado três meses desde a primeira anotação.

Foi utilizada a estatística *Kappa* (K) para medir a reprodutibilidade entre k anotadores sobre N itens e a estabilidade para um anotador. Para esses experimentos, itens são sentenças e o número de categorias é igual a sete. Os resultados mostraram que o esquema é reproduzível (K=0.69, N=320, k=3), considerando-se a subjetividade desse tipo de tarefa e as recomendações da literatura com relação a um valor aceitável do *Kappa*. Em um experimento similar, Teufel, Carletta, & Moens (1999) relatou a reprodutibilidade do seu esquema de anotação como sendo ligeiramente maior (K=0.71, N=4261, k=3), conforme comentado no Capítulo 4. Se agruparmos as categorias METODOLOGIA, RESULTADO e CONCLUSÃO em uma única categoria mais genérica (similar à categoria OWN de Teufel *et al.*), o valor do *Kappa* para esse experimento aumenta consideravelmente (K=0.82,

N=320, k=3). Esse resultado é um forte indicativo de que os anotadores têm problemas em distinguir essas três categorias. O esquema também se mostrou estável, uma vez que o mesmo anotador reproduziu uma anotação muito similar em tempos diferentes (K=0.79, N=320, k=2).

Tais resultados permitem concluir que anotadores humanos treinados conseguem distinguir as categorias propostas no esquema de anotação e, dessa forma, a anotação resultante de tais experimentos é confiável o suficiente para ser usada como material de treinamento para o classificador automático.

6.2.3 Descrição dos Atributos Utilizados

Assim como o AZ, o *AZPort* é um classificador bayesiano que atribui a cada sentença de entrada um conjunto de possíveis categorias retóricas com suas respectivas probabilidades estimadas. Assim como outros algoritmos de aprendizado de máquina, em vez de lidar diretamente com o objeto a ser classificado, o *AZPort* recebe as sentenças como vetores de atributos. Dessa forma, a extração de atributos é um passo crucial em tal cenário.

O ponto de partida para a seleção dos atributos que seriam utilizados pelos *AZPort* foi o conjunto de 16 atributos usados por Teufel & Moens (2002) (Tabela 4.4, Cap. 4). Com base nos desempenhos desses atributos e considerando que o objetivo do *AZPort* é classificar sentenças de resumos, foram derivados oito atributos, resumidos na Tabela 6.1.

Tabela 6.1: Resumo do conjunto de atributos

Atributo	Descrição	Valores possíveis
1. Tamanho	Qual é o tamanho da sentença em comparação aos dois limiares (20 e 40 palavras)?	curta, média ou longa
2. Localização	Qual a posição da sentença no resumo?	primeira, segunda, mediana, penúltima ou última
3. Citação	A sentença contém citações?	sim ou não
4. Expressão	Que tipo de expressão padrão a sentença contém?	C, L, P, M, R, Co ou *noexpr*
5. Tempo	Qual o tempo do primeiro verbo finito da sentença?	IMP, PRES, PAST, FUT, PRES-CPO, PAST-CPO, FUT-CPO, PRES-CT, PAST-CT, FUT-CT, PRES-CPO-CT, PAST-CPO-CT, FUT-CPO-CT ou *noverb*
6. Voz	Qual a voz do primeiro verbo finito da sentença?	passiva, ativa ou *noverb*
7. Modal	O primeiro verbo finito da sentença é modal?	sim, não ou *noverb*
8. Histórico	Qual é a categoria da sentença anterior?	_, C, L, P, M, R, Co ou E

Tamanho

O atributo **Tamanho** classifica uma sentença como *curta, média* ou *longa*, com base no número de palavras da sentença. São utilizados dois limiares para a decisão do valor do atributo – 20 e 40 palavras. Esses limiares foram estimados usando-se a média dos tamanhos de sentenças presentes no córpus.

Localização

O atributo **Localização** identifica a posição ocupada pela sentença no resumo. Quebras de parágrafo não foram consideradas e, dessa forma, todo resumo é processado como um parágrafo único. São utilizados cinco valores para este atributo: *primeira, segunda, mediana, penúltima* e *última*. Foram realizados vários experimentos para a determinação dos melhores valores para **Localização** e os resultados demonstraram que os cinco valores adotados caracterizam localizações comuns para algumas das categorias do esquema utilizado.

Citação

O atributo **Citação** identifica a presença de citações na sentença. Uma vez que apenas o resumo é processado e, dessa forma, a lista de referências bibliográficas não está disponível, não é possível identificar o tipo de citação (citação para outros autores ou para trabalhos anteriores do mesmo autor) como é feito no AZ original.

Expressão

O atributo **Expressão** identifica a presença de uma expressão padrão na sentença. O método utilizado para encontrar tais expressões é o casamento de padrões, com base em um conjunto fixo de expressões que podem ocorrer em textos que seguem o modelo de estruturação apresentado na Figura 6.2.2. Dessa forma, nosso conjunto de expressões é dividido de acordo com as categorias previstas nesse modelo. A Figura 6.11 apresenta exemplos dos tipos de expressões representadas no atributo **Expressão**.

Tempo, Voz e Modal

Os atributos **Tempo, Voz** e **Modal**, chamados de atributos sintáticos, descrevem propriedades sintáticas do primeiro verbo finito[1] da sentença, no modo indicativo ou imperativo. Devido à alta probabilidade de verbos no modo subjuntivo pertencerem a

[1]Conforme já definido no Capítulo 4, diz-se *finita* qualquer forma verbal dos modos indicativo, imperativo ou subjuntivo.

Categoria	Expressão
CONTEXTO	A partir do ano...
LACUNA	Contudo, é necessário...
PROPÓSITO	Esta tese apresenta...
METODOLOGIA	Foi utilizado o modelo...
RESULTADO	Os resultados mostram...
CONCLUSÃO	Concluímos que...
ESTRUTURA	Na seção seguinte...

Figura 6.11: Exemplos de expressões padrão

orações subordinadas, eles são considerados apenas quando nenhum outro verbo finito no indicativo ou imperativo é encontrado. Nos caso em que nenhum verbo finito é encontrado na sentença, os três atributos sintáticos assumem o valor *noverb*. É importante destacar que, na determinação dos atributos sintáticos, são considerados tanto os verbos simples (p.e. *"Os resultados demonstram..."*) quanto as locuções verbais, chamadas aqui de verbos complexos, que incluem um ou mais verbos auxiliares para expressar (i) aspecto contínuo (*estar* + gerúndio) ou perfeito (*ter* + particípio), como em *"Grande esforço* **tem** *sido feito para..."*; (ii) voz passiva (*ser* + particípio), como em *"tem* **sido** *feito"*; (iii) modalização (*dever/poder/precisar/ter que|de/etc.* + infinitivo). O verbo complexo pode ainda conter o pronome *se* como índice de indeterminação do sujeito ou partícula apassivadora. Neste trabalho, será utilizado o termo "complexo verbal" para designar verbos em geral, sejam eles simples ou complexos.

O atributo **Tempo** indica a flexão do verbo (simples ou complexo) e pode assumir 14 valores, incluindo o valor *noverb*. Originalmente, esse atributo incluía todos os tempos simples possíveis no português, mas como muitos desses tempos são raros ou mesmo inexistentes no córpus de resumos, alguns tempos foram agrupados. Dessa forma, é utilizado um único valor de passado e futuro em detrimento dos três tempos de passado e dois tempos de futuro possíveis, e também foi neutralizado o modo, resultando nos 14 valores utilizados (veja Tabela 6.1). Os valores do atributo **Tempo** têm a forma "TempoSimples(-CPO)(-CT)", sendo que "CPO" sinaliza a presença do auxiliar *ter* no complexo verbal, expressando o aspecto perfeito, e "CT" sinaliza a presença do auxiliar *estar* expressando o aspecto contínuo.

O atributo **Voz** indica a voz do verbo e pode assumir os valores *ativa, passiva* ou *noverb*. A voz passiva é entendida aqui em um sentido mais amplo, abrangendo certas formas e construções verbais que são geralmente utilizadas para omitir um agente, a saber: (i) voz passiva analítica (verbo *ser* + particípio), (ii) voz passiva sintética (feita com a partícula *se*) e (iii) sujeito indeterminado indicado pela flexão de terceira pessoa do singular + *se*.

O atributo **Modal** indica se há um auxiliar modal no complexo verbal e pode assumir

os valores *sim, não* e *noverb*. São considerados como modais os seguintes verbos: *ter (que|de), dever, precisar* e poder.

Histórico

O atributo **Histórico** leva em conta a categoria da sentença anterior à sentença em classificação. Conforme dito anteriormente, sabe-se que algumas zonas argumentativas tendem a seguir outras zonas específicas. Essa propriedade é ainda mais evidente em textos autocontidos, como resumos. Em nosso córpus, algumas seqüências de categorias são muito freqüentes. Por exemplo, o padrão CONTEXTO (C) seguido de LACUNA (L), com repetição ou não, e então seguido de PROPÓSITO (P), isto é, ((CL)|(LC)+)P, ocorre em 30.7% do córpus.

Durante o treinamento, a determinação do atributo **Histórico** é feita por simples observação do córpus. Para sentenças novas, no entanto, o atributo **Histórico** tem que ser estimado como um segundo passo durante o processo de teste. Assim como em Teufel & Moens (2002), essa estimativa é feita por meio de uma *beam search* de profundidade três (Manning & Schütze, 2001) entre as categorias candidatas, de forma que se encontre a classificação mais provável.

6.2.4 Extração dos Valores dos Atributos

Os oito atributos descritos anteriormente são determinados automaticamente a partir do texto de entrada, por meio de um processo implementado em *Perl*. A Figura 6.12 apresenta as etapas do processo de determinação dos atributos, bem como os recursos empregados em cada etapa. Também são mostrados os atributos resultantes de cada etapa.

A seguir é detalhada cada uma das etapas do processo de determinação de atributos.

Tokenização e Delimitação de Sentenças

Para a realização das duas primeiras etapas do processo, **Tokenização** e **Delimitação de sentenças**, foi utilizada a ferramenta *Sentencer*, desenvolvida no NILC, por Jorge Marques Pelizzoni. Essa ferramenta utiliza informações contidas no léxico do NILC[2] juntamente com um conjunto de heurísticas para tokenizar e também para decidir sobre os limites das sentenças. Foi adicionado código *Perl* para que as sentenças fossem delimitadas por etiquetas XML contendo os respectivos identificadores. As etapas de tokenização e

[2]Disponível para consulta como a base de dados lexicais Diadorim (http://www.nilc.icmc.usp.br/nilc/tools/intermed.htm)

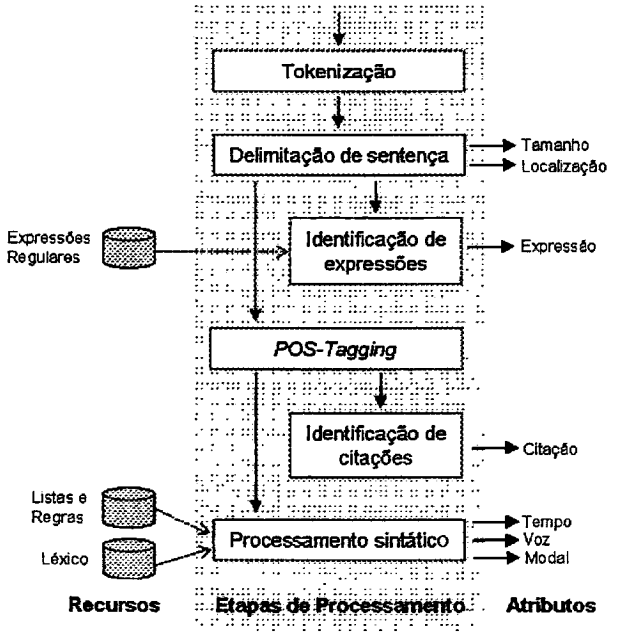


Figura 6.12: Etapas do processo de determinação dos atributos

delimitação de sentenças provêem as informações necessárias para o cálculo dos atributos **Tamanho** e **Localização**.

O principal problema dessa ferramenta é sua limitação no tratamento de sentenças que possuem texto entre parênteses, colchetes ou chaves. Caso o texto delimitado por esses sinais contenha pontuação de final de sentença, a ferramenta quebrará nesse ponto a sentença em que o texto se insere.

Para evitar esse problema, o texto é pré-processado antes de ser submetido ao *Sentencer*, retirando-se qualquer pontuação que possa causar a quebra de texto delimitado por parênteses, colchetes ou chaves.

Identificação de Expressões

Para fazer o reconhecimento das expressões padrão, foi construído um conjunto de 377 expressões regulares, divididas em seis categorias: CONTEXTO, LACUNA, PROPÓSITO, METODOLOGIA, RESULTADO e CONCLUSÃO. Definido o conjunto de expressões regulares, foi utilizado o `flex`[3] para a geração de *scanners* que fazem o reconhecimento das possíveis expressões no texto. O `flex` foi utilizado devido a sua facilidade de uso e eficiência no reconhecimento de padrões regulares, de ordem de complexidade de tempo linear ao tamanho do texto de entrada.

As fontes para a construção das expressões regulares foram observações feitas no córpus

[3]Ferramenta tradicional do Unix, mas também disponível para Windows.

e também expressões mencionadas na literatura como características do gênero acadêmico. A partir dessas observações, foi feito um processo manual de generalização das expressões para que cobrissem outras expressões semelhantes, mesmo que estas não estivessem presentes no córpus. Exemplos de expressões regulares (no formato `flex`) e de expressões geradas pelas respectivas expressões regulares são apresentados na Tabela 6.2. A parte superior da tabela (**Constantes**), descreve as constantes utilizadas na realização das expressões regulares apresentadas na parte inferior da tabela. O conjunto completo de constantes e expressões regulares, no formato `flex`, é apresentado no Apêndice D.

POS-Tagging

A etiquetação morfossintática, ou *POS-tagging*, é uma etapa fundamental no processo de determinação dos atributos, pois fornece informações tanto para o processamento sintático quanto para a identificação de citações. O *POS-tagging* é realizado pelo MXPOST (Ratnaparkhi, 1996), de acordo com o *tagset* reduzido do NILC (Tabela 6.3). O MXPOST foi treinado com o córpus do NILC, com precisão geral de 97%[4].

Para diminuir a taxa de erros nos verbos, é feito um pós-processamento de correção de erros óbvios. Também são feitas correções mais complexas, envolvendo a análise de informações gramaticais e de contexto do verbo (p.e., verbos no particípio etiquetados como adjetivos e vice-versa). A correção e a avaliação da precisão da etiquetação dos verbos serão apresentadas a seguir, na subseção *Processamento Sintático*.

Identificação de Citações

A identificação de citações é feita por meio de código *Perl* e utiliza o texto etiquetado na etapa anterior do processo. O texto etiquetado é utilizado por facilitar a identificação de números. A busca por citações é muito simples e segue a seguinte heurística:

> Se encontrar [ ], ou ( ) contendo números de 2 ou 4 dígitos, é citação.

Embora bastante superficial, essa heurística mostrou-se eficiente na identificação das citações em sentenças de resumos. Quando uma citação é encontrada, o texto é marcado com uma etiqueta XML do tipo <REF>.

Processamento Sintático

Grande parte do esforço de implementação da extração de atributos se aplicou aos atributos relacionadas aos verbos: **Tempo, Voz** e **Modal**. Isso ocorreu devido, por

[4]Informações em http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html

Tabela 6.2: Exemplos de expressões regulares no formato `flex` e possíveis expressões reconhecidas

Constantes	
NSPC (não-espaço)	[^ \n\t]*
SPC (espaço)	[\t]+
SER	(é|são|foi|foram)
EST	(este|esse|estes|esses|esta|essa|estas|essas)
DEST	(deste|desse|destes|desses|desta|dessa|destas|dessas)
NOMES	(documento|artigo|doutorado|estudo|estudo{SPC}proposto|mestrado|texto|trabalho{SPC}proposto| trabalho|dissertação|pesquisa|trabalho{SPC}de{SPC}doutorado|trabalho{SPC}de{SPC}mestrado| dissertação{SPC}de{SPC}mestrado|pesquisa{SPC}de{SPC}doutorado|pesquisa{SPC}proposta| pesquisa{SPC}de{SPC}mestrado|projeto{SPC}de{SPC}doutorado|projeto{SPC}de{SPC}mestrado| projeto{SPC}proposto|projeto|tese|tese{SPC}de{SPC}doutorado)
C-NOMES	(atividade|tarefa|técnica|pesquisa|estudo|alternativa){NSPC}
C-VERBOSCPX	(aplicad|empregad|explorad|usad|utilizad|aceit|estudad){NSPC}
L-ADJ	(essencia|fundamenta|vita|necessári|preciso|difíc|complex|complicad|desafiador|problemátic|critíc|inviáve| incomplet|inconsistente|imprecis|impraticáve|indispensáve){NSPC}
P-VERBOS	(introd|test|model|verific|valid|projet|realiz|constr|desenvolve|dedic|explor|concentr|abord|analis| apresent|busc|consist|contribu|defin|descrev|discut|enfatiz|enfoc|estend|estud|focaliz|fornec|implement| indic|investig|mostr|objetiv|procur|prop|question|relat|report|represent|revis|traç|trat|vis|formul){NSPC}
M-VERBOS	(adot|empreg|usa|usou|utiliz){NSPC}
R-NOMES	(análise|estudo|pesquisa|trabalho|obtidos?)
R-VERBOSCPX	(analisar|demonstrar|determinar)
CO-INI	(a{SPC}conclusão|a{SPC}contribuição|a{SPC}maior{SPC}contribuição|a{SPC}maior{SPC}virtude| a{SPC}originalidade|a{SPC}outra{SPC}conclusão|a{SPC}outra{SPC}contribuição| a{SPC}principal{SPC}conclusão|a{SPC}principal{SPC}contribuição|a{SPC}relevância| a{SPC}relevância{SPC}e{SPC}originalidade|as{SPC}conclusões|as{SPC}contribuições| as{SPC}maiores{SPC}contribuições|as{SPC}principais{SPC}conclusões| as{SPC}principais{SPC}contribuições)

Categoria	Expressão regular	Exemplos de expressões reconhecidas
Contexto	{SER}{SPC}{C-NOMES}{SPC}{C-VERBOSCPX}	*são atividades empregadas* *são alternativas utilizadas* *são técnicas exploradas* ...
Lacuna	{SER}{SPC}{L-ADJ}	*é necessário* *são difíceis* *é problemático* ...
Propósito	{EST}{SPC}{NOMES}{SPC}{P-VERBOS}	*este trabalho investiga* *este trabalho de mestrado explora* *esta dissertação apresenta* ...
Metodologia	{EST}{SPC}{NOMES}{SPC}{M-VERBOS}	*este estudo empregou* *esta pesquisa adotou* *este trabalho usa* ...
Resultado	{R-NOMES}{SPC}permitiu{SPC}{R-VERBOSCPX}	*estudo permitiu determinar* *análise permitiu demonstrar* *trabalho permitiu analisar* ...
Conclusão	{CO-INI}{SPC}{DEST}{SPC}{NOMES}	*a maior contribuição deste trabalho* *a conclusão desta pesquisa* *as contribuições desta dissertação* ...

Tabela 6.3: *Tagset* utilizado pelo MXPOST

Tagset reduzido	
Etiqueta	**Descrição**
I	interjeição
LOCU	locução
PREP	preposição
N	nome
NP	nome próprio
VERB	verbo
ADV	advérbio
PRON	pronome
CONJ	conjunção
NUME	numeral
ART	artigo
RES	resíduo
PDEN	palavra denotativa
Contrações	
Etiqueta	**Descrição**
PREP+ART	preposição + artigo
PREP+PD	preposição + pronome demonstrativo
PREP+PPR	preposição + pronome pessoal reto
PREP+N	preposição + nome
PREP+PPOT	preposição + pronome pessoal oblíquo tônico

um lado, à grande produtividade flexional da morfologia do português e, por outro, ao fato de a saída do etiquetador morfossintático usar uma etiqueta única para verbos, sem distinção entre os diversos tipos de auxiliar e principais ou mesmo entre verbos com ou sem clíticos. O mesmo ocorre com os pronomes, não havendo subclassificação para os oblíquos átonos ou ainda as diversas funções do pronome *se*. Em suma, o etiquetador não só deixa de reconhecer os complexos verbais, como também não especifica suficientemente seus componentes e, por vezes, nem mesmo corretamente. É o caso dos particípios, muitas vezes impropriamente etiquetados como adjetivos, não sendo assim considerados como parte de complexos verbais.

Para refinar e corrigir os dados disponíveis, foi implementado o *pipeline* apresentado na Figura 6.13, que compreende os seguintes passos:

1. para cada palavra etiquetada como VERB ou ADJ, são recuperadas suas possíveis acepções como verbo ou adjetivo, desconsiderando o contexto. Cada acepção é constituída de um lema[5] e traços morfológicos (gênero, número, pessoa, tempo, modo, aspecto) e morfossintáticos aplicáveis (classe gramatical, subclasse [infinitivo, particípio, gerúndio, pronominal, auxiliar, modal], presença de clíticos [tem_clítico,

[5]Dentre as várias flexões de uma palavra, o lema é a forma convencional com que se faz referência à palavra desconsiderando-se sua flexão. Por exemplo, entre os verbos, o lema é usualmente o infinitivo impessoal.

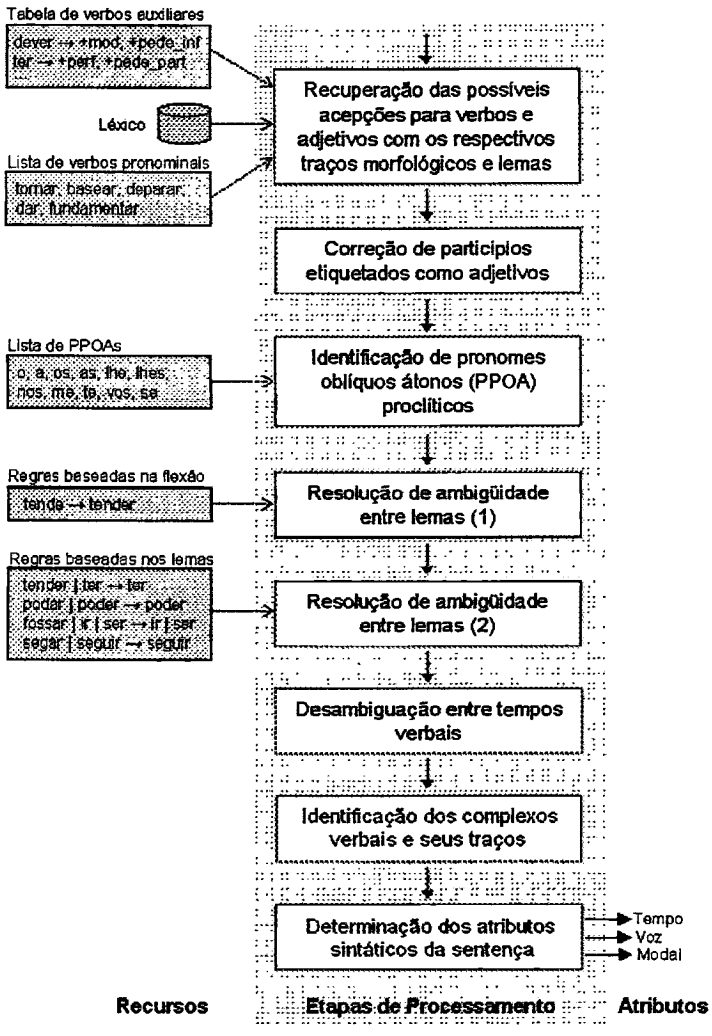


Figura 6.13: Etapas do processamento sintático

tem_se] e, para o caso de verbos auxiliares, requisitos sobre o verbo seguinte no complexo verbal [pede_infinitivo, pede_particípio] e traços funcionais [aux_contínuo, aux_perfeito, aux_passiva]). Essas informações provêm de quatro fontes distintas: o léxico do NILC, um algoritmo de isolamento de enclíticos e mesoclíticos, uma tabela com informações sobre os verbos auxiliares e uma lista de verbos pronominais. Em vez de procurar se ater a definições tradicionais de verbo pronominal, essa lista é definida neste trabalho pelos verbos que usualmente são aplicados com partícula *se*, mas que não parecem, com isso, estabelecer oposição semântica com uma forma em voz ativa com agente explícito;

2. para cada palavra com acepções concorrentes de adjetivo e particípio, uma ou outra é descartada considerando-se as informações disponíveis para a palavra precedente, tais como (i) se é um determinante, (ii) se pode ser um verbo auxiliar e, nesse caso,

(iii) quais os requisitos desse verbo específico sobre o particípio;

3. para cada palavra etiquetada como PRON, verifica-se sua pertinência a uma lista de pronomes oblíquos átonos. Em caso positivo, é etiquetada como pronome proclítico (PPOA), recebendo ainda a etiqueta adicional PPOA-SE, caso se trate do pronome *se*;

4. para cada palavra etiquetada como VERB e que tenha mais de um possível lema, tenta-se resolver essa ambigüidade aplicando-se regras simples que estabelecem prioridades na seleção de lemas. Essas regras são de dois tipos, a saber: (i) regras baseadas nos lemas, no formato $Lema_1|Lema_2|...|Lema_n \rightarrow Lema$, significando que, se houver ambigüidade entre os lemas do lado esquerdo da regra, então prevalece o lema do lado direito; (ii) regras baseadas na flexão, no formato $Forma \rightarrow Lema$, significando que, se a palavra em questão é literalmente $Forma$, então o lema selecionado deve ser $Lema$. Como apresentado na Figura 6.13, as regras baseadas na flexão têm prioridade de aplicação sobre as baseadas nos lemas e foram concebidas para tratar exceções a estas últimas;

5. para cada palavra etiquetada como VERB e que, dentro da acepção selecionada no passo anterior, tenha mais de um possível conjunto de traços de flexão (número, pessoa, tempo e/ou modo), tenta-se resolver essa ambigüidade por meio de algumas heurísticas simples. Os conflitos tratados são os seguintes: (i) futuro do subjuntivo vs. infinitivo impessoal, (ii) imperativo vs. presente do subjuntivo, (iii) pretérito perfeito vs. mais-que-perfeito, (iv) presente do indicativo vs. imperativo de 2ª pessoa do singular. As heurísticas que tratam de (i) e (ii) analisam o segmento da sentença o qual separa o verbo analisado de um outro verbo ou do início da sentença e buscam palavras-chave que "ativem" uma ou outra flexão. No caso do futuro do subjuntivo, procuram-se as conjunções *se* ou *quando*; no caso do presente do subjuntivo, o advérbio *talvez*, a palavra *que* em qualquer classe gramatical ou ainda uma conjunção qualquer. O infinitivo impessoal e o imperativo, respectivamente, são selecionados quando essas condições não são satisfeitas. Por sua vez, a resolução dos conflitos (iii) e (iv) não recorre à qualquer análise contextual, simplesmente eliminando os casos (extremamente) mais raros (mais-que-perfeito e imperativo de 2ª pessoa do singular, respectivamente);

6. para cada seqüência de palavras etiquetadas como VERB, podendo conter eventuais pronomes oblíquos átonos (proclíticos, etiquetadas como PPOA no passo 3), advérbios ou preposições, tenta-se formar complexos verbais. A tarefa mais importante desse passo consiste em verificar, para cada verbo marcado como potencialmente auxiliar no passo 1, se este é de fato auxiliar na sentença considerada. Para isso, verifica-se a satisfação, por parte do verbo seguinte na seqüência, dos traços

"pede_infinitivo" ou "pede_particípio" do verbo auxiliar. Para alguns auxiliares em específico (ex. *ter que|de, começar a, etc.*), permite-se e verifica-se a presença de uma preposição requerida. Por fim, os complexos verbais são delimitados e transformados em um único token, aglutinando-se a cada verbo principal toda a seqüência de verbos auxiliares (incluindo eventuais advérbios, pronomes e preposições) que o precede. Cada complexo verbal resultante conta com traços finais: (i) tempo-modo do componente finito, (ii) ±aspecto contínuo (estar + gerúndio), (iii) ±aspecto perfeito (ter|haver + particípio), (iv) ±modal, (v) ±voz passiva analítica (ser + particípio) e (vi) ±voz passiva sintética ou indeterminação do sujeito (verbo principal não-pronominal + flexão de 3ª pessoa + clítico *se*).

7. para cada sentença, determinam-se seus atributos sintáticos a partir dos traços gramaticais observados para o complexo verbal da oração principal da sentença. A identificação deste complexo é realizada de forma muito simples, sem envolver análise sintática, por meio da seguinte heurística: selecionar o primeiro verbo finito da sentença que não esteja no modo subjuntivo; caso só haja verbos no subjuntivo, selecionar o primeiro. Os verbos no subjuntivo são preteridos por geralmente participarem de orações subordinadas.

Para avaliar a precisão do processamento sintático descrito acima foi realizado um experimento com 100 sentenças extraídas aleatoriamente do córpus de treinamento. Foi observada a precisão no reconhecimento dos complexos verbais, na determinação da voz passiva (analítica e sintética|indeterminação do sujeito (I.s.)) e na identificação de verbos modais. Os números resultantes da avaliação são apresentados na Tabela 6.4. Vale lembrar que grande parte dos erros observados, em especial os casos não reconhecidos, são conseqüência de erros do etiquetador morfossintático.

Tabela 6.4: Resultados da avaliação do processo sintático

	Complexos verbais finitos	Voz passiva analítica	Voz passiva sintética ou I.s.	Modal
Total na amostra	181	52	15	17
Total de reconhecidos corretos	175	49	15	15
Total de reconhecidos errados	4	0	1	0
Total de reconhecidos parcial	1	–	–	–
Total de não reconhecidos	5	3	0	2

Na Tabela 6.5, são apresentados os mesmos resultados, expressos em termos das medidas *Precision*, *Recall* e *F-Measure*. Para esse caso, *Precision* é o número de casos corretamente reconhecidos sobre o total de casos reconhecidos, enquanto *Recall* é o número de casos corretamente reconhecidos sobre o total de casos presentes na amostra.

Tabela 6.5: Resultados da avaliação do processo sintático em termos de *Precision*, *Recall* e *F-Measure*

	Precision	Recall	F-Measure
Complexos verbais finitos	0.97	0.96	0.96
Voz passiva analítica	1	0.94	0.97
Voz passiva sintética ou l.s.	0.93	1	0.96
Modal	1	0.88	0.93

6.2.5 Descrição do Classificador Estatístico

Foi implementado um classificador Naive Bayes simples para estimar a probabilidade que uma sentença S tenha a categoria C dados os valores de seus atributos. A categoria com a maior probabilidade é escolhida como saída para S. Esse classificador é semelhante ao utilizado por Teufel & Moens (2002) e é apresentado na Figura 6.14.

$$P(C^i \mid V_{0,x}, ..., V_{n-1,y}) = P(C^i)\frac{P(V_{0,x},...,V_{n-1,y}|C^i)}{P(V_{0,x},...,V_{n-1,y})} \approx P(C^i)\frac{\prod_{j=0}^{n-1} P(V_{j,r}|C^i)}{\prod_{j=0}^{n-1} P(V_{j,r})}$$

$P(C^i \mid V_{0,x}, ..., V_{n-1,y})$:	Probabilidade de uma sentença ter categoria C^i, dados os valores $V_{0,x}, ..., V_{n-1,y}$ de seus atributos;
$P(C^i)$:	Probabilidade de uma sentença ter categoria C^i(prior);
$P(V_{j,r} \mid C^i)$:	Probabilidade do par atributo-valor $V_{j,r}$ ocorrer em uma sentença de categoria C^i;
$P(V_{j,r})$:	Probabilidade do valor de atributo $V_{j,r}$ (r-ésimo valor do Atributo A_j);

Figura 6.14: Classificador Naive Bayes (Teufel & Moens, 2002)

O aprendizado é supervisionado, isto é, durante a fase de treinamento, o classificador aprende as associações entre os atributos e as categorias provenientes do córpus manualmente anotado. Durante a fase de teste, o modelo previamente treinado fornece a probabilidade de cada categoria para cada sentença de entrada, com base nos atributos identificados para a sentença.

6.2.6 Avaliação Intrínseca do AZPort

Os resultados da classificação foram computados aplicando-se 13-13-*fold cross-validation* aos 52 resumos anotados no experimento descrito na Seção 6.2.2. Dessa forma, em cada rodada o *AZPort* foi treinado com 48 resumos e testado com 4 resumos. Como *Baseline* 1, foi utilizada a escolha aleatória de categorias, considerando-se suas distribuições no

córpus. Como *Baseline* 2, foi utilizada a atribuição da categoria mais freqüente a todas as sentenças. A distribuição de categorias no córpus é apresentada na Tabela 6.6.

Tabela 6.6: Distribuição das categorias

Categoria	%
CONTEXTO	21
LACUNA	10
PROPÓSITO	18
METODOLOGIA	12
RESULTADO	32
CONCLUSÃO	5
ESTRUTURA	2

Comparando-se o *AZPort* (treinado com o conjunto completo de atributos) a um anotador humano, a concordância entre eles, medida por meio do *Kappa*, é K=0.65. Esse resultado é bastante encorajador se comparado com o resultado relatado por Teufel & Moens (2002) (K=0.45). A diferença de desempenho entre os dois sistemas pode ser justificada pelo fato de estarmos trabalhando apenas com resumos (em vez de artigos completos) e também por todos os resumos pertencerem ao mesmo domínio (Ciência da Computação). O resultado do *AZPort* também supera os resultados das duas *baselines*, conforme apresentado na Tabela 6.7.

Tabela 6.7: Desempenho do *AZPort* e das *baselines*, calculado em comparação a um anotador humano

	AZPort	*Baseline* 1	*Baseline* 2
Kappa	0.65	0	0.26
Acurácia	0.72	0.20	0.32

Uma análise mais aprofundada dos resultados mostrou que, exceto pela categoria ESTRUTURA, o classificador apresenta um bom desempenho em todas as categorias, conforme pode ser observado na matriz de confusão mostrada na Tabela 6.8. Os resultados da classificação feita pelo *AZPort* estão representados nas colunas da matriz enquanto os resultados do anotador humano estão representados nas linhas.

Para analisar o comportamento do *AZPort* com relação a cada categoria, novamente foram utilizadas as medidas *Precision*, *Recall* e *F-Measure*. Para esse caso, dada uma categoria C, *Precision* é o total de sentenças corretamente classificadas como C sobre o total de sentenças classificadas como C. *Recall* é o total de sentenças corretamente classificadas como C sobre o total de sentenças pertencentes à categoria C presentes no conjunto. Na Tabela 6.9 são apresentados os valores das três medidas para cada uma das categorias.

Conforme comentado anteriormente, o classificador tem um desempenho ruim para a categoria ESTRUTURA (*F-Measure* =0). Isso era esperado, uma vez que apenas seis

Tabela 6.8: Matriz de Confusão: anotação manual vs. anotação automática

	AZPort								
	Categoria	C	L	P	M	R	Co	E	Total
Humano	C	**57**	10	2	1	7	0	0	77
	L	11	**23**	0	0	2	0	0	36
	P	6	1	**49**	0	8	1	0	65
	M	5	0	0	**26**	14	0	0	45
	R	2	2	0	9	**101**	3	0	117
	Co	0	0	0	0	9	**10**	1	20
	E	0	0	0	0	5	1	**0**	6
	Total	81	36	51	36	146	15	1	366

Tabela 6.9: *Precision*, *Recall* e *F-Measure* por categoria

Categoria	Precision	Recall	F-Measure
CONTEXTO	0.704	0.74	0.722
LACUNA	0.639	0.639	0.639
PROPÓSITO	0.961	0.754	0.845
METODOLOGIA	0.722	0.578	0.642
RESULTADO	0.692	0.863	0.768
CONCLUSÃO	0.667	0.5	0.571
ESTRUTURA	0	0	0

sentenças do córpus de treinamento são rotuladas como ESTRUTURA. Muitos algoritmos de aprendizado de máquina, incluindo o Naive Bayes, têm um desempenho ruim com categorias pouco freqüentes, devido à insuficiência de material de treinamento. Com relação às outras categorias, o melhor desempenho do classificador foi para a categoria PROPÓSITO (*F-Measure* =0.845), seguida da categoria RESULTADO (*F-Measure* =0.769) (veja Tabela 6.9). O bom desempenho para a categoria PROPÓSITO pode ser atribuído à presença de marcadores discursivos fortes em sentenças dessa categoria (modelados pelo atributo **Expressão**). Quanto à categoria RESULTADO, o bom desempenho pode ser atribuído a alta freqüência desse tipo de sentença no córpus de treinamento e também à presença de expressões padrão específicas.

Analisando-se a contribuição de cada atributo isoladamente, foi constatado que **Expressão** é o atributo com maior poder de distinção. Observou-se também que considerar o contexto de ocorrência da sentença (modelado pelo atributo HISTÓRICO) é uma heurística bastante útil e melhora significativamente o desempenho do classificador em 12%. Os atributos sintáticos – **Tempo, Voz** e **Modal** – e o atributo **Citação** foram os mais fracos. É possível que o atributo **Citação** obtenha um melhor desempenho em outro tipo de texto, como introduções, que normalmente apresentam um numero maior de citações.

O potencial de distinção de cada atributo é apresentado na Tabela 6.10. Nessa tabela, a segunda coluna apresenta o potencial de distinção do atributo isoladamente, em termos do *Kappa* entre o classificador e um anotador humano. Como pode ser visto, com exceção

Tabela 6.10: Potencial de distinção de cada atributo em termos do Kappa entre um anotador humano e o *AZPort*

Atributo	Sozinho	Excluído
Tamanho	-0.106	0.62
Localização	-0.047	0.624
Citação	-0.272	0.63
Expressão padrão	0.557	0.345
Tempo	-0.166	0.642
Voz	-0.018	0.644
Modal	-0.287	0.65
Histórico	0.251	0.54
Baseline 1 (aleatório por distr.): K=0		
Baseline 2 (categoria mais freq.): K=0.26		

dos atributos **Expressão** e **Histórico**, todos os outros atributos têm um desempenho pior que o das duas *baselines*. A terceira coluna da Tabela 6.10 apresenta os valores *Kappa* para experimentos realizados com todo o conjunto de atributos menos um, isto é, deixando sempre um dos atributos de fora do conjunto. Conforme apresentado, todos os atributos, exceto os sintáticos, contribuem combinadamente para a distinção das categorias.

Os resultados da classificação automática estão razoavelmente de acordo com os resultados dos experimentos feitos anteriormente para anotação manual, isto é, as categorias tidas como "confusas" pelos anotadores humanos são similares às categorias confundidas pelo classificador automático. Conforme pode ser observado na Tabela 6.8, o classificador tem problemas de distinção com as categorias METODOLOGIA, RESULTADO e CONCLUSÃO, assim como os anotadores humanos. Conforme comentado na Seção 6.2.2, a transformação dessas três categorias em uma única aumenta consideravelmente a concordância entre os anotadores, o que sugere que existem problemas de distinção dessas categorias mesmo para humanos treinados.

6.2.7 Considerações sobre o AZPort

Os resultados da avaliação intrínseca do *AZPort* mostraram que, embora seu desempenho esteja bem abaixo do desempenho humano, é bastante promissor e encontra-se no nível de desempenho alcançado pelos outros classificadores retóricos da literatura. Atualmente, sua principal limitação é a manipulação restrita a textos curtos, como resumos. Testes preliminares com textos maiores, como introduções, mostraram que o desempenho do *AZPort* diminui consideravelmente. Outra limitação é a dependência da base de expressões padrão. Embora a base seja razoavelmente grande, a probabilidade de textos novos conterem expressões novas e não incluídas na base é alta. Uma vez que o atributo **Expressão** é o mais forte, esse fator afeta diretamente o desempenho do classificador. A avaliação extrínseca do *AZPort*, isto é, como parte do sistema de crítica do ambiente

SciPo, será apresentada na Seção 6.4.

6.3 Apoio à Escrita do Texto

Para apoiar a escrita do texto propriamente dito, foi incorporado ao ambiente *SciPo* um formulário de edição, representando a estrutura esquemática construída pelo usuário (ou detectada pelo *AZPort*). Nesse formulário, é criada uma caixa de texto para cada componente/estratégia retórica selecionada pelo usuário, de modo que a escrita do Resumo/Introdução seja orientada à estrutura escolhida. Dessa maneira, o usuário pode compor seu Resumo/Introdução ao mesmo tempo em que acessa os recursos do ambiente, como a recuperação de exemplos similares, a pesquisa de exemplos de componentes e estratégias retóricas e a navegação na Base de exemplos. A Figura 6.15 mostra como o formulário de edição e os outros recursos do ambiente *SciPo* se integram.

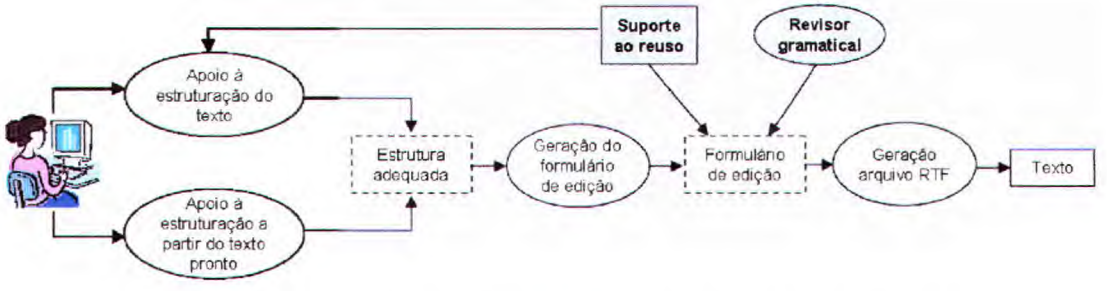


Figura 6.15: Visão geral do processo de apoio à escrita do texto

Como as caixas de texto do formulário de edição estão associadas a componentes/estratégias retóricas, o usuário pode visualizar diretamente exemplos de cada um desses componentes/estratégias por meio dos *links* denominados Ver Exemplos. Nesse contexto de escrita, foi implementado um recurso de inserção de trechos de texto reutilizáveis (anotados nos textos da Base de exemplos pela etiqueta XML **<Reutilizavel>**). Dessa forma, quando visualizando um exemplo, o usuário pode transferir para caixa de texto em foco as partes de texto reutilizável. Os trechos são separados por sentença e os trechos não-reutilizáveis são substituídos por lacunas do tipo "__". A Figura 6.16 exemplifica o formulário de edição e o mecanismo de reutilização de trechos de texto.

Outra ferramenta integrada ao ambiente *SciPo* para apoiar a escrita é o revisor ortográfico e gramatical ReGra[6] (Martins *et al.*, 1998). Foi criada uma interface Web para o acesso ao revisor, de forma que o usuário pode utilizá-lo no ambiente *SciPo* da mesma forma que é feito em um editor de texto comum (p.e., Microsoft Word). A Figura 6.17 exemplifica o uso da interface do revisor no formulário de edição do ambiente *SciPo*.

[6]O revisor ReGra foi desenvolvido no NILC em projeto de parceria com a Itautec/Philco.

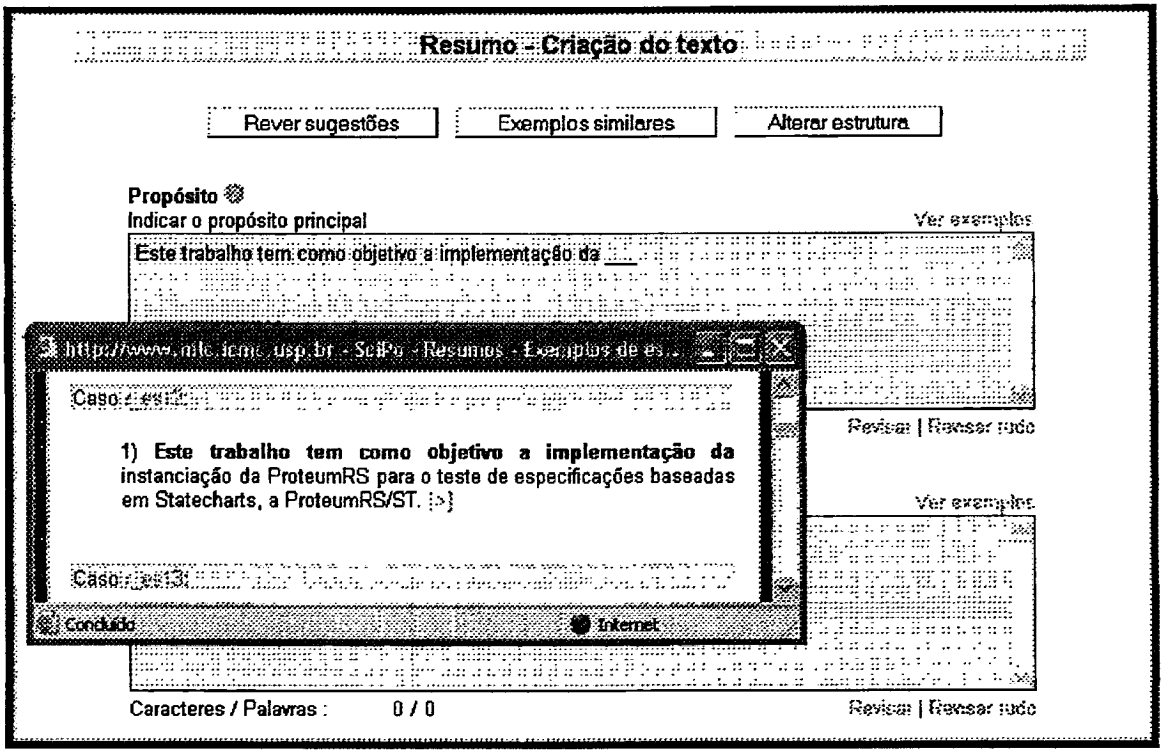


Figura 6.16: Exemplo de formulário de edição e de inserção de texto reutilizável

Uma vez que a escrita do Resumo/Introdução tenha sido finalizada, o usuário pode solicitar o agrupamento do texto redigido no formulário de edição em um arquivo do tipo RTF[7]. Essa função permite que o texto redigido no *SciPo* possa ser posteriormente manipulado em editores de texto comuns. Uma limitação, especialmente no caso das Introduções, é que essa função não prevê o tratamento de parágrafos, isto é, o texto do formulário de edição é sempre agrupado como um único parágrafo.

6.4 Avaliação do Ambiente SciPo

Conforme foi dito no início deste capítulo, o protótipo *SciPo* segue uma abordagem baseada em córpus e em sistemas de crítica e apóia a estruturação de Resumos e Introduções e a realização lingüística dessas estruturas de forma flexível, deixando o usuário livre para escolher entre dois modos de trabalho: (i) um processo *top-down*, que parte do planejamento estrutural para a escrita propriamente dita, ou (ii) um processo *bottom-up*, em que se submete um texto já escrito à análise automática da estrutura.

Os dois modos de interação foram considerados na avaliação, de forma que foram reali-

[7] *Rich Text Format*

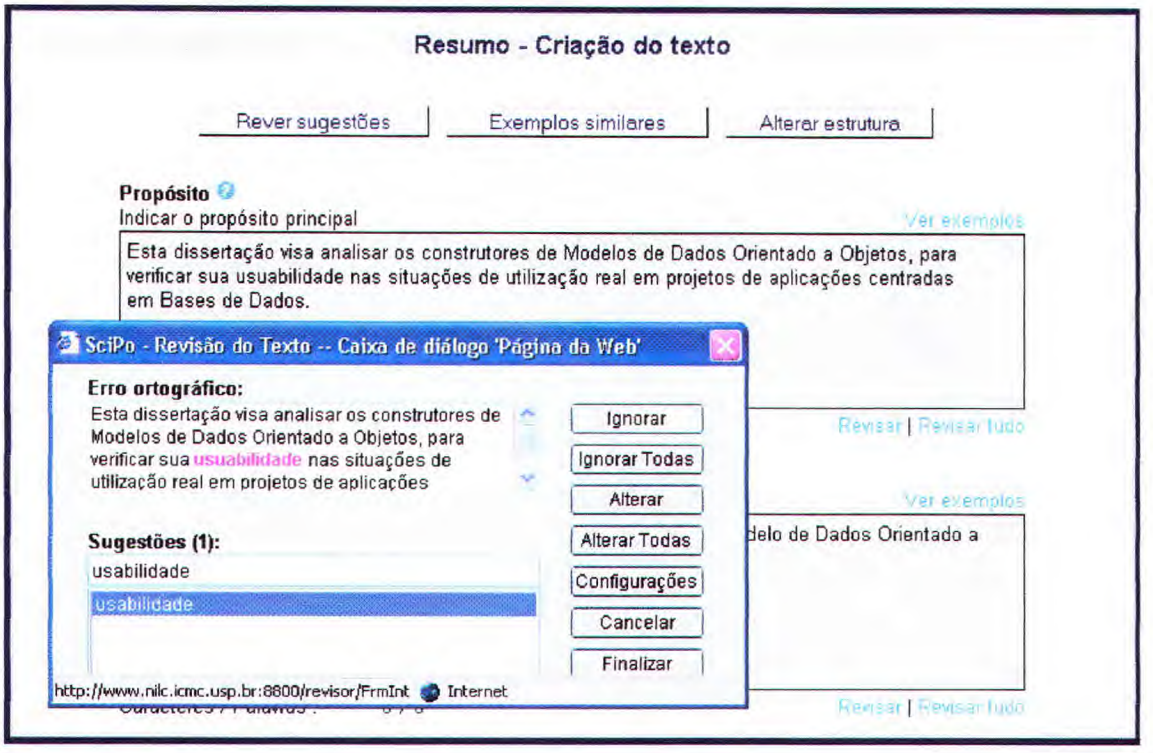


Figura 6.17: Exemplo de uso do revisor ortográfico e gramatical

zados dois experimentos. No Experimento 1, foi solicitado aos usuários que trabalhassem no modo *bottom-up*, enquanto no Experimento 2, foi solicitado que trabalhassem no modo *top-down*. Ambos os experimentos tiveram por objetivo maior avaliar o quanto o modelo subjacente ao *SciPo* é de fato efetivo no auxílio à escrita acadêmica.

Os experimentos foram realizados com estudantes voluntários dos cursos de graduação e pós-graduação em Ciência da Computação do ICMC. A escrita de Resumos foi o foco dos dois experimentos.

6.4.1 Experimento 1

O Experimento 1 foi realizado com quatro estudantes de pós-graduação do ICMC que escreveram suas dissertações de mestrado recentemente. Foi pedido aos estudantes que utilizassem o *SciPo* para reescrever o resumo de suas dissertações, iniciando a interação com o ambiente como um processo *bottom-up*, isto é, submetendo seus resumos já escritos à análise. Durante a execução do experimento, a autora do ambiente estava presente, mas só foram feitas intervenções quando requeridas pelo estudante. Antes do experimento ser iniciado, foi pedido aos estudantes que lessem um documento explicativo da interface e das funções do *SciPo*, uma vez que era nosso interesse avaliar apenas a eficácia do

ambiente no auxílio à escrita. Outro interesse desse experimento foi observar o impacto dos erros cometidos pelo *AZPort* no resultado final das críticas emitidas pelo *SciPo*. Após o término da avaliação, todos os estudantes preencheram um questionário relatando suas impressões sobre o ambiente. O questionário utilizado nesse experimento é apresentado no Apêndice E.

Foi pedido aos estudantes que utilizassem o resumo de suas dissertações como entrada para o sistema de detecção automática da estrutura esquemática, o *AZPort*. Antes de ser acionado o processo de crítica, era permitido aos estudantes corrigirem o resultado da detecção automática, se desejado. Dois estudantes fizeram correções, enquanto os outros dois acataram a estrutura detectada automaticamente como correta. Os quatro estudantes receberam críticas e sugestões do ambiente e modificaram seu resumo em algum nível.

Embora cientes da precisão do *AZPort*, os estudantes consideraram seus resultados bastante confiáveis e isso influenciou a forma como interagiram com o ambiente. O *AZPort* cometeu erros de classificação em três dos quatro resumos submetidos, com impactos diferentes nas críticas resultantes. Os erros de classificação observados foram:

(a) Lacuna vs. Contexto
(b) Propósito vs. Contexto
(c) Metodologia vs. Resultado
(d) Propósito vs. Conclusão

Os erros (a) e (c) ocorreram em um mesmo resumo (Estudante 1). Os erros (b) e (d) ocorreram em resumos diferentes (Estudantes 2 e 3, respectivamente). O Estudante 1 não corrigiu nenhum dos dois erros, enquanto os Estudantes 2 e 3 fizeram as devidas correções. Essa diferença de comportamento entre os estudantes pode ser atribuída ao tipo de erro de classificação cometido pelo *AZPort*. Em (b) e (d), os erros envolvem categorias bastante distintas e, dessa forma, facilmente percebidos pelo estudante. Já os erros (a) e (c), constituem um problema maior, uma vez que as categorias envolvidas apresentam dificuldades de distinção mesmo para anotadores humanos. Sendo assim, o Estudante 1 não percebeu o erro e aceitou a saída do *AZPort* como correta. Esse erro fez com que o ambiente emitisse críticas e sugestões de pouco efeito para o resumo do Estudante 1.

Quanto às respostas ao questionário, os quatro estudantes reportaram suas experiências com o ambiente *SciPo* como positiva. Conforme dito acima, todos consideraram o sistema de detecção automática da estrutura esquemática confiável. As críticas e sugestões emitidas pelo *SciPo* foram consideradas relevantes por todos, exceto por um estudante que classificou apenas as sugestões como irrelevantes. Todos os estudantes avaliaram o *SciPo* como útil e declararam suas intenções em utilizar o ambiente novamente em situações reais.

Para avaliar se houve melhoria no texto, foram comparados os dois conjuntos de resumos: (1) os originais e (2) os reescritos com o auxílio do ambiente *SciPo*. Os resumos foram avaliados por um juiz experiente em escrita acadêmica visando identificar melhorias no conjunto (2) com relação à estruturação. Foi informado ao juiz qual era o conjunto original e também o modelo de estruturação de resumos utilizado pelo *SciPo*.

Os resultados da análise do juiz mostraram que os resumos reescritos no *SciPo* (conjunto 2) são mais informativos, no sentido de fornecerem mais dados factuais do que os resumos originais (conjunto 1). Entretanto, esses resumos não podem ser classificados como resumos de "melhor qualidade", uma vez que outros problemas de escrita permaneceram. O ambiente *SciPo* foca especialmente na estruturação do resumo e existem outros fatores que contribuem para sua qualidade, como o uso correto da gramática, estilo, entre outros. De qualquer forma, o experimento mostrou que o ambiente *SciPo* é útil, no sentido de guiar o escritor na composição de resumos mais informativos e de acordo com o gênero acadêmico.

6.4.2 Experimento 2

O Experimento 2 foi realizado com quatro estudantes de graduação do ICMC que haviam realizado trabalhos de iniciação científica (IC) recentemente. Foi pedido aos estudantes que utilizassem o *SciPo* para escrever um resumo do seus trabalho de IC, iniciando a interação com o ambiente como um processo *top-down*, isto é, partindo do planejamento estrutural para a escrita propriamente dita.

O procedimento adotado na realização desse experimento foi similar ao descrito no Experimento 1. Novamente, nosso interesse foi avaliar a eficácia do ambiente no auxílio à escrita e os estudantes preencheram um questionário relatando suas impressões gerais sobre o *SciPo*. Esse questionário também é apresentado no Apêndice E.

Os quatro estudantes construíram suas estruturas esquemáticas guiados pelas críticas e sugestões emitidas pelo ambiente, obtendo ao final do processo de crítica, uma estrutura esquemática de Resumo adequada. Entretanto, na etapa de escrita do Resumo (preenchimento das caixas de texto do formulário de edição) um problema mostrou-se recorrente — a falta de conhecimento relativo ao significado dos componentes do modelo estrutural. Constatou-se que o uso de funções do ambiente, como a visualização de exemplos de componentes/estratégias e a navegação na Base de exemplos, minimizam esse problema, mas não o solucionam completamente. Todos os estudantes fizeram uso dessas funções e, ainda assim, dois deles tiveram problemas. O Estudante 1 interpretou o elemento Lacuna::Citar problemas/dificuldades como representando as dificuldades encontradas por ele próprio na realização do trabalho, em vez de citar problemas e dificuldades da área de pesquisa em questão. O Estudante 2 utilizou o elemento Contexto::Introduzir a pesquisa

a partir da grande área para introduzir o propósito da sua pesquisa, em vez entender esse elemento como uma estratégia de contextualização.

Quanto ao questionário, os resultados foram similares aos reportados para o Experimento 1. Os quatro estudantes reportaram sua experiência com o ambiente *SciPo* como altamente positiva. Novamente, um estudante considerou as sugestões dadas pelo ambiente como irrelevantes. Com relação às funções disponíveis no ambiente, a menos utilizada foi a recuperação de exemplos com estruturas similares, acionada por apenas um dos estudantes. A mais utilizada foi a visualização de exemplos de componentes/estratégias. Todos os estudantes reportaram sua intenção de utilizar o *SciPo* em situações reais.

Nesse experimento, não foi possível fazer uma análise comparativa dos resumos, como no Experimento 1, uma vez que não existiam resumos originais para a comparação. Analisando-se os resumos desse experimento isoladamente, foi possível notar que, embora o ambiente seja útil no auxílio à estruturação, não há como garantir que os textos resultantes sejam bons, principalmente quando o estudante tem pouco ou nenhum conhecimento sobre o gênero acadêmico.

6.5 Considerações Finais

O protótipo *SciPo* faz jus ao levantamento de requisitos realizado no capítulo anterior e não pode ser considerado uma mera transposição do AMADEUS para o português, como pode ser observado na Tabela 6.11, em que se compara resumidamente o *SciPo* com as demais ferramentas estudadas. A introdução do sistema de detecção automática de estrutura esquemática é um diferencial importante, permitindo levar a abordagem de crítica a um novo patamar de interatividade ainda a ser melhor explorado em versões futuras do protótipo.

Tabela 6.11: Resumo das ferramentas de auxílio à escrita investigadas, juntamente com o *SciPo*

	WE	Writer's Assistant	Academic Writer	Abstract Helper	Mover	AMADEUS	*SciPo*
Apoio à organização de idéias	√	√					
Brainstorm dirigido			√				
Hipertextos explicativos			√				
Exemplos autênticos				√		√	√
Modelos de estrutura esquemática						√	√
Identificação automática da estrutura					√		√
Crítica da estrutura						√	√
Indicação de estratégias retóricas						√	√
Base de padrões de escrita			√	√		√	√

Apesar de os usuários afirmarem que se sentem auxiliados pela ferramenta e que a consideram útil, deve-se admitir que seus benefícios são limitados a apenas alguns entre os fatores de qualidade de um texto, especialmente informatividade e conformidade com

o gênero. Talvez a maior crítica que se possa fazer ao protótipo seja relativa à sua usabilidade, já que está baseada na aquisição e no uso de uma linguagem de especificação (a estrutura esquemática) aparentemente problemática para a maioria dos usuários.

O protótipo *SciPo* está disponível por meio do endereço www.nilc.icmc.usp.br/ scipo.

Capítulo

7

Conclusões e Trabalhos Futuros

Este trabalho de pesquisa partiu da premissa de que a tarefa de produção de um texto acadêmico por escritores inexperientes é, por um lado, complexa devido à natureza específica do texto, à falta de familiaridade do autor, às influências de ordem cognitivas, naturais ao processo de escrita, à subjetividade nos julgamentos dos textos produzidos, entre outras razões. Por outro lado, trata-se de um processo cujo produto — o texto acadêmico — pode ser altamente padronizado e, portanto, possivelmente sujeito a procedimentos automatizados.

Com o objetivo de investigar a viabilidade de um ambiente de auxílio à escrita acadêmica em português, esse trabalho primeiramente constatou a existência, e analisou as características, de trabalhos similares para a língua inglesa, porém, não constatou nenhuma iniciativa para a língua portuguesa. Posteriormente, identificou, com usuários-alvo, seus anseios quanto a um ambiente computacional dessa natureza. De todas essas informações, levantou um conjunto de requisitos que, avaliados sob a luz das limitações tecnológicas atuais, convergiu para a proposta de um ambiente que conjuga auxílio baseado em conhecimento advindo de padrões levantados de córpus e um sistema interativo de crítica, possibilitando a criação livre de textos pelo usuário. A crítica é feita a partir da detecção automática dos componentes retóricos do texto do usuário. Adicionalmente, ferramentas de pós-edição de texto também são disponibilizadas.

Uma avaliação do protótipo resultante com usuários mostrou que essa abordagem híbrida de auxílio — liberdade na escolha de padrões e crítica do sistema — atende satisfatoriamente os requisitos iniciais, do ponto de vista do processo. Uma avaliação do

produto (texto), no entanto, exigirá, num trabalho futuro, um planejamento ainda mais elaborado. Embora contemplados parcialmente apenas dois componentes do modelo de texto acadêmico utilizado, o protótipo pode ser estendido naturalmente, uma vez que, com mais ou menos dificuldades, a metodologia de extração de conhecimento e de criação dos processos necessários às demais seções é exatamente a mesma aqui apresentada. Ressalva-se, no entanto, que nada se garante sobre o resultado da sucessiva aplicação do processo para a criação de todas as seções, ou seja, se o resultado obtido seria necessariamente melhor do que aquele gerado sem auxílio da ferramenta. Trata-se de hipótese a ser verificada futuramente. Já se sabe de antemão que a coesão do texto completo requer muito mais do que a justaposição de seções bem escritas.

Do ponto de vista da área de Lingüística Computacional para o português do Brasil, a principal contribuição é, sem dúvida, a criação de uma subárea de pesquisa riquíssima em desafios, com grande potencial para a geração de sistemas efetivamente úteis ao pesquisador brasileiro. Especificamente ao Processamento da Língua Portuguesa, esse trabalho contribui com um sistema de categorização retórica inédito, que pode ser utilizado em outras aplicações, como a sumarização automática, a categorização e a compreensão de textos.

Tal sistema, denominado *AZPort*, teve por base a técnica *Argumentative Zoning* (Teufel & Moens, 2002) e foi adaptado para classificar as sentenças de resumos acadêmicos em português em uma das 7 categorias retóricas definidas no projeto *SciPo:* Contexto, Lacuna, Propósito, Metodologia, Resultado, Conclusão e Estrutura. A precisão do *AZPort*, embora abaixo da precisão humana, está de acordo com o estado da arte para esse tipo de classificação, conforme demonstrado no Capítulo 4. Ainda, os resultados da avaliação extrínseca do *AZPort*, como parte do ambiente *SciPo*, demonstraram que a precisão atual é suficiente para sua utilização no contexto de uma ferramenta de crítica à escrita acadêmica.

No desenvolvimento do *AZPort*, foram implementados mecanismos para a extração automática dos valores dos atributos de classificação. Em especial, foi desenvolvido um *pipeline* para a extração de características sintáticas de verbos em português. Tal *pipeline*, que complementa as informações da etiquetação morfossintática, é utilizado na extração dos atributos de classificação **Tempo, Voz** e **Modal**, relativos ao primeiro verbo finito de cada sentença. Embora esse *pipeline* seja um produto intermediário do *AZPort*, não deixa de ser uma contribuição desta tese, uma vez que pode ser utilizado por qualquer aplicação que necessite extrair informações sintáticas de verbos. Ainda, esse mecanismo de extração de informações sintáticas pode ser estendido a outras classes de palavras que não verbos.

Outra contribuição deste trabalho são os resultados da análise do CorpusDT. Essa análise de córpus nos permitiu extrair padrões de estruturação, o que permitiu a defini-

ção de modelos estruturais mais específicos ao tipo de texto analisado. Também foram observados os desvios (com relação aos modelos prescritivos de boa escrita da literatura) recorrentes da comunidade acadêmica analisada – estudantes de pós-graduação em Ciências da Computação.

Como extensão imediata desse trabalho, pretende-se aplicar a metodologia proposta para a geração de conhecimento e recursos necessários para o completo tratamento das Introduções e Conclusões. Isso inclui, além da análise do córpus de Conclusões, a extensão do sistema de detecção automática da estrutura esquemática (*AZPort*), atualmente implementado para os Resumos. Tal tarefa envolve não apenas a adequação dos recursos utilizados no processo de extração dos atributos já implementados, mas também a implementação e análise de impacto de outros possíveis atributos. Além disso, seria interessante verificar o desempenho de tal sistema utilizando-se outros algoritmos de classificação que o Naive Bayes.

Conforme dito anteriormente, pretende-se investir em uma avaliação mais robusta do ambiente *SciPo*, visando uma avaliação mais extensa do ambiente e, principalmente, focando na avaliação do texto gerado por meio do uso do *SciPo*. Atualmente, já estamos trabalhando no desenvolvimento uma rubrica dedicada à avaliação de textos acadêmicos. Essa rubrica deverá contemplar, além dos aspectos estruturais já abordados pelo *SciPo*, outros aspectos que também influenciam na avaliação de um texto, tais como coerência, coesão e estilo.

Outra questão a ser abordada num trabalho futuro é a avaliação da interface do protótipo *SciPo*, com base em critérios de usabilidade, visando, em especial, investigar melhorias que possam minimizar o problema da interpretação dos modelos estruturais.

Referências Bibliográficas

Aluísio, S. M., and Gantenbein, R. E. 1997. Towards the application of systemic functional linguistics in writing tools. In *Proceedings of International Conference on Computers and their Applications*, 181–185.

Aluísio, S. M., and Oliveira Jr., O. N. 1995. A case-based approach for developing writing tools aimed at non-native english users. *Lecture Notes in Artificial Intelligence* 1010:121–132.

Aluísio, S. M., and Oliveira Jr., O. N. 1996. A detailed schematic structure of research papers introductions: An application in support-writing tools. *Revista de la Sociedad Espanyola para el Procesamiento del Lenguaje Natural* 19:141–147. Also available in <http://www.cica.es/sepln96/sepln96.html>.

Aluísio, S. M.; Barcelos, I.; Sampaio, J.; and Oliveira Jr., O. N. d. 2001. How to learn the many unwritten "Rules of the Game" of the Academic Discourse: A Hybrid Approach Based on Critiques and Cases. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, 257–260.

Aluísio, S. M. 1995. *Ferramentas para Auxiliar a Escrita de Artigos Científicos em Inglês como Língua Estrangeira.* Ph.D. Dissertation, Instituto de Física de São Carlos - USP.

Anthony, L., and Lashkia, G. V. 2003. Mover: A machine learning tool to assist in the reading and writing of technical papers. *IEEE Transactions on Professional Communication* 46(3):185–193.

Anthony, L. 1999. Writing research article introductions in software engineering: How accurate is a standard model? *IEEE Transactions on Professional Communication* 42(1):38–46.

Barrass, R. 1979. *Os Cientistas Precisam Escrever: Guia de Redação para Cientistas, Engenheiros e Estudantes.* São Paulo, SP: Editora da Universidade de São Paulo.

Barros, R. C. 2000. Modelagem de usuários para sistemas de auxílio à escrita técnica. Master's thesis, Instituto de Ciências Matemáticas e de Computação - USP.

Booth, W. C.; Colomb, G. G.; and Williams, J. M. 2000. *A Arte da Pesquisa.* São Paulo, SP: Ed. Martins Fontes.

Broady, E., and Shurville, S. 2000. Developing academic writer: Designing a writing environment for novice academic writers. In Broady, E., ed., *Second Language Writing in a Computer Environment.* London, UK: CiLT/AFLS. 131–152.

Burstein, J.; Chodorow, M.; and Leacock, C. 2003. Criterion online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the 15th Annual Conference on Innovative Applications of Artificial Intelligence.*

Burstein, J.; Marcu, D.; and Knight, K. 2003. Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing* 18(1):32–39.

Castro, C. d. M. 1981. *A Prática da Pesquisa.* São Paulo, SP: Ed. McGraw-Hill do Brasil.

Cohen, W. W. 1995. Fast effective rule induction. In Prieditis, A., and Russell, S., eds., *Proceedings of the 12th International Conference on Machine Learning,* 115–123. Tahoe City, CA, USA: Morgan Kaufmann.

Cohen, W. W. 1996. Learning trees and rules with set-valued features. In *AAAI/IAAI, Vol. 1,* 709–716.

Eco, U. 2000. *Como Se Faz uma Tese.* São Paulo, SP: Ed. Perspectiva.

Feitosa, V. C. 1991. *Redação de Textos Científicos.* Campinas, SP: Ed. Papirus.

Fischer, G.; Lemke, A. C.; Mastaglio, T. W.; and Mørch, A. I. 1991. Critics: An emerging approach to knowledge-based human-computer interaction. *International Journal of Man-Machine Studies* 35(5):695–721.

Fontana, N. M.; Caldeira, S. M. A.; De Oliveira, M. C. F.; and Oliveira Jr., O. N. 1993. Computer assisted writing — aplications to english as a foreign language. *CALL* 6(2):145–161.

Hartley, A., and Paris, C. 1996. Multilingual document production: From support for translating to support for authoring. ITRI Technical Report Series ITRI-96-17, ITRI - University of Brighton, Brighton, UK.

Hayes, J., and Flower, L. S. 1980. Writing as problem solving. *Visible Language* XIV(4).

Hegenber, L. 1976. *Etapas da Investigação Científica.* São Paulo, SP: Ed. Epu.

Huckin, T. N., and Olsen, L. A. 1991. *Technical Writing and Professional Communication For Nonnative Speakers of English.* New York, USA: McGraw-Hill.

Kriegsman, M., and Barletta, R. 1993. Building a case-based help desk application. *IEEE Expert: Intelligent Systems and Their Application* 8(6):18–26.

Kupiec, J.; Pedersen, J. O.; and Chen, F. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95)*, 68–73.

Lansman, M.; Smith, J. B.; and Weber, I. 1993. Using the writing environment to study writer´s strategies. *Computers and Composition* 10(2):71–92.

Lewis, D. D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In Nédellec, C., and Rouveirol, C., eds., *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, 4–15. Chemnitz, DE: Springer Verlag, Heidelberg, DE.

Mani, I. 2001. *Automatic Summarization.* Amsterdam, The Netherlands: John Benjamins Publishing Company.

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

Manning, C. D., and Schütze, H. 2001. *Foundations of Statistical Natural Language Processing.* Cambridge, Mass.: The MIT Press.

Mantaras, R. L., and Plaza, E. 1997. Case-based reasoning: An overview. *AI Communications Journal* 10(1):21–29.

Marcu, D. 2000. *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge, Mass.: The MIT Press.

Martins, R. T.; Hasegawa, R.; Nunes, M. d. G. V.; Montilha, G.; and Oliveira Jr., O. N. 1998. Linguistic issues in the development of regra: a grammar checker for brazilian portuguese. *Natural Language Engineering* 4:287–307.

McCallum, A. K. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

Mikheev, A. 1998. Feature lattices for maximum entropy modelling. In *COLING-ACL*, 848–854.

Mizuta, Y., and Collier, N. 2004. An annotation scheme for a rhetorical analysis of biology articles. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*.

Motta-Roth, D., and Hendges, G. 1998. Uma análise transdiciplinar do gênero abstract. *Revista Intercâmbio* VII:125–134.

Narita, M. 2000a. Constructing a tagged e-j parallel corpus for assisting japanese software engineers in writing english abstracts. In *Proceedings of Second International Conference on Language Resources and Evaluation (LREC'2000)*, 1187–1191.

Narita, M. 2000b. Corpus-based english language assistant to japanese software engineers. In *Proceedings of MT-2000 Machine Translation and Multilingual Applications in the New Millennium*, 24–1–24–8.

Oliveira Jr., O. N.; Caldeira, S. M. A.; and Fontana, N. M. 1992. Chusaurus: A writing tool resource for non-native users of english. In Baeza-Yates, R., and Manber, U., eds., *Computer Science: Research and Application*. New York, USA: Plenum Press. 63–72.

Oliveira, S. L. 2001. *Tratado de Metodologia Científica*. São Paulo, SP: Pioneira.

Paris; C., and Evans, R. 1994. Preliminary proposal for the gist architecture.

Pemberton, L.; Shurville, S.; and Hartley, T. 1996. Motivating the design for a computer assisted environment for writers in a second language. *Lecture Notes in Computer Science* 1108:141–148.

Ratnaparkhi, A. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*.

Rey, L. 1972. *Como Redigir Trabalhos Científicos*. São Paulo, SP: Ed. da Universidade de São Paulo.

Rodrigues, B. B. 1998. *Estratégias de condução de informações em resumos de dissertações*. Ph.D. Dissertation, Programa de Pós-Graduação em Letras/Lingüística, UFSC.

RSI, R. S. I. 1992. The n.1 grammar & style checker gram.mat.ik 5 (for windows) - user's guide.

Salomon, D. V. 1977. *Como Fazer uma Monografia: Elementos de Metodologia do Trabalho Científico*. Belo Horizonte, MG: Ed. Interlivros.

Santos, M. B. d. 1996. The textual organisation of research paper abstracts. *Text* 16(4):481–499.

Severino, A. J. 1996. *Metodologia do Trabalho Científico*. São Paulo, SP: Cortez Editora.

Sharples, M., and Pemberton, L. 1992. Representing writing: external representations and the writing process. In Holt, P., and Williams, N., eds., *Computers and Writing: State of the Art*. Oxford, UK: Intellect. 319–336.

Sharples, M.; Goodlet, J.; and Clutterbuck, A. 1994. A comparison of algorithms for hypertext notes network linearization. *International Journal of Human-Computer Studies* 40(4):727–752.

Shurville, S.; Hartley, A. F.; and Pemberton, L. 1997. A development methodology for composer: a computer support tool for academic writing in a second language. In Jakobs, E.-M., and Knorr, D., eds., *Textproduktion in elektronischen Umgebungen [Text Production in Electronic Environments]*. Frankfurt, Germany: Peter Lang Verlag. 171–182.

Siegel, S., and Castellan, N. J. J. 1988. *Nonparametric Statistics for the Behavioral Sciences*. Berkeley, CA, USA: McGraw-Hill, 2nd edition.

Silva, M. H. B.; Pelizzoni, J. M.; and Aluísio, S. M. 1998. Uma abordagem híbrida baseada em críticas e casos para a construção de ferramentas colaborativas de ensino da escrita de artigos científicos. In *Anais do IX Simpósio Brasileiro de Informática na Educação*.

Silva, M. H. B. 1999. A abordagem de críticas para a construção de sistemas de aprendizado da escrita técnica. Master's thesis, Instituto de Ciências Matemáticas e de Computação - USP.

Silva, E. C. F. d. 2002. A estrutura genérica e as escolhas léxico-gramaticais em introduções de dissertação de mestrado na área de lingüística. In *Caderno de Resumos do 12o InPLA*, 235–236. São Paulo, SP: LAEL/PUC-SP.

Smith, J. B., and Lansman, M. 1988. A cognitive basis for a computer writing experiment. Technical Report 87-032, Chapel Hill.

Swales, J. 1990. *Genre Analysis: English in Academic and Research Settings. Chapter 7: Research articles in English*. Cambridge, UK: Cambridge University Press. 110–176.

Tachizawa, T., and Mendes, G. 2000. *Como Fazer Monografia na Prática*. Rio de Janeiro, RJ: Ed. Fundação Getúlio Vargas.

Taylor, G., and Tingguang, C. 1991. Linguistic, cultural, and subcultural issues in contrastive discoure analysis: Anglo-american and chinese scientific texts. *Applied Linguistics* 12(3):319–336.

Teufel, S., and Moens, M. 2000. What's yours and what's mine: Determining intellectual attribution in scientific text. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Teufel, S., and Moens, M. 2002. Summarising scientific articles — experiments with relevance and rhetorical status. *Computational Linguistics* 28(4):409–446.

Teufel, S.; Carletta, J.; and Moens, M. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99)*, 110–117.

Teufel, S. 1999. *Argumentative Zoning: Information Extraction from Scientific Text.* Ph.D. Dissertation, School of Cognitive Science, University of Edinburgh, Edinburgh, UK.

Trimble, L. 1985. *English for science and technology: a discourse approach.* Cambridge, UK: Cambridge University Press.

Vander Linden, K., and Scott, D. 1995. Raising the interlingual ceiling with multilingual text generation. ITRI Technical Report Series ITRI-95-12, ITRI - University of Brighton, Brighton, UK.

Vieira, S. 1999. *Como Escrever uma Tese.* São Paulo, SP: Ed. Pioneira.

W3C. 2001. Xml schema part 0: Primer.

Weissberg, R., and Buker, S. 1990. *Writing up Research: Experimental Research Report Writing for Students of English.* Prentice Hall.

WTG, W. T. G. I. 1991. Using correct grammar for windows.

Apêndice

A

Questionário Aplicado aos Potenciais Usuários de uma Ferramenta de Auxílio à Escrita Acadêmica

Este questionário tem por objetivo o levantamento de requisitos para um sistema computacional de auxílio à escrita acadêmica. Um texto acadêmico geralmente possui uma estrutura padronizada. Nessa estrutura, entendemos por seções/capítulos o Resumo, Abstract, Introdução, Revisão Bibliográfica, Metodologia, Resultados e Conclusões, sendo que cada um deles pode ser ainda subdividido em componentes com funções específicas (p.e., os componentes de uma introdução podem ser contextualização, motivação e relevância, propósitos, entre outros). Suas respostas não serão divulgadas individualmente, pois os questionários serão avaliados por grupos de usuários. Qualquer dúvida entre em contato com Valéria D. Feltrim (vfeltrim@icmc.usp.br).

1. Como você classificaria o seu conhecimento sobre a redação de textos acadêmicos (p.e. relatório técnico, artigo, monografias de qualificação, dissertação, tese, projeto científico) quanto à estrutura geral das seções/capítulos, seus componentes e suas funções?
   ( ) Excelente ( ) Bom ( ) Razoável ( ) Pouco ( ) Insuficiente

2. Quantos textos desse tipo você já produziu individualmente?
   ( ) Nenhum ( ) Até 5 ( ) Até 10 ( ) Mais que 10

3. Considerando as seguintes seções de um texto acadêmico típico, assinale seu nível de dificuldade ao escrevê-las. Use uma escala de 1 a 5 para indicar o grau de dificuldade, sendo que 1 indica maior dificuldade.

( ) Resumo ( ) Abstract ( ) Introdução ( ) Revisão Bibliográfica ( ) Metodologia
( ) Resultados ( ) Conclusão

4. Quantas revisões, além das suas, são feitas em média em seus textos acadêmicos até que você os considere prontos?

( ) Nenhuma ( ) 1 ( ) 2 ( ) Outro. Neste caso, indique o número.

5. Qual o grau de interferência de alguém mais experiente (p.e. seu orientador) na produção de seus textos acadêmicos?

( ) Nenhuma ( ) Pouca ( ) Significativa ( ) Muita

6. Você se baseia em algum modelo de texto para produzir o seu?

( ) Não ( ) Sim. Neste caso, qual?
( ) Texto anterior seu
( ) Texto de um colega
( ) Texto anterior de seu orientador ou de quem foi orientado por ele
( ) Texto acadêmico da literatura nacional
( ) Texto acadêmico da literatura em inglês

Considere as características abaixo (7 a 18) que poderiam estar presentes em uma ferramenta computacional para auxiliar a escrita de um texto acadêmico. Utilize a seguinte escala para classificar cada uma das características:

1. Indispensável: sem essa característica não valeria a pena usar a ferramenta
2. Desejável: a presença da característica motivaria o uso da ferramenta
3. Interessante: a característica é interessante, mas não influenciaria sua motivação em usar a ferramenta
4. Dispensável: a ferramenta não perderia valor sem essa característica

7. ( ) Apoio à organização das idéias através de brainstorming livre (sem interferência do sistema), por exemplo, fornecendo um espaço onde o usuário possa escrever suas idéias e relacioná-las (p.e. um grafo com nós e arcos).

8. ( ) Apoio à organização das idéias através de brainstorming dirigido (p.e., através de perguntas que estimulem a geração das idéias iniciais) com o foco nos principais componentes de um texto acadêmico.

9. ( ) Disponibilidade de explicações sobre o gênero acadêmico, indicando o papel de cada componente de uma seção/capítulo.

10. ( ) Disponibilidade de modelos de textos acadêmicos, indicando os possíveis componentes e a ordem em que aparecem.

11. ( ) Disponibilidade de exemplos autênticos de bons textos acadêmicos que instanciem os modelos.

12. ( ) Disponibilidade de guidelines sobre como escrever bem cada componente.

13. ( ) Disponibilidade de padrões de escrita para cada componente, i.e., expressões comumente utilizadas em um determinado componente.

14. ( ) Alerta do sistema sobre erros comumente encontrados na escrita de cada componente.

15. ( ) Disponibilidade de uma checklist para que o usuário possa verificar se o seu texto apresenta as principais características de um bom texto acadêmico.

16. ( ) Apoio à correção ortográfica.

17. ( ) Apoio à correção gramatical e de estilo.

18. ( ) Revisão do texto depois de pronto com sugestões de melhoria.

19. Existem características que não foram listadas aqui que você considera importantes? Em caso afirmativo, faça sua sugestão sem se preocupar com a viabilidade da implementação.

20. Assinale qual o estilo de interação com o usuário de sua preferência para a ferramenta para auxiliar a escrita de um texto acadêmico.

( ) Um ambiente tutor, sob o controle do sistema, que especifica uma seqüência de passos para a escrita do texto.

( ) Um ambiente livre, em que o usuário aciona as funções do sistema (solicitando ajuda) quando achar necessário.

( ) Um ambiente misto, em que o usuário é livre para acionar funções do sistema e o sistema também interfira quando julgar necessário.

( ) Nenhuma das anteriores. Neste caso, faça sua sugestão.

21. Indique a sua motivação em usar a ferramenta para auxiliar a escrita de um texto acadêmico acoplada ao seu processador de texto.

( ) Não a usaria caso não fosse acoplada ao editor de texto

( )Usaria mesmo não sendo acoplada ao editor de texto, desde que o formato dos arquivos sejam compatíveis.

( ) Usaria mesmo não sendo acoplada ao editor de texto, pois essa característica é irrelevante.

( ) Não gostaria de uma ferramenta acoplada ao editor de texto.

Apêndice

B

XML Schema para a Anotação do CorpusDT

```
<!- edited with XML Spy v4.0.1 U (http://www.xmlspy.com) by Valéria Feltrim (USP - NILC) ->
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">

    <xsd:annotation>
        <xsd:documentation xml:lang="pt">
            XML Schema para anotação de teses e dissertações (01-09-01)
            Copyright 2001 Valeria D. Feltrim. All rights reserved.
        </xsd:documentation>
    </xsd:annotation>

    <xsd:element name="Dissertacao" type="TipoDissertacao"/>

    <xsd:complexType name="TipoDissertacao">
        <xsd:sequence>
            <xsd:element name="PaginaRosto" type="TipoPaginaRosto"/>
            <xsd:element name="ElemIniciais" type="TipoElemIniciais" minOccurs="0"/>
            <xsd:element name="CorpoDissertacao" type="TipoCorpoDissertacao"/>
            <xsd:element name="RefBibliografica" type="xsd:string" minOccurs="0"/>
            <xsd:element name="Bibliografia" type="xsd:string" minOccurs="0"/>
            <xsd:element name="Anexo" type="TipoAnexo" minOccurs="0" maxOccurs="unbounded"/>
            <xsd:element name="Apendice" type="TipoApendice" minOccurs="0" maxOccurs="unbounded"/>
            <xsd:element name="Rodape" type="xsd:string" minOccurs="0" maxOccurs="unbounded"/>
        </xsd:sequence>
        <xsd:attribute name="Nivel" type="TipoNivel" use="required"/>
        <xsd:attribute name="Area" type="TipoArea" use="required"/>
    </xsd:complexType>
```

```
<xsd:simpleType name="TipoArea">
    <xsd:restriction base="xsd:string">
        <xsd:enumeration value="ES"/>
        <xsd:enumeration value="IC"/>
        <xsd:enumeration value="BD"/>
        <xsd:enumeration value="HIP"/>
        <xsd:enumeration value="CGPI"/>
        <xsd:enumeration value="SDG"/>
        <xsd:enumeration value="SDPC"/>
    </xsd:restriction>
</xsd:simpleType>

<xsd:simpleType name="TipoNivel">
    <xsd:restriction base="xsd:string">
        <xsd:enumeration value="MS"/>
        <xsd:enumeration value="DO"/>
    </xsd:restriction>
</xsd:simpleType>

<xsd:complexType name="TipoPaginaRosto">
    <xsd:all>
        <xsd:element name="Instituicao" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Titulo" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Autor" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Orientador" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Comentario" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Local" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Data" type="xsd:string" minOccurs="0"/>
    </xsd:all>
</xsd:complexType>

<xsd:complexType name="TipoElemIniciais">
    <xsd:all>
        <xsd:element name="Citacao" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Dedicatoria" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Agradecimentos" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Sumario" type="xsd:string" minOccurs="0"/>
        <xsd:element name="ListadeFigs" type="xsd:string" minOccurs="0"/>
        <xsd:element name="ListadeTabs" type="xsd:string" minOccurs="0"/>
        <xsd:element name="ListadeGraficos" type="xsd:string" minOccurs="0"/>
        <xsd:element name="ListadeAbrev" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Resumo" type="TipoResumo" minOccurs="0"/>
        <xsd:element name="Abstract" type="TipoResumo" minOccurs="0"/>
        <!- "Citacao" pode ser uma frase, uma poesia, um conto, etc ->
    </xsd:all>
</xsd:complexType>

<xsd:complexType name="TipoAnexo" mixed="true">
    <xsd:sequence>
        <xsd:element name="Elemento" type="TipoElemento" minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>

<xsd:complexType name="TipoApendice" mixed="true">
    <xsd:sequence>
        <xsd:element name="Elemento" type="TipoElemento" minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>
```

```
<xsd:complexType name="TipoCorpoDissertacao">
    <xsd:sequence>
        <xsd:element name="Introducao" type="TipoIntroducao"/>
        <xsd:element name="Capitulo" type="TipoCapitulo" maxOccurs="unbounded"/>
        <xsd:element name="Conclusão" type="TipoConclusao"/>
    </xsd:sequence>
</xsd:complexType>

<xsd:complexType name="TipoIntroducao">
    <xsd:sequence>
        <xsd:element name="Titulo" type="xsd:string"/>
        <xsd:element name="SecaoI" type="TipoSecaoI" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>

<xsd:complexType name="TipoConclusao">
    <xsd:sequence>
        <xsd:element name="Titulo" type="xsd:string"/>
        <xsd:element name="SecaoC" type="TipoSecaoC" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>

<xsd:complexType name="TipoCapitulo">
    <xsd:sequence>
        <xsd:element name="Titulo" type="xsd:string"/>
        <xsd:element name="Secao" type="TipoSecao" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>

<xsd:complexType name="TipoSecaoI">
    <xsd:sequence>
        <xsd:element name="Titulo" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Bloco" type="TipoBloco" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element name="SubsecaoI" type="TipoSecaoI" minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>

<xsd:complexType name="TipoSecaoC">
    <xsd:sequence>
        <xsd:element name="Titulo" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Bloco" type="TipoBloco" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element name="SubsecaoC" type="TipoSecaoI" minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>

<xsd:complexType name="TipoSecao" mixed="true">
    <xsd:sequence>
        <xsd:element name="Titulo" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Elemento" type="TipoElemento" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element name="Subsecao" type="TipoSecao" minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>
```

```
<xsd:complexType name="TipoElemento">
    <xsd:choice>
        <xsd:element name="Figura" type="xsd:string"/>
        <xsd:element name="Tabela" type="xsd:string"/>
        <xsd:element name="Grafico" type="xsd:string"/>
        <xsd:element name="Quadro" type="xsd:string"/>
        <xsd:element name="Equacao" type="xsd:string"/>
        <xsd:element name="Rodape" type="xsd:string"/>
    </xsd:choice>
</xsd:complexType>

<xsd:complexType name="TipoBloco">
    <xsd:choice>
        <xsd:element name="Componente" type="TipoComponente"/>
        <xsd:element name="Figura" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Tabela" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Grafico" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Quadro" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Equacao" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Rodape" type="xsd:string" minOccurs="0"/>
    </xsd:choice>
</xsd:complexType>

<xsd:complexType name="TipoComponente">
    <xsd:sequence>
        <xsd:element name="Nome" type="xsd:string"/>
        <xsd:element name="Estrategia" type="TipoEstrategia" minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>

<xsd:complexType name="TipoEstrategia" mixed="true">
    <xsd:sequence>
        <xsd:element name="Nome" type="xsd:string"/>
        <xsd:element name="Reutilizavel" type="TipoReutilizavel" minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>

<xsd:complexType name="TipoReutilizavel" mixed="true">
    <xsd:sequence>
        <xsd:element name="Marcador" type="xsd:string" minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>

<xsd:complexType name="TipoResumo">
    <xsd:sequence>
        <xsd:element name="Componente" type="TipoComponente" maxOccurs="unbounded"/>
    </xsd:sequence>
</xsd:complexType>

</xsd:schema>
```

Apêndice

C

Regras de Crítica Utilizadas no Ambiente *SciPo*

C.1 Regras de Crítica para Resumos

C.1.1 Críticas de conteúdo

Crítica: Faltam componentes essenciais

— Falta *Indicar o propósito principal*
O resumo deve conter o propósito principal! Sem informação, o leitor não poderá identificar qual foi o objetivo do seu trabalho. Acrescente a estratégia *Indicar o propósito principal.*

— Falta *Metodologia*
O resumo deve indicar objetivamente ao leitor a metodologia empregada para a realização do seu trabalho. Acrescente pelo menos uma das 3 estratégias de metodologia. Escolha a que for mais adequada ao seu resumo.

— Falta *Resultado*
O resumo deve apresentar ao leitor os principais resultados do seu trabalho. Acrescente a estratégia *Descrever o artefato* caso você queira enfocar o desenvolvimento de um artefato computacional como o principal resultado de seu trabalho. Acrescente a estratégia *Apresentar resultados* caso seus principais resultados sejam outros.

Crítica: Remova componentes desnecessários

— Remova o Outline
Deixe para apresentar a estrutura do seu trabalho na introdução. Utilize o resumo para apresentar informações mais relevantes.

Crítica: Repetição imprópria de componentes

— *Indicar o propósito principal* aparece mais de uma vez
O propósito principal do seu trabalho deve ser enunciado uma única vez e de forma clara e objetiva. Portanto, deixe apenas uma ocorrência da estratégia *Indicar o propósito principal* em sua estrutura. Caso queira fornecer mais informações sobre o propósito, acrescente a estratégia *Detalhar/Especificar o propósito.* Caso você queira acrescentar outros propósitos além do principal (propósitos secundários), utilize a estratégia *Introduzir mais propósitos.*

— *Introduzir a pesquisa a partir da grande área* aparece mais de uma vez
A estratégia *Introduzir a pesquisa a partir da grande área* fornece uma contextualização bastante abrangente e dispensa sua repetição. Portanto, é melhor que você utilize apenas uma ocorrência dessa estratégia em sua estrutura.

C.1.2 Críticas de ordem

Crítica: Reorganize a estrutura

— *Detalhar/Especificar o propósito* ou *Introduzir mais propósitos* aparece antes de *Indicar o propósito principal* ou do componente Propósito::*
A estratégia *Indicar o propósito principal* deve enunciar o objetivo principal do seu trabalho, por isso deve ser a primeira estratégia de propósito a aparecer no resumo. As estratégias *Detalhar/Especificar o propósito* e *Introduzir mais propósitos* são estratégias complementares e devem aparecer depois do propósito principal. Reorganize essas estratégias.

— O *Resultado* aparece antes do *Propósito*
Antes de dizer quais foram os resultados do seu trabalho, você deve dizer quais eram seus objetivos. Por isso, a estratégia *Indicar o propósito principal* deve aparecer antes de qualquer uma das estratégias de Resultado. Reorganize sua estrutura para que a estratégia *Indicar o propósito principal* venha antes das estratégias de Resultado.

— A *Metodologia* aparece antes do *Propósito*
Antes de dizer os métodos utilizados no seu trabalho é preciso que o leitor saiba quais eram seus objetivos. Por isso, a estratégia *Indicar o propósito principal* deve vir antes das estratégias de Metodologia. Reorganize sua estrutura para que a estratégia *Indicar o propósito principal* apareça antes das estratégias de Metodologia.

— A *Conclusão* aparece antes do *Propósito*
As estratégias de Conclusão servem para dar um "fecho" ao resumo. Por isso, é mais natural que elas apareçam no final. Antes de fazer qualquer conclusão, você precisa dizer ao leitor quais eram os objetivos do seu trabalho. Reorganize sua estrutura de forma que a estratégia *Indicar o propósito principal* apareça antes das estratégias de Conclusão.

— A *Conclusão* aparece antes da *Metodologia* ou do *Resultado*
As estratégias de Conclusão servem para dar um "fecho" ao resumo. Por isso, é mais natural que elas apareçam no final. Antes de fazer alguma conclusão, é melhor dizer quais foram seus métodos e seus resultados. Reorganize sua estrutura para que informações sobre a metodologia ou sobre os resultados apareçam antes das conclusões.

— A contextualização *Introduzir a pesquisa a partir da grande área* aparece depois do *Propósito*
A estratégia *Introduzir a pesquisa a partir da grande área* fornece uma contextualização bastante abrangente e tem por objetivo guiar o leitor da grande área de pesquisa para a sua área de trabalho mais específica. Essa estratégia prepara o leitor para entender melhor o propósito do trabalho. Reorganize sua estrutura para que a estratégia *Introduzir a pesquisa a partir da grande área* apareça antes da estratégia *Indicar o propósito principal*

— *Comentar/Discutir resultados* aparece antes de *Apresentar resultados* e/ou *Descrever o artefato*
Antes de discutir seus resultados você deve apresentá-los ao seu leitor. Por isso, a estratégia *Comentar/Discutir resultados* deve aparecer depois da estratégia *Apresentar resultados* ou então depois da estratégia *Descrever o artefato*. Reorganize suas estratégias de Resultado.

— Algum *Contexto* aparece depois do *Propósito*
As estratégias de *Contexto* servem para familiarizar o leitor com sua área de pesquisa e para ressaltar sua relevância. Em geral, essas estratégias aparecem no início do resumo e precedem a estratégia *Indicar o propósito principal*. Em sua estrutura existe uma estratégia de *Contexto* após a estratégia *Indicar o propósito principal*. Reorganize sua estrutura para que toda contextualização seja feita antes da indicação do propósito principal.

— Alguma *Lacuna* aparece depois do *Propósito*
As estratégias de *Lacuna* servem para indicar alguma lacuna numa área de pesquisa que vale a pena ser investigada. Dessa forma, o mais comum é que elas apareçam juntamente com a contextualização e que precedam a estratégia *Indicar o propósito principal*. Em sua estrutura existe uma ou mais estratégias de *Lacuna* após a estratégia *Indicar o propósito principal*. Reorganize sua estrutura para que as lacunas indicadas precedam o propósito principal.

C.1.3 Sugestões de conteúdo

Sugestão: Tente enriquecer sua estrutura

— Adicionando *Contexto* e *Lacuna*
Que tal acrescentar *Contexto* e *Lacuna*? As estratégias de contextualização ajudam o leitor a entender melhor o escopo do seu trabalho. Com essas estratégias você pode convencer o leitor de que sua área de pesquisa é relevante e merece ser explorada. Vale a pena acrescentar alguma estratégia de *Contexto*! A contextualização também prepara o leitor para compreender melhor o problema que você pretende atacar em sua pesquisa. Esse problema é enunciado através das estratégias de *Lacuna*. A declaração de uma lacuna em sua área de pesquisa complementa a contextualização e motiva o propósito do seu trabalho. Escolha as estratégias mais adequadas ao seu resumo.

— Adicionando *Contexto*
Você colocou uma ou mais estratégias de *Lacuna* na sua estrutura e isso é bom, pois apontando uma ou mais lacunas numa área de pesquisa você justifica os propósitos do seu trabalho. No entanto, você não usou nenhuma estratégia de *Contexto*. Apontar uma lacuna a ser preenchida motiva o propósito do seu trabalho, mas contextualizar essa lacuna em uma área de pesquisa também é importante. Por isso, estratégias de *Lacuna* geralmente são acompanhadas de estratégias de *Contexto*. Escolha as estratégias de contexto que sejam mais adequadas ao seu resumo.

— Adicionando *Lacuna*
Você utilizou uma ou mais estratégias de *Contexto*, o que contribuirá para que o leitor entenda em que área de pesquisa seu trabalho se situa. No entanto, você não utilizou nenhuma estratégia de *Lacuna*. Indicando uma lacuna numa linha de pesquisa, você pode justificar seus objetivos e convencer o leitor da relevância do seu trabalho. Por isso, é comum que estratégias de *Lacuna* acompanhem estratégias de *Contexto*. Escolha uma das estratégias de *Lacuna* que seja mais adequada ao seu resumo.

— Adicionando *Conclusão*
Você não utilizou nenhuma estratégia de *Conclusão* em sua estrutura. As estratégias de *Conclusão* podem ser utilizadas para dar um "fecho" ao resumo e, dessa forma, torná-lo um texto mais completo. Em um resumo de dissertação/tese, é especialmente bom destacar as contribuições do trabalho. Para isso, acrescente a estratégia *Apresentar contribuições/-valor do trabalho.*

C.1.4 Sugestões de ordem

Sugestão: Tente melhorar a organização da sua estrutura

— A *Lacuna* aparece intercalada entre *Propósitos*
As estratégias de *Lacuna* servem para indicar alguma lacuna numa área de pesquisa que vale a pena ser investigada. Dessa forma, o mais comum é que elas apareçam juntamente com a contextualização e que precedam a estratégia *Indicar o propósito principal.* Em sua estrutura existe uma ou mais estratégias de *Lacuna* após a estratégia *Introduzir mais propósitos.* Você pode estar utilizando essa lacuna para introduzir a estratégia *Introduzir mais propósitos.* Tome cuidado. A intercalação dessas estratégias pode confundir o leitor em relação ao seu propósito principal. Utilize marcadores que reforcem a distinção entre seus propósitos principais e seus propósitos secundários.

— O *Contexto* (exceto *Familiarização*) aparece intercalado entre *Propósitos*
As estratégias de *Contexto* servem para familiarizar o leitor com sua área de pesquisa e para ressaltar sua relevância. Em geral, essas estratégias aparecem no início do resumo e precedem a estratégia *Indicar o propósito principal.* Em sua estrutura existe uma estratégia de *Contexto* após a estratégia *Indicar o propósito principal.* Você pode estar utilizando essa nova contextualização para introduzir a estratégia *Introduzir mais propósitos.* Tome cuidado. A retomada da contextualização pode confundir o leitor em relação ao seu propósito principal. Utilize marcadores que reforcem a distinção entre seus propósitos principais e seus propósitos secundários.

C.2 Regras de Crítica para Introduções

C.2.1 Críticas de conteúdo

Crítica: Faltam componentes essenciais

— Falta *Contexto*
As estratégias de contextualização ajudam o leitor a entender melhor o escopo do seu trabalho. Com essas estratégias você pode convencer o leitor de que sua área de pesquisa é relevante e merece ser explorada. Além disso, a contextualização fornece ao leitor o *background* necessário para o entendimento do tópico particular da sua pesquisa em relação a uma área de estudo geral. Acrescente alguma estratégia de *Contexto*. Escolha a que achar mais adequada a sua introdução.

— Falta *Revisão da literatura*
A revisão da literatura dá continuidade ao processo iniciado pela contextualização, de dar ao leitor informação de *background* necessária para entender o estudo em foco no trabalho. A revisão da literatura também garante ao leitor que você está familiarizado com as pesquisas relevantes que têm sido publicadas na sua área de pesquisa. Portanto, insira alguma estratégia de *Revisão da literatura* em sua introdução. Escolha a estratégia que lhe parecer mais adequada.

— Falta *Lacuna*
As estratégias de *Lacuna* servem para indicar alguma lacuna numa área de pesquisa que vale a pena ser investigada, ou seja, um tópico importante que não foi investigado por outros autores ou que não tenha sido suficientemente desenvolvido. Indicando uma lacuna, você pode justificar seus objetivos e convencer o leitor da relevância do seu trabalho. Por isso, é comum que estratégias de *Lacuna* acompanhem estratégias de *Contexto*. Escolha uma das estratégias de *Lacuna* que seja mais adequada a sua introdução.

— Falta *Indicar o propósito principal*
A introdução deve conter o propósito principal. Sem essa informação, o leitor não poderá identificar qual foi o objetivo do seu trabalho. Acrescente a estratégia *Indicar o propósito principal*.

— Falta *Metodologia*
A introdução deve indicar objetivamente ao leitor a metodologia empregada para a realização do seu trabalho. Acrescente pelo menos uma das três estratégias de metodologia. Escolha a que for mais adequada a sua introdução.

— Falta *Resultado*
A introdução deve apresentar ao leitor os principais resultados do seu trabalho. Acrescente a estratégia *Descrever o artefato* caso você queira enfocar o desenvolvimento de um artefato computacional como o principal resultado de seu trabalho. Acrescente a estratégia *Apresentar resultados* caso seus principais resultados sejam outros. Acrescente também a estratégia *Comentar/Discutir resultados* se achar apropriado.

Crítica: Repetição imprópria de componentes

— *Indicar o propósito principal* aparece mais de uma vez
O propósito principal do seu trabalho deve ser enunciado uma única vez e de forma clara e objetiva. Portanto, deixe apenas uma ocorrência da estratégia *Indicar o propósito principal* em sua estrutura. Caso queira fornecer mais informações sobre o propósito, acrescente a estratégia *Detalhar/Especificar o propósito*. Caso você queira acrescentar outros propósitos além do principal (propósitos secundários), utilize a estratégia *Introduzir mais propósitos*.

— *Introduzir a pesquisa a partir da grande área* aparece mais de uma vez
A estratégia *Introduzir a pesquisa a partir da grande área* fornece uma contextualização bastante abrangente e dispensa sua repetição. Portanto, é melhor que você deixe apenas uma ocorrência dessa estratégia em sua estrutura.

C.2.2 Críticas de ordem

Crítica: Reorganize a estrutura

— *Detalhar/Especificar o propósito* ou *Introduzir mais propósitos* aparece antes do *Propósito*
A estratégia *Indicar o propósito principal* deve enunciar o objetivo principal do seu trabalho, por isso deve ser a primeira estratégia de propósito a aparecer na introdução. As estratégias *Detalhar/Especificar o propósito* e *Introduzir mais propósitos* são estratégias complementares e devem aparecer depois do propósito principal. Reorganize essas estratégias.

— O *Resultado* aparece antes do *Propósito*
Antes de dizer quais foram os resultados do seu trabalho, você deve dizer quais eram seus objetivos. Por isso, a estratégia *Indicar o propósito principal* deve aparecer antes de qualquer uma das estratégias de *Resultado*. Reorganize sua estrutura para que a estratégia *Indicar o propósito principal* venha antes das estratégias de *Resultado*.

— A *Metodologia* aparece antes do *Propósito*
Antes de indicar a metodologia utilizada no seu trabalho é preciso que o leitor saiba quais eram seus objetivos. Por isso, a estratégia *Indicar o propósito principal* deve vir antes das estratégias de *Metodologia*. Reorganize sua estrutura para que a estratégia *Indicar o propósito principal* apareça antes das estratégias de *Metodologia*.

— *Introduzir a pesquisa a partir da grande área* aparece depois do *Propósito*
A estratégia *Introduzir a pesquisa a partir da grande área* fornece uma contextualização bastante abrangente e tem por objetivo guiar o leitor da grande área de pesquisa para a sua área de trabalho mais específica. Essa estratégia prepara o leitor para entender melhor a motivação e o propósito do seu trabalho. Reorganize sua estrutura para que a estratégia *Introduzir a pesquisa a partir da grande área* apareça antes da estratégia *Indicar o propósito principal*.

— *Comentar/Discutir resultados* aparece antes de *Apresentar resultados* e *Descrever o artefato*
Antes de discutir ou comentar seus resultados você deve apresentá-los ao seu leitor. Por isso, a estratégia *Comentar/discutir resultados* deve aparecer depois da estratégia *Apresentar resultados* ou então depois da estratégia *Descrever o artefato*. Reorganize suas estratégias de *Resultado*.

— Algum *Contexto* aparece depois do *Propósito*
As estratégias de *Contexto* servem para familiarizar o leitor com sua área de pesquisa e para ressaltar a relevância de seu trabalho. Em geral, essas estratégias aparecem no início da introdução e precedem a estratégia *Indicar o propósito principal*. Em sua estrutura existe uma ou mais estratégias de *Contexto* após a estratégia *Indicar o propósito principal*. Reorganize sua estrutura para que toda contextualização seja feita antes da indicação do propósito principal.

— Alguma *Lacuna* aparece depois do *Propósito*
As estratégias de *Lacuna* servem para indicar alguma lacuna numa área de pesquisa que vale a pena ser investigada. Dessa forma, o mais comum é que elas apareçam juntamente com a contextualização e que precedam a estratégia *Indicar o propósito principal*. Em sua estrutura existe uma ou mais estratégias de *Lacuna* após a estratégia *Indicar o propósito principal*. Reorganize sua estrutura para que as lacunas indicadas precedam o propósito principal.

— Alguma *Revisão da literatura* aparece depois do *Propósito*
A *Revisão da literatura* serve para dar continuidade ao processo iniciado pela contextualização e mostrar ao leitor que você está familiarizado com as pesquisas feitas na sua área. O mais comum é que a revisão da literatura apareça próxima à contextualização e que preceda a estratégia *Indicar o propósito principal.* Reorganize sua estrutura para que a *Revisão da literatura* preceda propósito principal.

— *Contribuições/Valor do trabalho* aparece antes do *Propósito*
Antes de dizer quais foram suas contribuições, você deve dizer quais eram seus objetivos. Por isso, a estratégia *Indicar o propósito principal* deve aparecer antes de qualquer uma das estratégias de *Contribuições/Valor do trabalho.* Reorganize sua estrutura para que a estratégia *Indicar o propósito principal* venha antes das estratégias de *Contribuições/Valor do trabalho.*

— *Indicar os capítulos/seções do trabalho* ou *Apresentar lista dos tópicos abordados* aparece sem ser no final da estrutura
Utilize estratégias de *Estrutura do trabalho* somente no final da introdução, já que essas estratégias servem para descrever resumidamente o que o leitor encontrará a seguir.

C.2.3 Sugestões de conteúdo

Sugestão: Tente enriquecer sua estrutura

— Adicionando *Contribuições/Valor do trabalho*
Você pode destacar na introdução quais foram as principais *Contribuições* do seu trabalho para a área de pesquisa na qual ele se insere. Também pode destacar a importância do seu trabalho, baseado em algum possível valor ou benefício que o estudo possa trazer para outros pesquisadores do campo de pesquisa ou para as pessoas que trabalham em situações práticas relacionadas a esse campo. Principalmente em teses e dissertações, é importante destacar suas contribuições.

— Adicionando *Estrutura do trabalho*
Você pode dar ao leitor um roteiro do que será encontrado no restante do trabalho. Isso o ajudará a conhecer, mais precisamente, o tipo de informação que ele encontrará a partir da introdução.

C.2.4 Sugestões de ordem

Sugestão: Tente melhorar a organização da sua estrutura

— A *Revisão da literatura* aparece depois do *Propósito* e antes de algum outro *Propósito*, *Metodologia* ou *Resultado*
A *Revisão da literatura* serve para dar continuidade ao processo iniciado pela contextualização e mostrar ao leitor que você está familiarizado com as pesquisas feitas na sua área. O mais comum é que a revisão da literatura apareça próxima à contextualização e que preceda a estratégia *Indicar o propósito principal*. Em sua estrutura existe uma ou mais estratégias de *Revisão da literatura* após a estratégia *Indicar o propósito principal*. Tome cuidado. A intercalação de estratégias de *Revisão* com outras estratégias de *Propósito*, *Metodologia* e *Resultado* pode confundir o leitor em relação ao seu propósito principal. Utilize marcadores que reforcem a distinção entre seus propósitos principais e seus propósitos secundários.

— Alguma estratégia de *Contribuições/Valor do trabalho* aparece depois do *Propósito* e antes da *Metodologia* e/ou *Resultado*
Em sua estrutura existe uma ou mais estratégias de *Contribuições/Valor do trabalho* antes de estratégias de *Metodologia* e/ou *Resultado*. O ideal é que você descreva todo seu trabalho antes de indicar *Contribuições*, pois assim o leitor poderá compreendê-las melhor. Tome cuidado para que esse tipo de organização estrutural não comprometa a interpretação do texto. Utilize marcadores que reforcem a distinção entre propósitos principais e propósitos secundários, metodologia e resultados.

— *Indicar estrutura do capítulo/seção* aparece sem ser no início da estrutura
Se você for utilizar essa estratégia para descrever os assuntos tratados na introdução como um todo, ela deve aparecer no início da estrutura. Entretanto, pense se tal estratégia é realmente necessária, uma vez que, em geral, a introdução é uma seção/capítulo único e dispensa explicações de conteúdo.

— Alguma *Lacuna* aparece depois do *Propósito* e antes de algum outro *Propósito*, *Metodologia* ou *Resultado*
As estratégias de *Lacuna* servem para indicar alguma lacuna numa área de pesquisa que vale a pena ser investigada. Dessa forma, o mais comum é que elas apareçam juntamente com a contextualização e que precedam a estratégia *Indicar o propósito principal*. Em sua estrutura existe uma ou mais estratégias de *Lacuna* após a estratégia *Indicar o propósito principal*. Tome cuidado. A intercalação de estratégias de *Lacuna* com outras estratégias de *Propósito*, *Metodologia* e *Resultado* pode confundir o leitor em relação ao seu propósito principal. Utilize marcadores que reforcem a distinção entre seus propósitos principais e seus propósitos secundários.

— Algum *Contexto* (exceto *Familiarização*) aparece depois do *Propósito* e antes de outro *Propósito*

As estratégias de *Contexto* servem para familiarizar o leitor com sua área de pesquisa e para ressaltar sua relevância. Em geral, essas estratégias aparecem no inicio da introdução e precedem a estratégia *Indicar o propósito principal.* Em sua estrutura existe uma ou mais estratégias de *Contexto* após a estratégia *Indicar o propósito principal.* Tome cuidado. A retomada da contextualização pode confundir o leitor em relação ao seu propósito principal. Utilize marcadores que reforcem a distinção entre seus propósitos principais e seus propósitos secundários.

Apêndice

D

Lista de Expressões Regulares Utilizadas pelo *AZPort*

Constantes

NSPC (não-espaço)	[^ \n\t]*
SPC (espaço)	[\t]+
ART	(o|a|os|as|à|às)
SER	(é|são|foi|foram)
TER	(têm|tem|teve|tiveram)
VIR	(vêm|vem|vieram|veio)
ESTAR	(está|estão|esteve|estiveram)
EST	(este|esse|estes|esses|esta|essa|estas|essas)
NEST	(neste|nesse|nestes|nesses|nesta|nessa|nestas|nessas)
DEST	(deste|desse|destes|desses|desta|dessa|destas|dessas)
NOMES	(documento|artigo|doutorado|estudo|estudo{SPC}proposto|mestrado|texto| trabalho{SPC}proposto|trabalho|dissertação|pesquisa|trabalho{SPC}de{SPC}doutorado| trabalho{SPC}de{SPC}mestrado|dissertação{SPC}de{SPC}mestrado| pesquisa{SPC}de{SPC}doutorado|pesquisa{SPC}proposta|pesquisa{SPC}de{SPC}mestrado| projeto{SPC}de{SPC}doutorado|projeto{SPC}de{SPC}mestrado|projeto{SPC}proposto| projeto|tese|tese{SPC}de{SPC}doutorado)
B-CPTO	(benefícios|enormes{SPC}benefícios|grandes{SPC}avanços|grandes{SPC}benefícios| um{SPC}enorme{SPC}benefício|um{SPC}grande{SPC}avanço|

	um{SPC}grande{SPC}benefício)
B-NOMES	(atividade|tarefa|técnica|pesquisa|estudo){NSPC}
B-ADV	(amplamente|largamente|muito|bastante|bem)
B-ADJ	(valios|desafiador|importante|essencia|fundamenta|indispensáve|vita){NSPC}
B-VERBOSCPX	(aplicad|empregad|explorad|usad|utilizad|aceit|estudad){NSPC}
B-CPTO2	(atuais|da{SPC}comunidade|dos{SPC}pesquisadores|do{SPC}mercado|das{SPC}pesquisas| dos{SPC}trabalhos|dos{SPC}usuários)
G-NOMES	(atividade|tarefa|pesquisa|análise|estudo|trabalho){NSPC}
G-NOMES2	(obstáculo|problema|dificuldade|necessidade|complexidade|limitaç|desafio|prejuízo){NSPC}
G-ADJ	(essencia|fundamenta|vita|necessári|preciso|difíc|complex|complicad|desafiador|problemátic| critíc|inviáve|incomplet|inconsistente|imprecis|impraticáve|indispensáve){NSPC}
G-ADJ2	(trivia|fác|simples){NSPC}
G-ADV	(extremamente|muito|bastante|mais|inúmer){NSPC}
M-CPTO	(a{SPC}partir|através|através{SPC}do{SPC}uso|baseando-se| com{SPC}a{SPC}utilização|com{SPC}auxílio|com{SPC}base|com{SPC}o{SPC}auxílio| na{SPC}forma|por{SPC}meio|segundo|usando|usando-se|utilizando|utilizando-se|via)
M-CPTO2	(como{SPC}base|como{SPC}estudos?{SPC}de{SPC}caso|como{SPC}exemplo| como{SPC}fundamento|como{SPC}ponto{SPC}de{SPC}partida|como{SPC}referencial| como{SPC}um{SPC}estudo{SPC}de{SPC}caso)
M-NOMES	(metodologia|técnica|conceito|conceitos{SPC}e{SPC}princípios|método|princípio| modelo){NSPC}
M-INI	(em{SPC}adição|para{SPC}isso|para{SPC}isto|para{SPC}tanto)
M-VERBOS	(adot|empreg|usa|usou|utiliz){NSPC}
M-VERBOSCPX	(avaliad|construíd|desenvolvid|estruturad|feit|implementad|investigad|manipulad| ministrad|modelad|obtid|realizad){NSPC}
M-VERBOSCPX2	(adotad|empregad|usad|utilizad){NSPC}
M-VERBOSCPX3	(base|fundament){NSPC}
P-VERBOS	(introd|test|model|verific|valid|projet|realiz|constr|desenvolve|dedic|explor| concentr|abord|analis|apresent|busc|consist|contribu|defin|descrev|discut|enfatiz|enfoc| estend|estud|focaliz|fornec|implement|indic|investig|mostr|objetiv|procur|prop|question| relat|report|represent|revis|traç|trat|vis|formul){NSPC}
P-VERBOSCPX	(introduzid|testad|modelad|verificad|validad|projetad|realizad|construíd| desenvolvid|explorad|criad|abordad|analisad|apresentad|buscad|definid|descrit|discutid| enfatizad|enfocad|estendid|estudad|feit|focalizad|fornecid|implementad|indicad|investigad| mostrad|propost|questionad|relatad|reportad|representad|revist|traçad|tratad|visad){NSPC}
P-VERBOSIRREG	(faz|fez|tem{SPC}como{SPC}objetivo|tem{SPC}como{SPC}principal{SPC}objetivo| tem{SPC}por{SPC}objetivo|teve{SPC}como{SPC}objetivo| teve{SPC}como{SPC}principal{SPC}objetivo|teve{SPC}por{SPC}objetivo|

possui{SPC}como{SPC}objetivo|possui{SPC}como{SPC}principal{SPC}objetivo|
possui{SPC}por{SPC}objetivo|possuiu{SPC}como{SPC}objetivo|
possuiu{SPC}como{SPC}principal{SPC}objetivo|possuiu{SPC}por{SPC}objetivo)

P-VERBOSIRREGCPX (fazemos|faz-se|fez-se|fizemos|temos{SPC}como{SPC}objetivo|
temos{SPC}como{SPC}principal{SPC}objetivo|temos{SPC}por{SPC}objetivo|
tem-se{SPC}como{SPC}objetivo|tem-se{SPC}como{SPC}principal{SPC}objetivo|
tem-se{SPC}por{SPC}objetivo|teve-se{SPC}como{SPC}objetivo|
teve-se{SPC}como{SPC}principal{SPC}objetivo|teve-se{SPC}por{SPC}objetivo|
tivemos{SPC}como{SPC}objetivo|tivemos{SPC}como{SPC}principal{SPC}objetivo|
tivemos{SPC}por{SPC}objetivo)

R-NOMES (da{SPC}análise|da{SPC}pesquisa|do{SPC}estudo|do{SPC}trabalho|obtidos?)

R-NOMES2 (análise|estudo|pesquisa|trabalho|obtidos?)

R-VERBOS (apresent|demonstr|evidenci|mostr|permit|suger){NSPC}

R-VERBOSCPX (analisar|demonstrar|determinar)

R-VERBOSCPX2 (analisar|demonstrar|determinar)

R-VERBOSCPX3 (analisad|implementad|testad|modelad|projetad|criad|construíd|estudad|feit|
investigad|explorad|computad|verificad|validad|definid|desenvolvid|realizad){NSPC}

C-INI (a{SPC}conclusão|a{SPC}contribuição|a{SPC}maior{SPC}contribuição|
a{SPC}maior{SPC}virtude|a{SPC}originalidade|a{SPC}outra{SPC}conclusão|
a{SPC}outra{SPC}contribuição|a{SPC}principal{SPC}conclusão|
a{SPC}principal{SPC}contribuição|a{SPC}relevância|
a{SPC}relevância{SPC}e{SPC}originalidade|as{SPC}conclusões|as{SPC}contribuições|
as{SPC}maiores{SPC}contribuições|as{SPC}principais{SPC}conclusões|
as{SPC}principais{SPC}contribuições)

C-INI2 (cabe{SPC}salientar{SPC}que|como{SPC}conclusão|como{SPC}conclusões|
como{SPC}contribuição|como{SPC}contribuições|como{SPC}principais{SPC}conclusões|
como{SPC}principais{SPC}contribuições|como{SPC}principal{SPC}conclusão|
como{SPC}principal{SPC}contribuição|concluiu-se{SPC}que)

C-VERBOS (apresenta contribuições|contribu{NPSC}|também{SPC}contribu{NSPC})

C-VERBOS2 (contribuir|melhorar|auxiliar)

Expressões de Contexto

{NEST}{SPC}contexto{SPC}é{SPC}que{SPC}se{SPC}(encontra|apresenta)
{SER}{SPC}{B-NOMES}{SPC}{B-VERBOSCPX}
{SER}{SPC}{B-NOMES}{SPC}que{SPC}{TER}{SPC}sido{SPC}{B-VERBOSCPX}
{SER}{SPC}comprovada{SPC}a{SPC}importância
{SER}{SPC}considerad{NSPC}{SPC}{B-ADJ}
{SER}{SPC}considerad{NSPC}{SPC}um{ART}{SPC}{B-ADJ}
{SER}{SPC}de{SPC}grande{SPC}(relevância|importância)
{SER}{SPC}de{SPC}{B-ADJ}{SPC}importância
{SER}{SPC}fator(es)?{SPC}motivador(es)?
{SER}{SPC}{B-VERBOSCPX}{SPC}para{SPC}solucionar
{SER}{SPC}{B-ADV}{SPC}{B-VERBOSCPX}
{TER}{SPC}apostado{SPC}cada{SPC}vez{SPC}mais
{TER}{SPC}(sido{SPC})?enfatizado
{TER}{SPC}ganbado{SPC}destaque
{TER}{SPC}se{SPC}(aberto|expandido|tornado)

{TER}{SPC}sido{SPC}{B-VERBOSCPX}{SPC}com{SPC}sucesso
{TER}{SPC}sido{SPC}({B-ADV}{SPC})?{B-VERBOSCPX}
{TER}{SPC}sido{SPC}alvo{SPC}de{SPC}investigação
{TER}{SPC}sido{SPC}conduzid{NSPC}{SPC}pela{SPC}comunidade
{TER}{SPC}sido{SPC}{ART}{SPC}objeto{SPC}de{SPC}estudo
{TER}{SPC}sido{SPC}propost{NSPC}{SPC}na{SPC}literatura
{TER}{SPC}trazido{SPC}{B-CPTO}
{VIR}{SPC}sendo{SPC}conduzid{NSPC}{SPC}pela{SPC}comunidade
{VIR}{SPC}trazendo{SPC}{B-CPTO}
a{SPC}busca{SPC}(pel{ART}|por)
a{SPC}crença{SPC}(de{SPC})?que
a{SPC}demanda{SPC}(pel{ART}|por)
a{SPC}maioria{SPC}daS{SPC}{B-NOMES}{SPC}{TER}{SPC}sido
a{SPC}maioria{SPC}{DEST}{SPC}{B-NOMES}{SPC}{TER}{SPC}sido
a{SPC}partir{SPC}da{SPC}análise{SPC}de{SPC}(diversos{SPC})?trabalhos{SPC}publicados
a{SPC}partir{SPC}das?{SPC}décadas?
a{SPC}partir{SPC}dos?{SPC}anos?
{ART}{SPC}possibilidade{NSPC}{SPC}de{SPC}(utilizaç{NSPC}|aplicaç{NSPC}){SPC}{SER}{SPC}imensas?
além{SPC}{DEST}{SPC}soluç{NSPC}{SPC}particular(es)?
alguns{SPC}estudos{SPC}(também{SPC})?têm{SPC}demonstrado
{B-ADV}{SPC}discutid{NSPC}{SPC}na{SPC}literatura
apesar{SPC}d(o|a){SPC}crescente
após{SPC}{NSPC}{SPC}anos{SPC}de{SPC}experiência
apresent{NSPC}{SPC}como{SPC}ótim{ART}
até{SPC}então{SPC}não{SPC}usua(l|is)({SPC}na{SPC}área)?
atingi{NSPC}{SPC}(o|um){SPC}estado{SPC}da{SPC}arte
{B-ADV}{SPC}{B-VERBOSCPX}
{B-ADV}{SPC}conhecid{NSPC}({SPC}e{SPC}{B-ADV}{SPC}{B-VERBOSCPX})?
colabor{NSPC}{SPC}para{SPC}a{SPC}massificação
começ{NSPC}{SPC}a{SPC}investir
constitu{NSPC}{SPC}uma{SPC}iniciativa
dentre{SPC}{ART}{SPC}vári(as|os)
desempenha{NSPC}{SPC}um{SPC}importante{SPC}papel
desempenha{NSPC}{SPC}um{SPC}papel{SPC}{B-ADJ}
diversos{SPC}trabalhos({SPC}publicados)?
durante{SPC}(a|as){SPC}(últimas?{SPC})?décadas?
é{SPC}importante{SPC}(criar-se|se{SPC}criar)
é{SPC}um{SPC}(dos{SPC})?fator(es)?{SPC}motivador(es)?
é{SPC}uma{SPC}{B-NOMES}{SPC}B-ADJ
é{SPC}uma{SPC}(das{SPC})?{B-NOMES}{SPC}que{SPC}tem{SPC}sido{SPC}{B-VERBOSCPX}
é{SPC}uma{SPC}(das{SPC})?{B-NOMES}{SPC}que{SPC}tem{SPC}sido{SPC}{B-VERBOSCPX}{SPC}{B-ADV}
é{SPC}uma{SPC}(das{SPC})?{B-NOMES}{SPC}que{SPC}tem{SPC}sido{SPC}{B-ADV}{SPC}{B-VERBOSCPX}
é{SPC}uma{SPC}(das{SPC})?{B-NOMES}{SPC}que{SPC}vem{SPC}sendo{SPC}{B-VERBOSCPX}
é{SPC}uma{SPC}(das{SPC})?{B-NOMES}{SPC}que{SPC}vem{SPC}sendo{SPC}{B-VERBOSCPX}{SPC}{B-ADV}
é{SPC}uma{SPC}(das{SPC})?{B-NOMES}{SPC}que{SPC}vem{SPC}sendo{SPC}{B-ADV}{SPC}{B-VERBOSCPX}
é{SPC}uma{SPC}(das{SPC})?áreas?{SPC}de{SPC}estudos?
estudos{SPC}(empíricos{SPC})?{TER}{SPC}(demonstrado|mostrado)
estudos{SPC}(empíricos{SPC})?também{SPC}{TER}{SPC}(demonstrado|mostrado)
existem?{SPC}(hoje{SPC})?na{SPC}literatura
existem{SPC}estudos{SPC}que{SPC}englobam
fornece{NSPC}{SPC}o{SPC}arcabouço{SPC}teórico{SPC}apropriado
fornece{NSPC}{SPC}o{SPC}contexto{SPC}(em{SPC})?que{SPC}está{SPC}inserido{SPC}{EST}{SPC}{NOMES}
muitas?{SPC}pesquisas?{SPC}{TER}{SPC}sido{SPC}feitas?
nas?{SPC}(últimas?{SPC})?década?
as?{SPC}(últimas?{SPC})?décadas?
o{SPC}mundo{SPC}moderno{SPC}exige
passou{SPC}a{SPC}ter{SPC}importância{SPC}central
percebe{NSPC}{SPC}({EST}|a){SPC}necessidade
pesquisadores{SPC}têm{SPC}apresentado

pode{SPC}ser{SPC}vist(a|o){SPC}hoje{SPC}como
pode{NSPC}{SPC}ser{SPC}considerad{NSPC}{SPC}parte{SPC}de{SPC}uma{SPC}evolução
represent{NSPC}{SPC}(a{SPC}mais|uma){SPC}nova{SPC}tentativa
tendo{SPC}em{SPC}vista{SPC}a{SPC}grande{SPC}utilização
um{SPC}domínio{SPC}de{SPC}trabalho{SPC}({NSPC}{SPC})?desafiador
um{SPC}dos{SPC}domínios{SPC}de{SPC}trabalho{SPC}interdisciplinar{SPC}mais{SPC}desafiadores
um{SPC}dos{SPC}temas{SPC}centrais{SPC}de{SPC}pesquisas?
um{SPC}tema{SPC}central{SPC}de{SPC}pesquisas?
uma?{SPC}{NSPC}{SPC}B-ADJ
{SER}{SPC}{B-NOMES}{SPC}{B-ADJ}
é{SPC}uma{SPC}das{SPC}(principais{SPC})?preocupações{SPC}{B-CPTO2}
alternativas?{SPC}interessantes?
torn{NSPC}{SPC}cada{SPC}vez{SPC}mais
{VIR}{SPC}se{SPC}tornando
trouxe{SPC}{B-CPTO}
a{SPC}discussão{SPC}sobre
(em{SPC})?constante{SPC}evolução
(em{SPC})?evolução{SPC}constante
{SER}{SPC}uma{SPC}tentativa{SPC}de
divers(as|os){SPC}(técnicas|métodos){SPC}têm{SPC}sido
uma?{SPC}{B-NOMES}{SPC}{B-ADV}{SPC}{B-VERBOSCPX}
uma{SPC}área{SPC}de{SPC}pesquisa{SPC}({B-ADV}{SPC})?{B-ADJ}
com{SPC}o{SPC}avanço{SPC}d(a|o){SPC}(tecnologia|conhecimento)
entre{SPC}os{SPC}interesses{SPC}de{SPC}pesquisa{SPC}do{SPC}grupo
entre{SPC}as{SPC}pesquisas{SPC}do{SPC}grupo
é{SPC}de{SPC}grande{SPC}interesse
a{SPC}crescente{SPC}(utilização|disponibilidade)
um{SPC}campo{SPC}de{SPC}pesquisa{SPC}recente
{TER}{SPC}experimentado{SPC}um{SPC}aumento
{TER}{SPC}sido{SPC}alvo{SPC}de{SPC}(grandes{SPC})?(pesquisas|estudos)
é{SPC}um{SPC}tópico{SPC}(novo|recente){SPC}de{SPC}pesquisa
é{SPC}um{SPC}tópico{SPC}de{SPC}pesquisa{SPC}(recente|novo)
a{SPC}tendência{SPC}é
entre{SPC}{ART}{SPC}principais
está{SPC}popularizando-se
um{SPC}estudo{SPC}relevante
uma{SPC}alternativa{SPC}interessante
uma{SPC}forma{SPC}fundamentada
um{SPC}procedimento{SPC}comum
permite{SPC}uma?{SPC}melhor
outro{SPC}ponto{SPC}importante
no{SPC}contexto{SPC}atual
nos{SPC}últimos{SPC}anos
no{SPC}brasil
todo{SPC}o{SPC}mundo
nos{SPC}dias{SPC}atuais
o{SPC}grande{SPC}desafio
no{SPC}mundo
comunidade{SPC}científica
estado{SPC}da{SPC}arte
por{SPC}exemplo
popularizou-se
ultimamente
atualmente

Expressões de Lacuna

{SER}{SPC}ainda{SPC}{NSPC}{SPC}{G-ADJ}

{SER}{SPC}comprometid{NSPC}{SPC}devido{SPC}a(o|os|s)?{SPC}{G-NOMES2}
{SER}{SPC}considerad{NSPC}{SPC}uma?{SPC}{G-NOMES2}
{SER}{SPC}pouco{SPC}conhecid{NSPC}
{TER}{SPC}apresentado{SPC}resultados{SPC}abaixo
{TER}{SPC}motivado({SPC}diversas|{SPC}a{SPC}proposição{SPC}de{SPC}diversas)?
antes{SPC}de{SPC}definir(-se)?{SPC}(qual{SPC})?{ART}{SPC}melhor
antes{SPC}de{SPC}se{SPC}definir{SPC}(qual{SPC})?{ART}{SPC}melhor
apresent{NSPC}{SPC}{G-ADV}{SPC}{G-NOMES2}
até{SPC}então{SPC}não{SPC}(obteve|obtiveram)
busca{NSPC}{SPC}responder{SPC}ao{SPC}seguinte{SPC}problema{SPC}de{SPC}investigação
cuja{SPC}solução{SPC}é{SPC}{G-ADJ}
cujas?{SPC}distorç{NSPC}{SPC}{SER}{SPC}evidentes?
dada{SPC}a{SPC}crescente{SPC}{G-NOMES2}
deve{SPC}ser{SPC}realizad{NSPC}{SPC}com{SPC}o{SPC}apoio
devido{SPC}{ART}{SPC}{G-NOMES2}
{SER}{SPC}{G-ADJ}{SPC}(analisar|desenvolver|definir|o{SPC}desenvolvimento|que{SPC}se{SPC}defina|que)
é{SPC}uma?{SPC}{G-NOMES2}({SPC}para{SPC}pesquisadores{SPC}da{SPC}área)?
é{SPC}uma{SPC}{G-NOMES}{SPC}{G-ADJ}
existem{SPC}poucos{SPC}trabalhos{SPC}relacionados
faz{NSPC}{SPC}{G-ADJ}{SPC}estudos
{ART}{SPC}maior(es)?{SPC}{G-NOMES2}{SPC}encontrad{NSPC}
mostra{SPC}uma?{SPC}grande{SPC}{G-NOMES2}
não{SPC}é{SPC}uma?{SPC}{G-NOMES}{SPC}{G-ADJ2}
não{SPC}é{SPC}{G-NOMES}{SPC}{G-ADJ2}
não{SPC}{SER}{SPC}ainda{SPC}{G-NOMES}{SPC}{B-VERBOSCPX}
não{SPC}{SER}{SPC}ainda{SPC}{B-VERBOSCPX}
não{SPC}pode{NSPC}{SPC}ser{SPC}considerad{NSPC}
{ART}{SPC}{G-NOMES2}{SPC}em{SPC}questão
observam-se{SPC}vári{NSPC}{SPC}{G-NOMES2}
para{SPC}amenizar{SPC}{ART}{SPC}{G-NOMES2}
pode{SPC}refletir{SPC}negativamente
pode{SPC}representar{SPC}grandes{SPC}{G-NOMES2}
pouco{SPC}{B-VERBOSCPX}{SPC}(na|pela){SPC}literatura
que{SPC}dificilmente({SPC}se{SPC}ajusta{SPC}{ART}{SPC}necessidades)?
se{SPC}faz{SPC}{G-ADJ}
se{SPC}sentem?{SPC}incapaz(es)?
sente{NSPC}{SPC}incapaz(es)?
tal{SPC}{G-NOMES2}{SPC}não{SPC}é{SPC}{G-ADJ2}
tal{SPC}{G-NOMES2}{SPC}não{SPC}é{SPC}uma?{SPC}{G-NOMES}{SPC}{G-ADJ2}
torn{NSPC}{SPC}uma{SPC}necessidade{SPC}básica
uma?{SPC}d{ART}{SPC}(grandes{SPC})?{G-NOMES2}
uma?{SPC}d{ART}{SPC}{G-NOMES2}{SPC}{G-ADV}{SPC}{G-ADJ}
uma?{SPC}grande{SPC}{G-NOMES2}
uma?{SPC}{G-NOMES2}{SPC}{G-ADV}{SPC}{G-ADJ}
torna{NSPC}{SPC}{G-ADV}{SPC}{G-ADJ}
{SER}{SPC}({G-ADV}{SPC})?{G-ADJ}
enfrenta{NSPC}{SPC}{G-NOMES2}
{ART}{SPC}principa(l|is){SPC}{G-NOMES2}
{SER}{SPC}(um{SPC})?aspectos?{SPC}{G-ADJ}
uma?{SPC}d{ART}{SPC}{G-NOMES2}{SPC}centra(l|ais)
uma?{SPC}{G-NOMES2}{SPC}importante
necessita{NSPC}{SPC}de{SPC}{M-NOMES}{SPC}complex(as|os)
nem{SPC}sempre{SPC}{G-ADJ2}
resulta{NSPC}{SPC}em{SPC}grandes{SPC}{G-NOMES2}
uma?{SPC}{G-NOMES2}{SPC}{G-ADJ}
{SER}{SPC}obrigad{NSPC}{SPC}a
se{SPC}defrontam?{SPC}com
busca{NSPC}{SPC}evitar
busca{NSPC}{SPC}responder

{SER}{SPC}agravad{NSPC}
principalmente{SPC}pela{SPC}falta
problema{SPC}de{SPC}investigação
torna{NSPC}{SPC}{G-ADJ}
que{SPC}dificulta
acarreta{SPC}o{SPC}aumento
não{SPC}é{SPC}{G-ADJ2}
informações{SPC}{G-ADJ}
há{SPC}(a{SPC})?necessidade
não{SPC}está{SPC}de{SPC}acordo
situações{SPC}frustrantes
{SER}{SPC}{G-ADJ}
embora
entretanto
contudo
todavia
dificilmente
dificulta
impõem
no{SPC}entanto
porém

Expressões de Propósito

{EST}{SPC}{NOMES}{SPC}{P-VERBOS}
{EST}{SPC}{NOMES}{SPC}{P-VERBOSIRREG}
{NEST}{SPC}{NOMES}{SPC}{P-VERBOS}
{NEST}{SPC}{NOMES}{SPC}{P-VERBOSIRREGCPX}
{NEST}{SPC}{NOMES}{SPC}{SER}{SPC}{P-VERBOSCPX}
{NOMES}{SPC}aqui{SPC}reportad{ART}{SPC}{P-VERBOSIRREG}
{SER}{SPC}{P-VERBOSCPX}{SPC}{NEST}{SPC}{NOMES}
{ART}{SPC}{NOMES}{SPC}{P-VERBOS}
{ART}{SPC}{NOMES}{SPC}{P-VERBOSIRREG}
{ART}{SPC}presente{SPC}{NOMES}{SPC}{P-VERBOS}
{ART}{SPC}presente{SPC}{NOMES}{SPC}{P-VERBOSIRREG}
{ART}{SPC}seguinte{SPC}{NOMES}{SPC}{P-VERBOS}
{ART}{SPC}seguinte{SPC}{NOMES}{SPC}{P-VERBOSIRREG}
metas?{SPC}{DEST}{SPC}{NOMES}{SPC}{SER}{SPC}({P-VERBOS}|fazer)
n{ART}{SPC}presente{SPC}{NOMES}{SPC}{P-VERBOS}
n{ART}{SPC}presente{SPC}{NOMES}{SPC}{P-VERBOSIRREGCPX}
n{ART}{SPC}presente{SPC}{NOMES}{SPC}{SER}{SPC}{P-VERBOSCPX}
noss{ART}{SPC}{NOMES}{SPC}{P-VERBOS}
noss{ART}{SPC}{NOMES}{SPC}{P-VERBOSIRREG}
{ART}{SPC}{NOMES}{SPC}apresentad{ART}{SPC}{NEST}{SPC}{NOMES}{SPC}{P-VERBOSIRREG}
o{SPC}foco{SPC}{DEST}{SPC}{NOMES}{SPC}({SER}|{ESTAR})
o{SPC}foco{SPC}central{SPC}de{SPC}análise{SPC}({DEST}{SPC}{NOMES}{SPC})?{SER}
o{SPC}objetivo{SPC}central{SPC}d{ART}{SPC}{NOMES}{SPC}consiste
objetivos?{SPC}{DEST}{SPC}{NOMES}{SPC}{SER}{SPC}({P-VERBOS}|fazer)
objetivos?{SPC}principa(l|is){SPC}{DEST}{SPC}{NOMES}{SPC}{SER}{SPC}({P-VERBOS}|fazer)
principa(l|is){SPC}objetivos?{SPC}{DEST}{SPC}{NOMES}{SPC}{SER}{SPC}({P-VERBOS}|fazer)
objetivos?{SPC}{DEST}{SPC}{NOMES}
objetivos?{SPC}principa(l|is){SPC}{DEST}{SPC}{NOMES}
principa(l|is){SPC}objetivos?{SPC}{DEST}{SPC}{NOMES}
objetivos?{SPC}(d(a|o){SPC}presente|d(a|o){SPC}seguinte){SPC}{NOMES}
objetivos?{SPC}principa(l|is){SPC}(d(a|o){SPC}presente|d(a|o){SPC}seguinte){SPC}{NOMES}
principa(l|is){SPC}objetivos?{SPC}(d(a|o){SPC}presente|d(a|o){SPC}seguinte){SPC}{NOMES}
o{SPC}foco{SPC}principal{SPC}{DEST}{SPC}{NOMES} metas?{SPC}{DEST}{SPC}{NOMES}
{NOMES}{SPC}{SER}{SPC}uma{SPC}tentativa{SPC}de

{SER}{SPC}{P-VERBOSCPX}{SPC}(e{SPC}P-VERBOSCPX{SPC})?{NEST}{SPC}{NOMES}
{NEST}{SPC}{NOMES}{SPC}nos{SPC}preocupamos
{NEST}{SPC}{NOMES}{SPC}a{SPC}preocupação{SPC}{SER}
concentramos{SPC}noss(o|a){SPC}{G-NOMES}
nossos{SPC}objetivos{SPC}(são|foram)
{NEST}{SPC}{NOMES}{SPC}(,{SPC})?{SER}{SPC}{P-VERBOSCPX}
{NEST}{SPC}{NOMES}{SPC}(,{SPC})?o{SPC}enfoque{SPC}({SER}|{ESTAR})
{EST}{SPC}{NOMES}{SPC}{TER}{SPC}como{SPC}foco{SPC}(principal|central)
{EST}{SPC}{NOMES}{SPC}se{SPC}prop{NSPC}
uma{SPC}das{SPC}vertentes{SPC}d{ART}{SPC}{NOMES}{SPC}{P-VERBOS}
de{SPC}forma{SPC}geral{SPC},{SPC}{P-VERBOS}
de{SPC}forma{SPC}geral{SPC},{SPC}{SER}{SPC}{P-VERBOSCPX}
{SER}{SPC}{P-VERBOSCPX}{SPC}(ainda|aqui)
busc{NSPC}{SPC}também{SPC}{P-VERBOS}
além{SPC}disso({SPC},)?{SPC}{P-VERBOS}
discute{SPC},{SPC}ainda{SPC},
define{SPC},{SPC}também{SPC},
especificamente{SPC},{SPC}{P-VERBOS}
dessa{SPC}forma{SPC},{SPC}pretendese
apresent{NSPC}{SPC}uma{SPC}proposta
também{SPC}{SER}{SPC}{P-VERBOSCPX}
assim{SPC},{SPC}{P-VERBOS}
pretende{NSPC}{SPC}(dar|fornecer)
uma{SPC}proposta

Expressões de Metodologia

{EST}{SPC}{NOMES}{SPC}{M-VERBOS}
{EST}{SPC}{NOMES}{SPC}(se{SPC})?{M-VERBOSCPX3}
{EST}{SPC}{NOMES}{SPC}{SER}{SPC}{M-VERBOSCPX3}
{EST}{SPC}{NOMES}{SPC}{TER}{SPC}sido{SPC}{M-VERBOSCPX3}
{ART}{SPC}{NOMES}{SPC}{TER}{SPC}sido{SPC}{M-VERBOSCPX3}
{ESTAR}{SPC}{M-VERBOSCPX3}
{EST}{SPC}{M-NOMES}{SPC}{M-VERBOS}
{EST}{SPC}{M-NOMES}{SPC}{SER}{SPC}{M-VERBOS}
{M-INI}{SPC},{SPC}{M-VERBOS}
{M-INI}{SPC}{M-VERBOS}
{M-INI}{SPC},{SPC}{SER}{SPC}{M-VERBOSCPX2}
{M-INI}{SPC}{SER}{SPC}{M-VERBOSCPX2}
{M-VERBOSCPX}{SPC}{M-CPTO}
{M-VERBOS}{SPC}{M-CPTO2}
{M-VERBOS}{SPC}({ART}{SPC})?{M-NOMES}
{NOMES}{SPC}aqui{SPC}{R-VERBOSCPX}{SPC}(se{SPC})?{M-VERBOSCPX3}
{SER}{SPC}{M-VERBOSCPX}{SPC}{M-CPTO}
{SER}{SPC}{M-VERBOSCPX2}{SPC}{M-CPTO2}
{SER}{SPC}{M-VERBOSCPX2}{SPC}({ART}{SPC})?{M-NOMES}
{SER}{SPC}{M-VERBOSCPX3}
{TER}{SPC}como{SPC}base{SPC}(teórica|metodologia)
a{SPC}metodologia{SPC}básica{SPC}{M-VERBOSCPX2}{SPC}para
o{SPC}referencial{SPC}teórico{SPC}{DEST}{SPC}{NOMES}{SPC}(se{SPC})?{M-VERBOSCPX3}
para{SPC}ilustrar{SPC}(a{SPC}relevância|e{SPC}validar)?
{M-VERBOSCPX2}{SPC}{NEST}{SPC}{NOMES}
para{SPC}realizar(mos|-se){SPC}EST{SPC}{G-NOMES}
{M-VERBOSCPX2}{SPC}como{SPC}base{SPC}(teórica|metodologia)
{SER}{SPC}{R-VERBOSCPX3}{SPC}com{SPC}(o{SPC})?auxílio
{SER}{SPC}{R-VERBOSCPX3}{SPC}{M-CPTO}
{ART}{SPC}{M-NOMES}{SPC}({NSPC}{SPC})?{M-VERBOSCPX2}
{ART}{SPC}{M-NOMES}{SPC}{R-VERBOSCPX3}{SPC}{M-VERBOS}

que{SPC}funciona{SPC}segundo
trabalh{NSPC}{SPC}inicialmente{SPC}com
estudos?{SPC}de{SPC}caso
serv(e|iu){SPC}de{SPC}suporte
definid{NSPC}{SPC}com{SPC}base

Expressões de Resultado

{EST}{SPC}análise{SPC}resultou
{R-NOMES2}{SPC}nos{SPC}permitiu{SPC}{R-VERBOSCPX2}
{R-NOMES2}{SPC}permitiu{SPC}{R-VERBOSCPX2}
{R-VERBOSCPX}{SPC}resultados{SPC}bastante{SPC}significativos
{R-VERBOSCPX}{SPC}resultados{SPC}significativos
{R-VERBOSCPX}{SPC}(as{SPC})?vantagens
{R-VERBOSCPX}{SPC}reflexões{SPC}sobre
através{SPC}dos?{SPC}resultados?{SPC}obtidos?
como{SPC}resultados?{SPC}{R-VERBOSCPX}
destaca-se{SPC}como{SPC}resultados?
é{SPC}{R-VERBOSCPX}{SPC}uma?{SPC}nov(a|o)
os?{SPC}resultados?{SPC}{ESTAR}{SPC}entre
os?{SPC}resultados?{SPC}{R-NOMES}
os?{SPC}resultados?{SPC}{R-NOMES}{SPC}{R-VERBOS}
os?{SPC}resultados?{SPC}{R-NOMES}{SPC}nos{SPC}{R-VERBOS}
os?{SPC}resultados?{SPC}{R-VERBOS}
os?{SPC}resultados?{SPC}{SER}{SPC}{R-VERBOSCPX}
os?{SPC}resultados?{SPC}nos{SPC}{R-VERBOS}
pode{NSPC}{SPC}melhorar{SPC}significativamente
por{SPC}meio{SPC}dos?{SPC}resultados?{SPC}obtidos?
result{NSPC}{SPC}(em{SPC})?um{SPC}estudo
{SER}{SPC}{R-VERBOSCPX}{SPC}algumas
{SER}{SPC}{R-VERBOSCPX}{SPC}algumas{SPC}considerações
{SER}{SPC}{R-VERBOSCPX}{SPC}considerações
{SER}{SPC}{R-VERBOSCPX}{SPC}alguns
{EST}{SPC}{NOMES}{SPC}também{SPC}{P-VERBOS}
{SER}{SPC}(também{SPC})?{R-VERBOSCPX3}
também{SPC}{SER}{SPC}{R-VERBOSCPX3}
possui{SPC}como{SPC}(principal{SPC})?característica
como{SPC}resultado{SPC}pôde-se
os?{SPC}resultados?{SPC}(de{SPC}desempenho{SPC})?{SER}{SPC}{R-VERBOSCPX}
estudos?{SPC}((teóricos?|empíricos?){SPC})?{SER}{SPC}{R-VERBOSCPX3}
os?{SPC}resultados?{SPC}preliminar(es)?
^ (com{SPC}base{SPC}NEST{SPC}dados{SPC},)
^ (como{SPC}resultado{SPC},)
sendo{SPC}obtidos{SPC}resultados{SPC}relevantes
os{SPC}benefícios{SPC}medidos
(foi|ficou){SPC}evidenciado{SPC}que
(foi|ficou){SPC}evidente{SPC}que
pudemos{SPC}verificar

Expressões de Conclusão

{C-INI}{SPC}{DEST}{SPC}{NOMES}
{C-INI2}{SPC},{SPC}{EST}{SPC}{NOMES}
{C-INI2}{SPC}{EST}{SPC}{NOMES}
{EST}{SPC}{NOMES}{SPC}{C-VERBOS}
{SER}{SPC}contribuiç{NSPC}{SPC}{DEST}{SPC}{NOMES}
{SER}{SPC}(uma{SPC})?contribuiç{NSPC}{SPC}importantes?{SPC}{DEST}{SPC}{NOMES}

{SER}{SPC}(uma{SPC})?importantes?{SPC}contribuiç{NSPC}{SPC}{DEST}{SPC}{NOMES}
{ART}{SPC}{NOMES}{SPC}aqui{SPC}{R-VERBOSCPX}{SPC}colaborou
{ART}{SPC}{NOMES}{SPC}procurou{SPC},{SPC}principalmente{SPC},{SPC}contribuir
{ART}{SPC}{NOMES}{SPC}procurou{SPC}contribuir
apresent{NSPC}{SPC}contribuiç{NSPC}{SPC}distint{NSPC}
apresent{NSPC}{SPC}outr{NSPC}{SPC}contribuiç{NSPC}
apresent{NSPC}{SPC}várias{SPC}contribuições
contribuiç{NSPC}{SPC}{DEST}{SPC}{NOMES}
contribuiç{NSPC}{SPC}{SER}{SPC}consideradas?{SPC}principa{NSPC}
contribuiç{NSPC}{SPC}consideradas?{SPC}principa{NSPC}
os{SPC}resultados{SPC}obtidos{SPC}são{SPC}promissores
outra{SPC}conclusão{SPC}é
permitiu{SPC}concluir{SPC}que
pode{SPC}vir{SPC}a{SPC}contribuir
pode{NSPC}{SPC}concluir{SPC}que
pôdese{SPC}concluir{SPC}que
pode{NSPC}{SPC}subdividir{SPC}os{SPC}resultados
tais{SPC}contribuições{SPC}consistem
nos{SPC}permit{NSPC}concluir
foi{SPC}possível{SPC}constatar
tendo{SPC}sido{SPC}obtidos{SPC}resultados{SPC}que{SPC}indica(m|ram)
a{SPC}apresentação{SPC}clara{SPC}e{SPC}concisa
{C-INI}
esperase{SPC}{C-VERBOS2}
contribuiu?{SPC}para
resultados{SPC}promissores
resultados{SPC}melhores

Apêndice

E

Questionários Utilizados na Avaliação do Ambiente *SciPo*

E.1 Experimento 1

1. Você tinha conhecimento prévio sobre a estruturação de um resumo acadêmico (em termos de componentes esperados)?
   ( ) não ( ) sim, mas mediano ( ) sim

2. Você teve algum problema para utilizar o ambiente?
   ( ) não ( ) sim Por quê?

3. Você corrigiu a classificação dada pelo ambiente para as sentenças do resumo? Em quantas sentenças você corrigiu a classificação?
   ( ) não ( ) sim

4. Quão difícil foi corrigir a classificação dada pelo ambiente?
   ( ) não corrigi a classificação
   ( ) não tive problemas para corrigir a classificação
   ( ) tive algumas dúvidas para corrigir a classificação
   ( ) tive muitas dúvidas para corrigir a classificação

5. Você consideraria a classificação dada pelo ambiente:

( ) muito confiável ( ) confiável ( ) pouco confiável

6. Na sua opinião, qual foi a relevância das críticas e sugestões emitidas pelo ambiente?

Críticas:	Sugestões:
( ) muito relevante	( ) muito relevante
( ) relevante	( ) relevante
( ) pouco relevante	( ) pouco relevante

7. Você acatou as sugestões emitidas pelo ambiente?
   ( ) integralmente ( ) parcialmente ( ) não acatei

Durante a construção (reconstrução) da estrutura:

8. Você utilizou a recuperação de exemplos de estruturas similares?
   ( ) não ( ) sim

9. Você utilizou a visualização de exemplos de componentes e/ou estratégias?
   ( ) não ( ) sim

10. Se sim, eles foram úteis/informativos?
    ( ) não ( ) sim

Durante a fase de redação (escrita):

11. Você utilizou a recuperação de exemplos de estruturas similares?
    ( ) não ( ) sim

12. Você utilizou a visualização de exemplos de componentes/estratégias?
    ( ) não ( ) sim

13. Se sim, eles foram úteis/informativos?
    ( ) não ( ) sim

14. Você utilizou texto reutilizável?
    ( ) não ( ) sim

15. Você utilizou a lista de marcadores discursivos?
    ( ) não ( ) sim

16. Você utilizou o revisor ortográfico e gramatical?

( ) não ( ) sim

17. Como você classifica o tipo de auxílio oferecido pelo ambiente SciPo?
( ) muito útil ( ) útil ( ) pouco útil ( ) atrapalha

18. Considerando as funcionalidades e recursos disponíveis no SciPo, o que você considera mais importante (maior utilidade)?

19. E o que você considera pouco importante (pouca utilidade)?

20. O que você adicionaria ao SciPo (funcionalidades e/ou recursos)?

21. Como você classificaria a interface do SciPo quanto a facilidade do uso?
( ) difícil de usar ( ) razoável ( ) fácil de usar

22. Você utilizaria o SciPo novamente (em situações reais)?

E.2 Experimento 2

1. Você tinha conhecimento prévio sobre a estruturação de um resumo acadêmico (em termos de componentes esperados)?
( ) não ( ) sim, mas mediano ( ) sim

2. Você teve algum problema para utilizar o ambiente?
( ) não ( ) sim Por quê?

3. Na sua opinião, qual foi a relevância das críticas e sugestões emitidas pelo ambiente?

Críticas:	Sugestões:
( ) muito relevante	( ) muito relevante
( ) relevante	( ) relevante
( ) pouco relevante	( ) pouco relevante

4. Você acatou as sugestões emitidas pelo ambiente?
( ) integralmente ( ) parcialmente ( ) não acatei

Durante a construção da estrutura:

5. Você utilizou a recuperação de exemplos de estruturas similares?
   ( ) não ( ) sim

6. Você utilizou a visualização de exemplos de componentes e/ou estratégias?
   ( ) não ( ) sim

7. Se sim, eles foram úteis/informativos?
   ( ) não ( ) sim

Durante a fase de redação (escrita):

8. Você utilizou a recuperação de exemplos de estruturas similares?
   ( ) não ( ) sim

9. Você utilizou a visualização de exemplos de componentes/estratégias?
   ( ) não ( ) sim

10. Se sim, eles foram úteis/informativos?
    ( ) não ( ) sim

11. Você utilizou texto reutilizável?
    ( ) não ( ) sim

12. Você utilizou a lista de marcadores discursivos?
    ( ) não ( ) sim

13. Você utilizou o revisor ortográfico e gramatical?
    ( ) não ( ) sim

14. Como você classifica o tipo de auxílio oferecido pelo ambiente SciPo?
    ( ) muito útil ( ) útil ( ) pouco útil ( ) atrapalha

15. Considerando as funcionalidades e recursos disponíveis no SciPo, o que você considera mais importante (maior utilidade)?

16. E o que você considera pouco importante (pouca utilidade)?

17. O que você adicionaria ao SciPo (funcionalidades e/ou recursos)?

18. Como você classificaria a interface do SciPo quanto a facilidade do uso?
    ( ) difícil de usar ( ) razoável ( ) fácil de usar

19. Você utilizaria o SciPo novamente (em situações reais)?

Apêndice

F

Orientações para a Anotação Manual do Corpus de Resumos Acadêmicos

Estas orientações descrevem o esquema de classificação para resumos científicos do projeto SciPo. O esquema de classificação é mostrado na Figura 1. Cada uma das categorias está associada com uma cor e uma etiqueta, correspondente a(s) primeira(s) letra(s) do nome da categoria. Essas categorias serão utilizadas para a anotação manual dos textos.

Etiqueta	Categoria	Descrição
C	Contexto	Conhecimento aceito pela comunidade científica
L	Lacuna	Problema de pesquisa, necessidade, ...
P	Propósito	Propósito da pesquisa
M	Metodologia	Metodologia utilizada
R	Resultado	Resultados obtidos
Co	Conclusão	Conclusão, recomendação, contribuição, ...
E	Estrutura	Descrição das partes do texto / assuntos tratados

Figura F.1: Esquema de classificação com as etiquetas que devem ser utilizadas na anotação

F.1 Antes da Anotação

Leia o texto antes da anotação. Isto é importante, uma vez que em alguns textos a interpretação do papel de determinadas sentenças na argumentação como um todo se torna possível somente após uma visão geral do texto. Não tente entender a pesquisa relatada. Concentre-se em entender a estrutura da argumentação feita pelo autor.

F.2 Processo de Anotação

A anotação deve ser feita sentença a sentença e é mutuamente exclusiva, isto é, cada sentença pode ser classificada em apenas uma única categoria. Entretanto, é possível que algumas sentenças apresentem características de mais de uma categoria (sentenças em que os papéis argumentativos se sobrepõem, como por exemplo, sentenças indicando o propósito e também a metodologia). Nesses casos, você deve decidir por apenas uma categoria. Prefira sempre a categoria correspondente ao papel argumentativo que está mais fortemente caracterizado na sentença.

Para determinar o papel argumentativo da sentença, você deve tentar interpretá-la de acordo com o papel que você acha que o autor pretendia que ela tivesse na argumentação. Focalize sempre as intenções do autor e considere sempre o contexto e a localização da sentença.

Quando uma sentença apresentar mais de um papel argumentativo, em caso de dúvida, siga as seguintes instruções:

1. Dê preferência ao papel argumentativo "mais forte". A maior ocorrência de papéis argumentativos sobrepostos está nas sentenças que indicam propósito. Geralmente, a categoria **Propósito (P)** é mais forte do que as outras categorias. No caso de sobreposição de outras categorias, você deve decidir qual delas está mais caracterizada na sentença.
2. Dê preferência a categoria caracterizada pela maior string contida na sentença. Vamos pegar como exemplo um caso que apresente resultados e metodologia na mesma sentença. Digamos que 60% da sentença referem-se aos resultados e 40% indica metodologia. Então, nesse caso, classifique a sentença como **Resultado (R)**.

Você pode anotar mais de uma sentença com a mesma categoria. Essas sentenças podem ser consecutivas ou não. É comum que sentenças consecutivas sejam anotadas com a mesma categoria, desde que juntas elas preencham os critérios da categoria. Por exemplo, você pode marcar duas sentenças como propósito, se juntas elas descrevem o objetivo do texto. **Se você não conseguir atribuir nenhuma categoria a alguma**

sentença, anote o identificador da sentença e descreva as suas dificuldades. Não se esqueça de anotar todas as sentenças do texto.

F.3 Descrição das Categorias

F.3.1 Contexto (C)

Sentenças de contexto são aquelas que apresentam conhecimento já reconhecido numa determinada área de pesquisa. Essas sentenças servem para estabelecer o contexto da pesquisa apresentada. O *Contexto* pode incluir afirmações sobre a importância do campo, sobre sua evolução ao longo do tempo e familiarizações de termos e conceitos referentes ao campo de pesquisa. O mais comum é que as sentenças de contexto apareçam no início do texto.

Pode acontecer de sentenças de contexto aparecer no meio do texto. Nesses casos, pode ficar difícil decidir se a sentença é *Contexto*. Então, use o seguinte teste: se você acha que a sentença poderia aparecer no início do texto e que ela **não contém material próprio** da pesquisa apresentada pelo autor, anote como *Contexto*.

Sentenças de contexto podem conter citações. Em geral, essas citações são a "pioneiros" da área ou a trabalhos amplamente aceitos pela comunidade científica, ou então são colocadas apenas para "dividir" a responsabilidade do autor sobre a afirmação de contexto.

Pode acontecer do texto não conter nenhuma sentença de contexto, principalmente se o autor começa indicando os propósitos do seu trabalho.

F.3.2 Lacuna (L)

Sentenças que indicam uma área de pesquisa importante que não foi investigada por outros autores ou que não tenha sido suficientemente desenvolvida devem ser marcadas como *Lacuna*. Normalmente, o autor indica uma lacuna em apenas uma ou duas sentenças e as escreve usando o presente como tempo verbal. Entretanto, isso não é uma regra. Outro indicativo de que uma determinada sentença é uma sentença de lacuna é o uso de marcadores de contraste, como "no entanto", "porém", "contudo", etc.

Marque como *Lacuna* sentenças que indicam:

- que existe um problema em determinada área de pesquisa que ainda não está resolvido;
- que a literatura disponível é inadequada ou, simplesmente, que não existe literatura disponível;

- que há um conflito não resolvido entre os autores dos estudos prévios relacionados ao tópico de pesquisa em foco, isto é, existe uma controvérsia. Essa controvérsia pode ser um desentendimento teórico ou prático;
- que o exame da literatura sugere uma extensão do tópico, ou levanta uma nova questão de pesquisa não considerada previamente por outros pesquisadores em seu campo de atuação;
- que as soluções disponíveis até o momento são inadequadas ou apresentam fraquezas, ou seja, sentenças que indicam aspectos negativos de outros trabalhos/abordagens/-métodos.

As sentenças que indicam lacunas como aspectos motivadores do trabalho devem ser anotadas como *Lacuna*. Indicando algum tipo de falha deixada pelos estudos anteriores, a *Lacuna* prepara o leitor para focalizar o estudo em questão no trabalho, e de certa forma justifica a realização do estudo. Essas sentenças geralmente contem sinalizadores léxicos expressando dificuldades, necessidades, problemas, fraquezas, inadequação, etc.

Pode acontecer do texto não conter nenhuma sentença de lacuna. Em geral, isso acontece quando o texto também não apresenta sentenças de contexto.

F.3.3 Propósito (P)

Sentenças de propósito descrevem o objetivo principal do trabalho. A apresentação do propósito está diretamente ligada à questão da pesquisa na qual está baseado o estudo. Normalmente, todo texto contém pelo menos uma sentença de propósito.

Geralmente, o propósito principal do estudo é expresso em uma única sentença. No entanto, outras sentenças podem ser marcadas como *Propósito*, uma vez que o propósito principal pode ser detalhado em outras sentenças e ainda podem existir propósitos secundários, principalmente quando se tratam de teses e dissertações. Tanto a sentença apresentando o propósito principal, como as sentenças detalhando o propósito e/ou introduzindo propósitos secundários, devem ser marcadas como *Propósito*.

As sentenças de propósito podem aparecer tanto no passado como no presente, dependendo da orientação utilizada. Quando a orientação da apresentação do propósito é dirigida ao próprio trabalho, isto é, refere-se ao artigo, tese, dissertação ou relatório que vai comunicar a informação sobre a pesquisa em questão, usa-se o presente: *"O principal objetivo desta dissertação é ..."*. Quando a orientação do propósito é dirigida à atividade de pesquisa, em outras palavras, ao próprio estudo em vez do trabalho escrito, usa-se o passado: *"O principal objetivo deste trabalho de mestrado foi ..."*. Entretanto, nem sempre os autores seguem essa recomendação de uso do tempo verbal, o que muitas vezes torna a sentença ambígua. Fique atento a possíveis ambigüidades e tente detectar a real intenção

do autor.

Uma maneira fácil de identificar sentenças de propósito, principalmente indicando o propósito principal, é pela identificação das expressões utilizadas nessa categoria de sentenças. Em geral, sentenças de propósito começam com expressões explícitas, como: *"O propósito deste trabalho é...", "Neste trabalho, apresentamos...", "O principal objetivo dessa dissertação...", etc.*

Veja exemplos de expressões-padrões de *Propósito* abaixo:

- *"Desse modo,* **o objetivo deste trabalho é** *estudar modelos de defeitos de software e investigar métodos de injeção, baseado nos conceitos e princípios oriundos do critério Análise de Mutantes."*
- *"Assim,* **este trabalho visa** *a aplicar e avaliar empiricamente duas técnicas teste de que têm se mostrado promissoras: a Técnica baseada em Modificação (Wong et al., 1997a) e a Técnica baseada em Mutação Seletiva (Wong et al., 1997b)."*
- **"Este trabalho propõe e investiga** *o uso de técnicas de inteligência artificial para fusão de sensores com o objetivo de melhorar a precisão e acurácia de medidas de distância entre um robô e um objeto no seu ambiente de trabalho, obtidas com diferentes sensores."*
- *"bf Neste trabalho, isto é alcançado através do uso de um sistema interativo de recuperação e gerenciamento de informações projetado para facilitar o acesso a itens (ou parte deles) armazenados no servidor."*

Em alguns casos, pode ser difícil distinguir as sentenças de *propósito* das sentenças de *resultado*. Toda sentença que se refere ao objetivo do estudo/artigo/tese deve ser marcada como *Propósito*. Conforme comentado anteriormente, essas sentenças, em geral, são bem sinalizadas por expressões-padrões. Sentenças que apresentam expressões-padrões de propósito devem ser marcadas como *Propósito*, mesmo que sejam direcionadas ao artefato produzido pelo autor e não ao problema da pesquisa em questão.

Veja os exemplos de Propósito abaixo:

- **"Este trabalho apresenta** *o Animbs (Animation for MBS), um sistema capaz de visualizar dados gerados por um sistema de simulação de engenharia (SD/FAST) na forma de animações por computador."*
- *"Diante desses problemas,* **este trabalho apresenta** *a ferramenta VersionWeb que foi desenvolvida."*
- **"Este trabalho apresenta** *o Método para Projeto de Hiperdocumentos para Ensino, ou EHDM (Educational Hyperdocuments Design Method), que proporciona uma abordagem sistemática para apoiar o projeto e desenvolvimento de aplicações hipermídia para ensino."*

Sentenças que descrevem o artefato (software, método, técnica, etc.) desenvolvido pelo autor devem ser marcadas como *Resultado*, mesmo que estejam relacionadas ao propósito principal. Descrições das partes componentes, da funcionalidade, de resultados de avaliações, entre outras descrições, devem ser marcadas como *Resultado*.

Veja um exemplo de Resultado ligado ao Propósito abaixo:

- *"Este trabalho apresenta o Método para Projeto de Hiperdocumentos para Ensino, ou EHDM (Educational Hyperdocuments Design Method), que proporciona uma abordagem sistemática para apoiar o projeto e desenvolvimento de aplicações hipermídia para ensino. (...)* **As três fases que compõem o método - modelagem conceitual hierárquica, projeto navegacional de contextos e construção e teste - são apresentadas.***"*

Não se esqueça que, em geral, todo resumo acadêmico apresenta pelo menos uma sentença indicando o propósito. Por isso, procure atentamente uma sentença que possa ser classificada com essa categoria. Entretanto, caso você não consiga anotar nenhuma sentença como *Propósito* em um dos textos (o que pode acontecer!), tome nota do identificador do texto para que ele seja posteriormente revisado.

F.3.4 Metodologia (M)

Sentenças descrevendo a metodologia utilizada para a realização da pesquisa devem ser marcadas como *Metodologia*. Sentenças de metodologia geralmente aparecem após o propósito principal ou um resultado, mas isso não é uma regra. Marque como *Metodologia* apenas as sentenças relacionadas à metodologia utilizada pelo autor para a realização da pesquisa relatada.

Marque como *Metodologia*:

- Sentenças que indicam os **materiais e métodos** utilizados ou que **servem de base** para a pesquisa. Nos textos de Computação, o mais comum é que os métodos sejam descritos ou, pelo menos, indicados. Por métodos entendemos métodos/técnicas/abordagens/etc. Sentenças indicando trabalhos nos quais a pesquisa é baseada (e que devem ser marcadas como metodologia) podem conter citações;
- Sentenças **justificando a metodologia** utilizada;
- Sentenças que **indicam critérios e condições** para a realização da pesquisa;
- Sentenças que **descrevem conjuntos de dados** utilizados na pesquisa;
- Sentenças que indicam os procedimentos utilizados para a **avaliação/comprovação** dos resultados, como **estudo de casos, testes empíricos**, ou mesmo **software/-protótipos computacionais**, desde que o motivo para seu desenvolvimento tenha sido a comprovação/demonstração do propósito principal da pesquisa relatada.

Pode acontecer do texto não conter nenhuma sentença de metodologia, pois muitas vezes o autor já dá uma indicação da metodologia utilizada no propósito principal, ou, simplesmente, a metodologia não é mencionada.

F.3.5 Resultado (R)

As sentenças de resultado indicam os principais resultados da pesquisa. É mais comum que as sentenças de resultado apareçam após sentenças de propósito ou metodologia. Também é comum que essas sentenças estejam escritas no passado. Novamente, isso não é uma regra.

Marque como *Resultado* sentenças que:

- Descrevem um artefato. Um artefato pode ser um software (ou parte de um), um método, uma técnica, um conjunto de regras, ou seja, qualquer tipo de produto computacional desenvolvido pelo autor. Conforme comentado na seção sobre a categoria Propósito, a descrição do artefato pode envolver descrição das partes componentes do artefato, da funcionalidade, de resultados de avaliações, entre outras;
- Descrevem ou "indicam" resultados de experimentos;
- Descrevem ou "indicam" resultados de avaliações;
- Comentam/discutem os resultados específicos da pesquisa.

Perceba que sentenças que apenas indicam a existência de resultados também devem ser marcadas como *Resultado*.

Veja exemplos de resultado indicativo abaixo:

- *"As três fases que compõem o método - modelagem conceitual hierárquica, projeto navegacional de contextos e construção e teste - são apresentadas."*
- *"Uma nova semântica de comunicação é implementada ("at-least-once") e os resultados obtidos na comparação com a semântica original ("at-most-once") são analisados."*
- *"Três tipos de sistemas são definidos com níveis crescentes de complexidade e funcionalidade."*
- *"Diversos critérios são propostos para a avaliação do sistema e para a determinação da sua adequação às principais aplicações na agricultura."*

Outro aspecto relativo às sentenças da categoria *Resultado* que deve ser observado é a diferença entre *Resultado* e *Conclusão*, principalmente envolvendo "contribuições". Em geral, quando a sentença apresenta o sinal lexical "contribuições", ela deve ser classificada como *Conclusão* e **não** como *Resultado*, principalmente se a sentença estiver apresentando

as contribuições da pesquisa num contexto generalizado. Entretanto, podem ocorrer casos em que as palavras "contribuição(ões)" podem aparecer em outro contexto. Nesses casos, você deve usar seu bom senso e decidir qual papel argumentativo se caracteriza de forma mais forte na sentença.

Pode acontecer do texto não conter nenhuma sentença de *Resultado*, principalmente porque, muitas vezes, o resultado se encontra sobreposto com o propósito.

F.3.6 Conclusão (Co)

Podem ocorrer casos de sentenças que tem o papel de "encerrar" o texto. Marque essas sentenças como *Conclusão*. Essa categoria inclui sentenças que indicam recomendações, contribuições e que expressam o valor/importância do trabalho realizado. Em geral, são sentenças mais gerais, que situam os resultados específicos do trabalho do autor dentro de um contexto de pesquisa mais amplo.

Sentenças que indicam benefícios práticos que podem resultar da aplicação dos resultados da pesquisa devem ser marcadas como *Conclusão*, assim como sentenças que enfatizam a importância teórica do estudo no avanço do estado da arte em uma área de pesquisa específica.

Veja abaixo exemplos de conclusões em que o autor situa seus resultados no contexto mais amplo de sua pesquisa e também indica benefícios práticos do trabalho:

- *"Os resultados obtidos mostram que o sistema pode substituir, a baixo custo, os meios convencionais utilizados até então para a obtenção dessas imagens."*
- *"Com a realização desses experimentos, espera-se contribuir para o estabelecimento de estratégias de teste de regressão efetivas e de baixo custo."*

Em geral, sentenças de conclusão ocorrem no final do texto. Entretanto, isso não é uma regra. Um exemplo de ocorrência de sentenças de conclusão em outras posições do texto é quando o autor intercala resultados específicos e conclusões específicas aqueles resultados, fazendo um movimento do tipo "*Resultado - Conclusão - Resultado - Conclusão...*".

Nem sempre é fácil distinguir se uma sentença é de conclusão ou de resultado. Principalmente porque o autor pode "misturar" sinais léxicos, por exemplo, usando a expressão *"concluímos que..."* para descrever um resultado bastante específico e localizado. Outro exemplo é o uso de expressões como *"os resultados mostram..."* para introduzir um efeito generalizado dos resultados num contexto mais amplo. Embora os sinais léxicos sejam um importante aliado na classificação das sentenças e tenha muito peso na decisão das categorias, fique atento à ocorrência do uso incorreto desses sinais e escolha a categoria que mais condiz com o papel argumentativo da sentença.

Pode acontecer do texto não conter nenhuma sentença de conclusão.

F.3.7 Estrutura (E)

Essa categoria não é comum em resumos por ser indicativa e muito pouco informativa. O mais comum é que essa categoria de sentenças apareça no final das introduções, mas como em alguns resumos ela também ocorre, resolvemos incluí-la no esquema.

Sentenças de estrutura descrevem o que será apresentado no artigo/tese. Podem aparecer em duas formas: orientadas às seções/capítulos do texto ou orientadas aos assuntos tratados. No primeiro caso, as sentenças costumam apresentar expressões do tipo *"Na Seção 1 comentamos..."*. No segundo caso, é comum o uso de expressões do tipo *"Inicialmente, ... A seguir... Finalmente, ..."*, descrevendo, assim, a seqüência de assuntos tratados no trabalho.

F.4 Últimas observações

Por favor, mantenha anotações sobre as dificuldades que você encontrou durante a anotação dos textos. Indique os trechos/categorias problemáticas e descreva as razões pelas quais você encontrou dificuldade.