Universidade Federal de São Carlos

Um processo para a geração de recursos lingüísticos aplicáveis em ferramentas de auxílio à escrita científica

Vanessa Silva Marquiafável

São Carlos
2007

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

Um Processo para a Geração de Recursos Lingüísticos Aplicáveis em Ferramentas de Auxílio à Escrita Científica

Vanessa Silva Marquiafável

Dissertação apresentada ao Programa de Pós-Graduação em Lingüística da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do Título de Mestre em Lingüística. Orientadora: Profa. Dra. Sandra Maria Aluisio

São Carlos, São Paulo, Brasil
2007

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

M357pg

Marquiafável, Vanessa Silva.
Um processo para a geração de recursos lingüísticos aplicáveis em ferramentas de auxílio à escrita científica / Vanessa Silva Marquiafável. -- São Carlos : UFSCar, 2007.
273 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2007.

1. Lingüística – processamento de dados. 2. Lingüística de corpus. 3. Língua inglesa - ensino. 4. Ferramenta de apoio à escrita científica. 5. Gênero textual. I. Título.

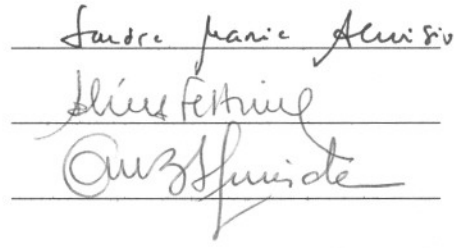
CDD: 410.285 (20ª)

BANCA EXAMINADORA

Profª. Dra. Sandra Maria Aluisio

Profª. Dra. Valéria Delisandra Feltrim

Profª. Dra. Gladis Maria de Barcellos Almeida



Agradecimentos

À Sandra, minha orientadora, pela paciência, dedicação, disponibilidade e carinho nesses seis anos de convivência. Também por seu profissionalismo, entusiasmo e ética que me proporcionaram crescer como ser humano e como profissional. Sem deixar de agradecer também pelas oportunidades, conquistas, alegrias e, sobretudo, pela confiança depositada em mim e em meu trabalho. E, claro, por ter me apresentado a Lingüística de Córpus.

À Universidade Federal de São Carlos, em especial, ao Programa de Pós-Graduação em Lingüística, a todos os professores que colaboraram com a minha formação e aos funcionários, pela atenção e profissionalismo.

À CAPES, pelo auxílio financeiro fundamental para a realização deste trabalho.

À todas as pessoas do NILC pelo constante aprendizado e amizade. Em especial, à Lívia, Carmen, Helena, Arnaldo, Leandro, Lucas, Luis e Marcelo que contribuíram direta e/ou indiretamente para que esta pesquisa fosse concretizada.

Aos amigos e amigas do programa de pós-graduação da UFSCar: Maristela, Marcela, Luciana, Marcelo, Ricardo, Valdete, Thiago, Denise, Cristiane e Andréa Monzon agradeço a amizade.

Aos meus pais, Wilmar e Maria José e à minha irmã, Flávia, pelo carinho, dedicação, paciência, incentivo e apoio incondicional nos momentos críticos e de muita alegria. E em especial ao Tiago, pela constante alegria da companhia e pelo carinho e apoio incondicionais.

À Deus, sem o qual nada disso teria sido possível.

"A vida é como jogar uma bola na parede:
Se for jogada uma bola azul, ela voltará azul;
Se for jogada uma bola verde, ela voltará verde;
Se a bola for jogada fraca, ela voltará fraca;
Se a bola for jogada com força, ela voltará com força.
Por isso, nunca "jogue uma bola na vida"
de forma que você não esteja pronto a recebê-la.
"A vida não dá nem empresta;
não se comove nem se apieda.
Tudo quanto ela faz é retribuir e
transferir aquilo que nós lhe oferecemos".
Albert Einstein

Lista de Tabelas

Lista de Figuras

Resumo

No ambiente acadêmico atual, a língua inglesa foi escolhida como a *lingua franca* da ciência nas mais variadas áreas do conhecimento. No entanto, sabe-se que a tarefa de produção de um texto científico adequado, no caso o artigo científico, não é fácil, principalmente se o escritor é iniciante nessa atividade e sua língua materna não é o inglês. Uma alternativa para esse problema é a utilização de ferramentas computacionais que apóiam as diferentes etapas do processo de escrita de um texto científico, cuja base seja formada por material lingüístico autêntico coletado de artigos científicos publicados e indexados de forma a facilitar seu rápido acesso. Dentre essas ferramentas, destacamos três em especial: o AMADEUS (*Amiable Article Development for User Support*), que apóia a escrita de artigos científicos em inglês nas áreas de Física e Computação, o SciPo, inspirado no AMADEUS, mas que apóia a escrita de teses e dissertações em português na área de Ciências da Computação e o SciPo-Farmácia, que dá suporte à escrita de artigos científicos em inglês na área de Ciências Farmacêuticas. O objetivo principal deste projeto de pesquisa foi formalizar um processo para a construção de recursos lingüísticos em inglês a serem usados em ferramentas de suporte à escrita científica semelhantes ao SciPo-Farmácia. A principal metodologia utilizada derivou da Lingüística de Corpus (usamos tanto a abordagem dirigida por córpus quanto baseada em córpus), pois a eficácia das ferramentas citadas, conforme experiências relatadas por seus desenvolvedores, está fortemente ligada ao fato de possuírem um córpus com textos da área de pesquisa do pesquisador-escritor, a partir do qual reutilizam-se trechos bem-escritos na escrita de um novo artigo. A avaliação do processo aqui proposto se deu em dois momentos: i) na avaliação da clareza e da completude dos manuais de construção de recursos lingüísticos, e ii) na avaliação da qualidade dos recursos lingüísticos produzidos e elaboração de uma estimativa do tempo gasto na construção dos recursos lingüísticos descritos por esses módulos. A estatística *Kappa* foi escolhida para medir a qualidade do material produzido nas duas etapas, a qual indicou valores entre k=0.72 e k=1,0. Esses bons resultados podem ser atribuídos ao entendimento do conteúdo dos manuais utilizados na avaliação das tarefas contidas no processo proposto. Dentre as contribuições desta pesquisa podemos citar: a possibilidade de construção de recursos lingüísticos para gerar uma ferramenta de suporte à escrita científica em inglês para várias áreas que possuem a pesquisa experimental como foco, utilizando apenas as informações contidas no processo proposto; o auxilio na divulgação, via Web, de ferramentas computacionais de suporte à escrita enquanto recurso didático a ser utilizado no ensino-aprendizado de inglês científico; a divulgação de métricas para avaliação de modelos de estruturas esquemáticas propostas; e disponibilização de córpus anotados em nível retórico para serem usados em ferramentas de processamento de língua natural ou ensino.

Abstract

Within the context of academic research, English is the *lingua franca* for various scientific disciplines. It is also widely acknowledged that producing an acceptable academic text is anything but a simple and easy task. This is particularly more acute if the author is a novice researcher and English is not his/her first language. One possible solution to minimize this difficulty is the use of writing tools to assist novice researchers during different stages of the writing process. This could involve, for instance, quick and easy access to a collection of authentic linguistic resources extracted from published scientific papers. AMADEUS (Amiable Article Development for User Support) and SciPo (Scientific Portuguese) are good examples of this type of writing tools. AMADEUS is a resource which was designed to help non-native English users write academic texts. It focuses on the fields of Physics and Computer Science specifically. SciPo is a Web critiquing system for writing theses in Portuguese and focuses on the discipline of Computer Science. A variation of Scipo is SciPo-Farmácia, which is a web-based tool to assist non-native speakers of English in writing scientific papers in the field of Pharmaceutical Sciences.

The main purpose of this dissertation is to elaborate a semi-automatic process to generate the necessary English linguistic resources required by supporting writing tools, such as the ones mentioned above. The primary aim is to enable researchers from various disciplines to develop their own aiding writing tool, customized to his/her specific field, with no need to refer to linguists, computer scientists and/or academic writing specialists for help. The semi-automatic process proposed here has been designed to include the knowledge which would be provided by these specialists. The main methodology adopted in this research derives from the discipline of Corpus Linguistics (we have used both corpus-based and corpus-driven approaches). This choice relies on the assumption that the success of such tools is strongly related to the corpus from which users collect well-written text extracts so that they can be recycled and reused in the text being produced. The semi-automatic process was evaluated in two ways: i) clearness and completeness of the manuals describing the linguistic resources and ii) quality of the linguistic resources generated and estimated time for developing all the necessary linguistic resources. For measuring the quality of the two evaluation stages, we have used the statistical system *Kappa.* The results ranged from k=0.72 e k=1.0. These figures can be interpreted as a good understanding of the tasks described in the manuals evaluated. The present research proves relevant in a number of aspects. It opens up the possibility of generating a computational tool to assist non-native English speakers in writing academic texts in any experimental field, by using the knowledge from the semi-automatic process only. It also promotes the use of supporting writing tools as didactic resource for teaching-learning scientific English and the use of metrics to evaluate rhetorical structure models. Last but not least, it produces a rhetorically annotated corpus which may be used for teaching-learning purposes or in natural language processing.

Sumário

1. Introdução

1.1 Contexto

No ambiente acadêmico atual, a língua inglesa foi escolhida como a *língua franca*[1] da ciência nas mais variadas áreas do conhecimento. No entanto, sabe-se que a tarefa de produção de um texto adequado, isto é, que atenda às expectativas da comunidade acadêmica, não é fácil, principalmente quando o escritor é iniciante na atividade de produção de escrita científica e sua língua materna não é o inglês. De fato, para que um artigo científico tenha sucesso na submissão, sendo aceito para publicações em boas conferências e revistas, o crivo dos pares é importante, pois segundo Swales (1990), são os membros mais especialistas da comunidade que ditam as convenções textuais que devem ser seguidas, desde a organização da estrutura textual até o conjunto de expressões lingüísticas que devem ser empregadas.

Segundo Aluísio (1995), a dificuldade acima citada pode ser explicada pelos seguintes fatores: (1) alta sobrecarga cognitiva sentida pelo escritor no momento de formulação de suas idéias quando tem de lidar com a complexidade naturalmente existente no processo de escrita em uma língua não-materna; (2) bloqueio na escrita do primeiro rascunho, uma vez que o autor, muitas vezes, escreve e não divulga seu artigo científico porque seus possíveis erros podem ser descobertos pelos demais, porém, segundo Secaf (2001), é assumido *que para aprender mais e poder crescer é necessário escrever, mesmo que as imperfeições existam e apareçam, pois só assim nos aperfeiçoamos e nos desenvolvemos*; (3) desconhecimento das convenções específicas do gênero[2] científico (retórica científica) ou o uso inadequado das mesmas, as quais fazem referência tanto ao tipo de informação que deve ser incluído no texto quanto ao formato de apresentação dessa mesma informação; e (4) não ciência das idiossincrasias existentes em cada comunidade de pesquisa, que dita algumas variações na estruturação e seleção dos conteúdos que devem ser adicionados ao texto.

Visando auxiliar principalmente pesquisadores iniciantes na produção escrita, foram escritos livros especializados e desenvolvidos *sites* e *softwares*, que podem corrigir erros

[1] Segundo Forattini (1997:4), língua *franca* significa a maneira de expressão escrita ou oral comum a falantes nativos de diferentes línguas que a elegem como meio de comunicação. Nas ciências, o inglês tem se destinado a agilizar a divulgação das pesquisas entre os cientistas, em vista da grande quantidade de conhecimento científico produzido pelos países falantes de inglês.

[2] No momento, compreendemos gênero como as formas convencionais de textos associadas a situações sociais específicas, reconhecidas como tais pela comunidade de usuários da língua que compartilham do contexto sócio-cultural do texto. Mais informação sobre esse conceito pode ser encontrada na Seção 3.4.

gramaticais e de estilo. Entretanto, estes não atacam a dificuldade principal de produção de uma primeira versão que consiga descrever adequadamente o cerne de um trabalho de pesquisa.

Uma alternativa para esse problema pode ser a utilização de ferramentas computacionais, que apóiam as diferentes etapas do processo de escrita (planejamento, composição e revisão) de um texto científico, cuja base seja formada por material lingüístico autêntico e indexado de forma a facilitar seu rápido acesso. Dentre as ferramentas existentes que apresentam tais recursos podemos citar duas em especial: o ambiente AMADEUS - *Amiable Article Development for User Support* (Aluísio, 1995), com suas ferramentas de Referência (Fontana *et al*, 1993), Suporte (Aluisio e Oliveira Jr., 1995) e de Crítica (Aluisio *et al*, 2001), que apóiam a escrita de artigos científicos em inglês nas áreas de Física e Computação, e o SciPo[3] – *Scientific Portuguese* - (Feltrim, 2004), ferramenta inspirada no AMADEUS, mas que apóia a escrita de teses e dissertações em português na área de Computação.

A utilização das mesmas tem trazido uma comprovada familiarização com a retórica do gênero científico e diminuição da sobrecarga cognitiva na fase de tradução das idéias em texto (Fontana, 1993; Feltrim, 2004), podendo tornar o texto produzido mais fluente. Esse tipo de construção e indexação das informações apresentadas por essas ferramentas computacionais de suporte à escrita científica possibilitaram aos seus usuários escritores-pesquisadores: (a) ter um insumo lingüístico adequado às suas necessidades; (b) adaptar as expressões que julgar adequadas ao seu texto e (c) reproduzir fórmulas ou expressões-padrão quando estiver escrevendo sobre determinados aspectos de seu trabalho. Esses fatos motivaram também a construção de uma segunda ferramenta Web de auxílio à escrita científica em inglês na área de Ciências Farmacêuticas, o SciPo-Farmácia[4], que vem sendo utilizada com sucesso em cursos de Escrita Científica[5] na Faculdade de Ciências Farmacêuticas da USP/São Paulo e também no Instituto de Física de São Carlos (USP).

Tanto o ambiente AMADEUS quanto as ferramentas SciPo e SciPo-Farmácia foram construídas conforme os seguintes passos:

1. Seleção de um córpus (conjunto de artigos científicos e no caso do SciPo, de dissertações) bem escritos da área de pesquisa na qual se deseja escrever;

[3] http://www.nilc.icmc.usp.br/~scipo/
[4] http://www.nilc.icmc.usp.br/scipo-farmacia/
[5] Material do curso de escrita científica da USP: http://www.nilc.icmc.usp.br/coteia/show.php?wikipage_id=14.

2. Identificação e anotação dos componentes da estrutura esquemática e das estratégias retóricas contidas no discurso científico em cada sentença de cada seção do córpus coletado;

3. Anotação das expressões-padrão e dos marcadores discursivos[6] nos textos, a fim de auxiliar na prática das convenções lingüísticas e de estilo desse gênero textual;

4. Inclusão de todos esses recursos lingüísticos (córpus anotado, componentes da estrutura esquemática e estratégias retóricas das seções) em uma ferramenta computacional.

A Figura 1.1 mostra as sentenças de um *abstract* cujos componentes retóricos (Background, Purpose, etc.) foram anotados. No SciPo-Farmácia, assim como no SciPo e no AMADEUS, esses componentes são sub-especificados em estratégias retóricas, tais como: (a) declarar relevância do tópico; (b) familiarizar termos, objetos ou processos; (c) listar critérios ou condições; (d) indicar/descrever materiais ou métodos. Alguns marcadores discursivos são destacados para mostrar a ligação entre os tipos de sentenças e cláusulas que eles propiciam.

[6] Nesta pesquisa, os marcadores discursivos são as palavras e expressões vindas principalmente da classe dos advérbios e locuções adverbiais, das conjunções e sintagmas preposicionais e que têm o papel duplo de indicar uma relação entre o segmento em que ele está presente e o anterior/posterior a ele, orientando o leitor para uma determinada direção na interpretação desses segmentos, por exemplo, *however*, *despite of this*, etc.

Antioxidants Inhibit Indoleamine 2,3-Dioxygenase in IFN-gamma-Activated Human Macrophages: Posttranslational Regulation by Pyrrolidine Dithiocarbamate Thomas SR, Salahifar H, Mashima R, Hunt NH, Richardson DR, Stocker R.
Background
Induction of the heme-containing indoleamine 2,3-dioxygenase (IDO) by IFN-gamma is implicated in anti-microbial and pro-inflammatory activities of human macrophages. Antioxidants can modulate the expression of immune and inflammatory genes, and pyrrolidine dithiocarbamate (PDTC) is a frequently used antioxidant to inhibit the transcription factor NF-kappaB.
Purpose
Here we show that IFN-gamma treatment of human monocyte-derived macrophages (hMDMs) increased the proportion of oxidized glutathione.
Main Results
PDTC attenuated this increase and inhibited IDO activity, **although** it increased IDO protein expression and did not affect IDO mRNA expression and enzyme activity directly. Other antioxidants, 2-ME, ebselen, and t-butyl hydroquinone, inhibited IDO protein expression. Similar to PDTC, the heme biosynthesis inhibitor succinylacetone (SA) and the iron-chelator pyridoxal isonicotinoyl hydrazone inhibited cellular IDO activity without affecting protein expression, **whereas** addition of hemin or the heme precursor delta-aminolevulinic acid increased IDO activity. **Also**, incubation of IFN-gamma-activated hMDM with delta-[14C]-aminolevulinic acid resulted in the incorporation of label into immunoprecipitated IDO, a process inhibited by PDTC and SA. **Furthermore**, supplementation of lysates from PDTC- or SA-treated hMDM with hemin fully restored IDO activity to control levels, and hemin also reversed the inhibitory action of SA but not PDTC in intact cells. Together these results establish a requirement for de novo heme synthesis for IDO activity in IFN-gamma-activated hMDM.
Conclusion
They show that, similar to other pro-inflammatory proteins, the activity of IDO is modulated by antioxidants **though** in the case of PDTC this takes place posttranslationally, in part by limiting the availability of heme for the formation of holo-IDO.

Figura 1.1: Abstract do J Immunol. 2001 May 15; 166(10): 6332-40, com suas sentenças segmentadas, apresentando quatro componentes esquemáticos (contexto, propósito, resultados principais e conclusão) e marcadores discursivos em negrito. Mais detalhes sobre os componentes retóricos podem ser encontrados na Seção 3.5.2. Sobre as estruturas retóricas, mais informações em 3.5.3.

1.2 Motivações

No conjunto das experiências realizadas com as ferramentas citadas, pôde ser constatada que a boa aceitação das mesmas por parte de seus usuários se deve fortemente ao fato de possuírem um córpus com textos específicos da área de pesquisa do usuário-escritor. A partir desse córpus puderam reutilizar trechos bem-escritos na elaboração de um novo artigo científico (Feltrim, 2004; Aluísio, 1995; Schuster *et al*, 2005).

O fato de divulgações de trabalhos relevantes se efetuarem prioritariamente em revistas especializadas e mundialmente veiculadas, que adotam o inglês como língua padrão, abriu caminhos para a possibilidade de se criar uma ferramenta computacional que auxiliasse o processo de escrita de trabalhos nesse idioma. A razão disso é que a comunidade acadêmica, em geral, necessita divulgar de maneira adequada e rápida o conhecimento científico por ela produzido. Entretanto, as ferramentas de suporte à escrita citadas a pouco podem atender adequadamente apenas três comunidades científicas, a Farmácia, a Computação e a Física. Assim, uma questão que se coloca é a possibilidade de estender esse auxílio computacional a pesquisadores de outras áreas do conhecimento.

Além disso, a construção desse tipo de ferramentas requer, em geral, um grupo de especialistas (cientista da computação, lingüista(s), especialista(s) em escrita científica, especialista(s) na área em que a ferramenta será construída) e um grupo de recursos lingüísticos (córpus anotado[7], teorias lingüísticas, teorias sobre escrita científica). Uma segunda questão que pode surgir aqui é sobre a possibilidade de se diminuir ou facilitar o acesso a essas variáveis envolvidas na construção de uma ferramenta de suporte à escrita.

Nesse contexto, surgiu no Núcleo Interinstitucional de Lingüística Computacional - NILC[8] - o projeto para criar um A*mbiente Web Gerador de Ferramentas Computacionais de Suporte à Escrita Científica em Inglês*. Para que tal projeto se concretizasse, foi necessária a colaboração de dois projetos de mestrado, um na área de Lingüística e outro na área de Ciências da Computação, desenvolvidos sob a mesma orientação. Um deles, que é o projeto descrito nesta dissertação, visou estabelecer um conjunto de etapas para se construir recursos lingüísticos (CECARL[9] doravante), assim como fornecer diretrizes para acoplá-los adequadamente em uma ferramenta de suporte à escrita genérica, semelhante à ferramenta SciPo-Farmácia, que foi baseada na Ferramenta de Suporte do AMADEUS. O segundo mestrando ficou responsável por construir as ferramentas computacionais necessárias na execução de duas etapas automáticas contidas no processo elaborado. Assim, para que esse Ambiente Web Gerador se concretize, resta, ainda, a um terceiro trabalho futuro a criação de uma interface gráfica na qual estariam automatizadas todas as etapas (CECARL) resultantes deste mestrado. Assim, o usuário desse Ambiente Web Gerador seria guiado de maneira automática desde a tarefa de compilação de um córpus até o momento de alocação, que será também automática, dos recursos lingüísticos em diretórios corretos de um dado servidor.

---

[7] Anotação consiste na inserção de etiquetas ou cabeçalhos em um dado texto. As etiquetas, por exemplo, podem fornecer informações nos níveis sintático, morfológico, semântico, etc de cada palavra, frase, oração, parágrafo, etc, de um dado texto. Os cabeçalhos também podem fornecer esse tipo de informação, bem como a autoria do texto, local de disponibilização do texto e assim por diante. Na seção "Uso de Córpus" há um exemplo de texto com cabeçalho e na última página dos Apêndices 1, 2, 5, 6, 7 e 8 exemplo de um texto anotado com informações retóricas.

[8] NILC - Núcleo Interinstitucional de Lingüística Computacional. Grupo interdisciplinar de lingüistas e cientistas da computação, criado em 1993 para desenvolver pesquisas e projetos relacionados com Lingüística Computacional e processamento de Língua Natural, tais como desenvolvimento de léxicos e córpus, sumarização automática, tradução automática e ferramentas de suporte à escrita. Localiza-se no Instituto de Ciências Matemáticas e de Computação da USP São Carlos e pode ser acessado pelo link: http://www.nilc.icmc.usp.br/nilc/index.html.

[9] CECARL – Conjunto de Etapas para Criação e Alocação de Recursos Lingüísticos.

1.3 Objetivo

Este projeto de pesquisa surge com o intuito de *formalizar um conjunto de etapas para a construção de recursos lingüísticos aplicáveis em um ambiente Web de suporte à escrita científica em inglês*.

Usando o CECARL, pesquisadores de diferentes áreas do conhecimento podem construir esse tipo de ambiente de suporte à escrita sem o auxílio de um grupo de especialistas – lingüistas, engenheiros do conhecimento e cientistas da computação – necessários em geral, dado que o conhecimento deles está incluído no CECARL. Sua formalização culminou em um processo para a geração de ferramentas Web de suporte à escrita, composto por uma seqüência de 11 passos. O CECARL apresenta a ordem e quais atividades devem ser realizadas para se obter uma ferramenta Web de suporte à escrita científica em língua inglesa, com funções semelhantes às apresentadas pelo SciPo-Farmácia. Algumas sugestões de novos recursos foram feitas para serem inseridas na ferramenta genérica (isto é, adaptável para qualquer área) gerada com o CECARL, tornando-a um pouco diferente do SciPo-Farmácia. Atualmente, tais etapas encontram-se disponíveis no ambiente Plonetarium da Fapesp, junto dos *links* para as ferramentas de tarefas automáticas[10].

Assim, o pressuposto a ser provado neste trabalho é o de que público-alvo deste trabalho consiga construir recursos lingüísticos para sua própria ferramenta computacional de suporte à escrita científica com a ajuda apenas das etapas e manuais elaborados neste projeto. Com o CECARL, ele pode construir e desfrutar dos benefícios de uma ferramenta customizada segundo as necessidades da comunidade ou área acadêmica da qual participa. Além disso, acreditamos também que não só o uso, mas também a confecção dos recursos lingüísticos (córpus anotado quanto aos componentes da estrutura esquemática de textos científicos, expressões-padrão e marcadores discursivos) necessários em tais ferramentas possam favorecer o aprendizado da escrita acadêmica pelo aprendiz-autor. Assim como ocorre na abordagem de ensino de Inglês com Propósitos Específicos (Swales, 2003), este passa a ter contato com o vocabulário, estruturas e gênero textual pertinentes à suas necessidades de aprendizado de escrita durante a confecção e no uso da ferramenta de suporte.

Acreditamos também que o público-alvo desta pesquisa seja, a princípio:

1) Professores, pesquisadores, especialistas envolvidos no ensino-aprendizagem de escrita científica. A ferramenta de suporte gerada com o CECARL poderia ser utilizada por esses

[10] http://gen-writingtool.incubadora.fapesp.br/portal.

profissionais enquanto mais uma opção de recurso didático a ser utilizado em aulas sobre escrita científica. Esses profissionais poderiam construir uma pequena base de dados para apresentação das funcionalidades, da ferramenta a seus alunos, a qual poderia ser acrescida de recursos produzidos pelos próprios alunos, pois, como dissemos, não só o uso como também a confecção de tal ferramenta pode auxiliar no aprendizado, aqui tratado.

2) Orientadores interessados na melhoria da escrita científica de seus alunos (Mestrandos ou doutorandos), que necessitam publicar em inglês e, portanto, adquirir uma noção adequada da organização/funcionamento de um artigo científico nesse idioma. Assim, o processo de construção de uma ferramenta de auxílio à escrita científica poderia funcionar como uma metodologia de ensino-aprendizagem da escrita científica focado nas necessidades específicas da área de atuação dos alunos. Segundo experimentos realizados com ferramentas de auxílio à escrita (Feltrim, 2004; Schuster *et al*, 2005), a tarefa de construção dessa base de casos já pode ser considerada o primeiro passo no aprendizado de escrita científica, uma vez que é necessário entender o funcionamento das partes das seções de um artigo científico e de seus constituintes para depois identificá-las e organizá-las na base de casos da ferramenta a ser construída.

3) Centros de escrita científica ou bibliotecas que visam dar auxílio especializado sobre escrita científica, aos quais os estudantes podem recorrer para obter auxílio e orientação sobre como redigir corretamente seus artigos científicos em inglês.

4) Escritor experiente (por exemplo, um pesquisador sênior, um doutor) que queiram desenvolver uma base de casos de textos científicos, a qual possam acessar de maneira organizada, isto é, por meio de uma interface amigável, sempre que desejarem escrever textos científicos em inglês. Se desejarem também, os pesquisadores experientes podem se restringir apenas à confecção dos recursos lingüísticos, que poderão ser salvos em diretórios organizados segundo a forma que lhes parecer mais conveniente.

E que também tenha os seguintes níveis de conhecimento:

1) Familiaridade com computador, isto é, que não tenha dificuldades em realizar comandos simples do Windows, como: copiar e colar, escrever uma linha de comando em DOS e fazer *downloads* de pacotes de arquivos que serão instalados no computador, como os citados na seção sobre o acoplamento de recursos lingüísticos em uma ferramenta de suporte à escrita genérica, isto é, sem recursos lingüísticos.

2) Familiaridade com os componentes da estrutura esquemática de artigos científicos. Isso implica dizer que é necessário ter uma noção mínima (superficial) da organização destes

elementos, pois há manuais (Apêndices 1, 2, 5, 6, 7 e 8) que os apresentam de maneira formalizada.

3) No mínimo, nível de inglês intermediário. Como se trata de recursos lingüístico para uma ferramenta de suporte à escrita científica em inglês, é necessário coletar textos escritos e inglês e uma dificuldade no entendimento do conteúdo destes textos pode prejudicar a identificação de recursos lingüísticos nos mesmos.

4) O usuário do CECARL seja da área de conhecimento na qual a ferramenta de suporte à escrita será construída, pois o conhecimento prévio do usuário sobre a área o auxiliará na tarefa de entendimento do conteúdo dos textos para posterior identificação de determinadas funções, características e recursos lingüísticos.

5) A última etapa do CECARL exige que todo o conhecimento levantado nas 10 etapas anteriores seja armazenado em diretórios adequados de um dado servidor, pois se trata de uma ferramenta Web de auxílio à escrita científica em inglês. Essa etapa é a única que exigiria, talvez, o auxílio de um cientista da computação ou então de uma pessoa que tenha as permissões e senhas do servidor no qual será alocada a ferramenta gerada, para a sua plena execução.

1.4 Metodologia

A primeira fase de elaboração do CECARL se constituiu na investigação de ferramentas de suporte à escrita científica existentes na literatura, as quais pudessem ser facilmente adaptadas para os fins deste trabalho. Para isso, considerou-se o nível de auxílio proporcionado por cada ferramenta, o tipo de categorização textual adotado por cada um desses sistemas e o custo/benefício de implementação de cada uma delas. A razão é que o público-alvo de nosso projeto são pessoas que construirão suas próprias ferramentas de auxílio à escrita e que, na maioria das vezes, não possuírem conhecimento especializado no domínio da computação. Dentre as ferramentas estudadas, o SciPo-Farmácia foi escolhido como inspiração para gerar o Scientific Writing, nossa ferramenta de suporte genérica, isto é, sem recurso lingüístico. Inspiração porque também propusemos algumas alterações na interface do SciPo-Farmácia, incrementando a sua funcionalidade. Essas alterações compreendem a inclusão de alguns recursos, a nosso ver, interessantes para os aprendizes de língua estrangeira: uma lista de expressões formulaicas organizadas por funções que podem

desempenhar em um artigo científico e um concordanciador[11] (recurso que poderá vir a ser adicionado a esse ambiente).

O próximo passo foi consultar a literatura especializada sobre as considerações históricas, conceituais e de aplicação da Lingüística de Córpus e de seu objeto de estudo, o córpus. Isso porque a confecção de um córpus com textos científicos é um dos maiores gargalos para o desenvolvimento de uma ferramenta de suporte à escrita, que segue o modelo da ferramenta SciPo-Farmácia, escolhido como inspiração deste trabalho.

A terceira fase de elaboração do CECARL se deu com a investigação sobre o que os principais teóricos de gêneros discursivos dizem a respeito do artigo científico, gênero que será abordado em nossa metodologia de geração de uma ferramenta de suporte à escrita. Durante essa investigação, procuramos delinear uma estrutura composicional recorrente em artigos científicos, indicando formas que lhe são peculiares independentemente da área científica escolhida, tais como os componentes da estrutura esquemática, as estratégias retóricas, as expressões formulaicas, os marcadores discursivos e os termos específicos de uma área que aparecem em textos do gênero científico.Isso se deu uma vez que essas escolhas lingüísticas ocorrem em função das idiossincrasias existentes dentro de um contexto sociocultural, no caso o acadêmico. Posteriormente, essas particularidades da composição textual (Swales, 1990) de artigos científicos foram apresentadas na forma de manuais (Apêndices de 1 a 8) e em linguagem simples, isto é, de modo que um não-especialista nas áreas desses elementos lingüísticos citados possa compreendê-los e identificá-los nos artigos científicos da área em que atua.

Por fim, com base nas teorias investigadas e em um estudo de caso - a construção e implementação da seção "Metodologia" do SciPo-Farmácia - deu-se a elaboração das etapas necessárias às etapas de construção de recursos lingüísticos para ferramentas de suporte à escrita científica, proposta desta pesquisa.

Após sua elaboração, foram realizadas duas avaliações de algumas etapas do CECARL. O objetivo da primeira fase foi avaliar a *Clareza e Completude dos manuais de construção (anotação) de recursos lingüísticos* a serem gerados. Para isso, três pessoas realizaram as etapas de: Identificação dos componentes da estrutura esquemática, 2) Avaliação da

[11] Concordanciador: é uma ferramenta informatizada que gera uma listagem na qual um dado item (palavra isolada, composta, estrutura, etc.) aparece com palavras (co-textos) ao seu redor (Berber-Sardinha, 2000).

qualidade[12] das seções de artigos científicos com o auxílio de uma rubrica[13] 3) Identificação de marcadores discursivos e 4) Identificação das estratégias retóricas em 5 Resumos da área de Ciências da Computação, com o auxílio de manuais com informações respectivas a cada um desses recursos lingüísticos citados. O propósito de se avaliar a clareza e completude desses manuais é observar se estão completos e claros o suficiente para serem utilizados pelo público-alvo deste projeto de pesquisa. Para isso, os avaliadores ficaram livres para anotarem nos próprios manuais os trechos considerados confusos, se os termos empregados não são claros e se há informação suficiente/insuficiente. Ao final dessa etapa, concluímos que as informações contidas nos manuais estão adequadas e suficientes para serem utilizadas pelo nosso público-alvo, uma vez que os resumos foram anotados com grau de concordância *Kappa*[14]= 0.835 para a tarefa de identificação dos componentes da estrutura esquemática e *Kappa* = 0.779 para a identificação das estratégias retóricas. Estes valores de *Kappa*, por serem maiores que 0.75 são considerados excelentes (Orwin, 1994). Entretanto, as sugestões dos avaliadores foram incluídas nos manuais para torná-los ainda mais claros e completos.

Já a segunda fase de avaliação objetivou avaliar a *Consistência da Anotação de uma dada seção de artigo científico e uma estimativa do Tempo Gasto a ser Gasto na Confecção de Recursos Lingüísticos* para uma ferramenta gerada com as nossas etapas. Assim, foram convidadas três duplas, cada uma de uma área diferente (Computação, Engenharia de Produção e Lingüística), para participarem do processo de replicação dos procedimentos de 1) Identificação dos componentes da estrutura esquemática, 2) Identificação das estratégias retóricas, 3) Identificação de marcadores discursivos e 4) Avaliação da qualidade textual, em 15 *abstracts* da área em que as duplas atuam, com o auxílio dos manuais testados na primeira fase e revistos. A verificação da consistência na anotação dos recursos lingüísticos produzidos nas duplas foi feita novamente com o auxílio da estatística *Kappa.* Os resultados apontaram valores entre k=0.72 e k=1.0, o que mostra que as duplas tiveram um grau de concordância entre bom e excelente, resultado do entendimento do conteúdo trazido pelos manuais

12 Neste trabalho, quando mencionarmos a avaliação da qualidade das seções de artigos científicos com o auxílio de uma rubrica, nos referimos a averiguação do modo como os componentes da estrutura esquemática de um artigo científico e a ordem lógica deles estão em acordo com as especificações de especialistas em escrita científica, como Swales (1990) e Weissberg (1990).

13 Rubrica: São critérios (grupo de dimensões) para se avaliar, no caso, um texto científico. Mais detalhes ver Capítulo 3, seção 3.5.7.

14 Método estatístico que foi utilizado pela primeira vez em 1995 por Isard e Carletta na análise de discurso e de diálogo. Essa estatística tem sido utilizada como teste para tarefas de classificação nas quais alguns ou vários anotadores ou juízes têm como função atribuir classes a um grupo de itens. Ela auxilia também a descobrir problemas de anotação surgidos durante o processo, bem como de teste de qualidade e abrangência do conjunto de etiquetas utilizadas, do manual de anotação consultado e do corpus de treinamento. Em suma, pode-se dizer que essa estatística Kappa auxilia a verificar o grau de replicabilidade de uma dada tarefa cf. Capítulo5).

utilizados na avaliação. Quanto à *Estimativa do Tempo Gasto na Construção desses Recursos*, essa foi realizada com base na anotação feita pelos próprios colaboradores do tempo gasto por eles para concluir os procedimentos requisitados pelos manuais em 15 *abstracts* de suas respectivas áreas. Conforme observado, a média de tempo gasto para anotar determinados recursos lingüísticos em abstracts ficou em 05h29min. Vale dizer, que nessa estimativa não foi considerado o tempo que seria gasto na compilação e formatação dos abstracts anotados.

1.5 Organização do Trabalho

Esta proposta de dissertação está organizada em seis capítulos. No Capítulo 2, é apresentada uma revisão das ferramentas de auxílio à escrita existentes, com a descrição de suas abordagens e respectivos resultados de avaliação de seus autores. No Capítulo 3, apresenta-se uma revisão das teorias que auxiliaram na execução deste projeto, entre elas a Lingüística de Córpus e suas metodologias de investigação lingüística, as concepções do conceito de gênero empregadas pelos principais teóricos bem como suas implicações para esse trabalho, as principais características de um artigo científico e formas de estruturar as informações padronizadas contidas nesse tipo de texto. No Capítulo 4, apoiando-nos nas reflexões, ferramentas computacionais e teorias trazidas por esse trabalho, temos a elaboração da proposta de trabalho deste projeto: um conjunto de etapas para a geração de ferramentas de suporte à escrita de artigos científicos em inglês de uma dada área de especialidade. Apresentamos também nesse capítulo a ferramenta de suporte à escrita genérica, Scientific Writing. No Capítulo 5, apresentamos os resultados das duas avaliações feitas sobre o CECARL, e no último, o Capítulo 6, as considerações finais, limitações e contribuições do projeto apresentado.

2. Ferramentas de Auxílio à Escrita Científica

2.1 Considerações Iniciais

A língua inglesa tem atualmente o status de *língua franca* da ciência (Forattini, 1997) e da tecnologia (Johns & Dudley-Evans, 1991), pois artigos científicos produzidos nessa língua se tornaram um dos principais meios de divulgação e distribuição de conhecimento entre pesquisadores de todo o mundo. *No entanto, a tarefa de produção de um texto científico em inglês não é fácil, principalmente quando o escritor ainda é iniciante e/ou inexperiente nessa atividade e sua língua materna não é o inglês* (Aluísio, 1995).

Para auxiliar os pesquisadores, principalmente os novatos, têm sido editados livros especializados no/para o ensino de escrita científica, como por exemplo, os trabalhos de Swales (1990) e de Weissberg & Buker (1990). Da mesma forma, têm sido desenvolvidas ferramentas computacionais de apoio aos processos de planejamento, composição e revisão de um texto. No entanto, o processo de escrita seja ela científica ou não, envolve diferentes etapas, como é discutido por Hayes & Flower (1980[1] *apud* Feltrim, 2004), de modo que projetar e desenvolver um ambiente que auxilie eficientemente todo o processo pode ser uma tarefa bastante complexa. Portanto, as ferramentas que até o presente momento têm sido desenvolvidas procuram auxiliar em alguns dos aspectos do processo de escrita, geralmente atacando um problema específico ou uma categoria de problemas aparentemente semelhantes e que, por isso, podem ser tratados em conjunto.

Há que se destacar neste trabalho que é reconhecida a utilidade de sistemas que auxiliam o pós-processamento de um texto, como os corretores ortográficos, estilísticos e gramaticais. Porém, para se ter acesso a esse tipo de ajuda, o autor de um texto precisa ser capaz de compor seu primeiro rascunho, tarefa esta que, segundo Swales (1990), mesmo para escritores experientes não é nada fácil. Sendo assim, a classe de ferramentas que visa auxiliar o processo de composição e não apenas o de revisão textual será destacada nesta pesquisa.

1 HAYES, J.R.; FLOWER, L.S. Writing as Problem Solving. In Visible Language, XIV v. 4, 1980.

2.2 Escrita de artigos científicos por não-nativos e ferramentas de auxílio a essa tarefa

Existem diferentes estudos a respeito das dificuldades enfrentadas por não-nativos do inglês ao escreverem textos científicos nessa língua, como os trabalhos de Bazerman, 1988; Swales, 1990b; Gosden, 1995; Mauranen, 1993 e Flowerdew, 1999a, por exemplo. As dificuldades levantadas por esses estudos apontam para dois fatores principais. O primeiro é a diversidade sociolingüística, já que as línguas diferem consideravelmente entre si em termos de fonologia, sintaxe, léxico e pragmática (Wolfson, 1989). Como conseqüência, existem diferentes línguas que utilizam diferentes padrões e elementos lexicais para organizar seu discurso (Henner-Stanchina, 1985[2] *apud* Mirahayuni, 2002). O segundo fator relacionado com o desempenho de uma boa escrita tem a ver com a relação existente entre compreensão e produção: é bem aceito o fato de que a compreensão é mais fácil do que a produção. De fato, os aprendizes de línguas estrangeiras ou maternas parecem estar aptos a entender as funções e formas da língua, as quais eles normalmente não utilizam em suas produções (Fontana *et a*l, 1993). Entre os tipos mais comuns de problemas podem ser citados: (1) a falta de encadeamento lógico e claro entre as sentenças; (2) o desenvolvimento dos tópicos entre as sentenças de maneira incoerente; (3) o uso de sentenças gramaticalmente incorretas e (4) a falta de habilidade ao manipular a linguagem utilizada para sustentar o que é dito (Gosden,1995:48[3] *apud* Mirahayuni, 2002). James (1984) é outro estudioso dos problemas relacionados à escrita de não-nativos e aponta entre esses, particularmente, a falta de coesão entre sentenças consecutivas, o que afeta além da coesão textual, a sua coerência. James (1984; 1989) também identificou em seus estudos que erros localizados, isto é, aqueles que podem ser tratados isoladamente, como por exemplo a regência verbal, não causam tantos danos à comunicação como aqueles que afetam o significado global de um texto ou de grandes trechos de texto.

Diante de tal situação, podem ser encontrados diferentes tipos de auxílios especializados como livros, *sites* na Web e ferramentas computacionais de auxílio ao pré e pós-processamento de textos. No entanto, é sabido que as dificuldades "globais" apresentadas impossibilitam a escrita de um texto que possua um mínimo de qualidade suficiente para ser apenas corrigido por corretores gramaticais e estilísticos ou dicionários. Um alternativa proposta por Oliveira Jr. *et al* (1992) e seguida nos trabalhos de Fontana *et al* (1993), Aluisio

[2] HENNER-STANCHINA, C. From reading to writing acts. In *Discourse and Learning.* RILEY, P (Ed.), Burnt Mill, Longman, 1985.
[3] GOSDEN, H. Success in research article writing and revision: A social-constructionist perspective. *English for Specific Purposes*, v. 14, p. 37-57, 1995.

& Oliveira Jr (1995), Aluísio (1995) entre outros do mesmo grupo de pesquisadores é a utilização de ferramentas computacionais que forneçam materiais lingüísticos autênticos indexados (categorizados) de acordo com os componentes da estrutura esquemática e estratégias retóricas de um texto científico de modo a facilitar um acesso rápido a esse tipo de informação para posterior reutilização. Esse tipo de reuso lingüístico tem sido a essência de determinadas ferramentas computacionais de auxílio à escrita, que serão descritas a seguir com mais detalhes.

2.3 Abstract Helper

O *Abstract Helper* (AH doravante) visa dar suporte na estruturação e realização lingüística de um resumo em inglês (*abstract)*. Segundo Narita (2000a e b), a motivação para a construção desse tipo de sistema se deveu a uma tentativa de solucionar problemas de escrita de textos científicos em inglês, provenientes não só do idioma como também da organização textual dos abstracts, enfrentados por escritores japoneses.

O AH utiliza uma abordagem de reutilização de textos autênticos, os quais servem como modelos organizacionais e estilísticos para a produção de novos textos. Essa abordagem permite ao usuário: (1) acessar um córpus paralelo Inglês/Japonês, anotado e constituído por 539 exemplos de *abstracts* provenientes de publicações relevantes na área de Ciência da Computação, considerados bons exemplos de organização e de estilo; (2) encontrar um bom modelo para elaborar seu *abstract* ou sentença; e (3) acessar rapidamente os recursos lingüísticos relevantes ao contexto do texto a ser produzido, por exemplo, os marcadores discursivos utilizados em um *abstract*.

De acordo com Narita *apud* Feltrim (2004), esse córpus paralelo foi anotado utilizando um conjunto de etiquetas organizadas em dois níveis: resumo e sentença. No primeiro nível foi privilegiada a identificação da macro-estrutura contida em um resumo, como o tipo de resumo (por exemplo, o resumo apresenta a proposta de um novo sistema, a proposta de um novo algoritmo, etc.) e o tipo de estrutura organizacional contido em um resumo: (1) resumos que iniciam com uma sentença-tópico; (2) resumos com a sentença-tópico no meio do texto; (3) resumos que terminam com a sentença-tópico e (4) resumos multiparágrafos. No nível de sentença foram identificados/anotados os papéis de cada sentença constitutiva de um resumo: (1) introdutória; (2) tópico; (3) explanatória; (4) verificação; (5) suplementar; (6) conclusão e

(7) fechamento. Outro recurso lingüístico disponibilizado por esse sistema é uma base de colocações[4] em inglês extraídas do córpus e checadas manualmente.

Esses recursos disponibilizados pelo AH podem ser acessados por diferentes tipos de busca: (1) por resumos, (2) por sentença, (3) por padrão de sentença - que considera além do papel atribuído à função da sentença, as características sintáticas e lexicais das mesmas - e (4) por colocação. Não existe uma ordem pré-estabelecida de acesso a esses recursos, apesar de se esperar que o usuário os utilize por meio de um processo descendente, isto é, primeiramente busque um modelo de resumo, a seguir exemplos de sentenças e por fim, informações sintáticas e lexicais.

Em um experimento realizado com usuários acadêmicos (Narita 2000b) foi constatada uma boa aceitação do AH por parte de seus usuários. Foi apontada também a necessidade de melhorias quanto à diversificação do domínio do córpus utilizado e a inclusão de um número maior de exemplos.

2.4 AMADEUS – Amiable Article Development for User Support

Um outro tipo de abordagem de suporte à escrita científica em inglês como língua estrangeira, que também se baseia em um repositório de recursos lingüísticos, sugere a seguinte proposta de auxílio à escrita do primeiro rascunho:

1. Seleção de um conjunto de artigos científicos bem escritos da área de pesquisa para a qual se pretende escrever;
2. Indexação (anotação) dos componentes da estrutura esquemática e das estratégias retóricas contidas no discurso científico;
3. Reutilização das expressões-padrão e/ou formulaicas existentes, a fim de se praticar as convenções lingüísticas e de estilo exigidas pelo gênero acadêmico;
4. Colocar os itens acima em uma ferramenta computacional.

Esse tipo de abordagem apresentada possibilita ao escritor em situações de dificuldades de escrita:

1. Obter um insumo lingüístico adequado as suas necessidades, uma vez que será exposto ao léxico e às estruturas textuais pertinentes à área na qual o texto será produzido;
2. Adaptar as expressões que julgar adequadas ao seu texto;

[4] Colocações podem ser definidas como combinações lexicais recorrentes, não idiomáticas e coesas cujos elementos são contextualmente restritos e co-ocorrem arbitrariamente (Tagnin 1998: 41). Ex: ‘stark’ (adv. extremely, totally) se associa a ‘contrast’ e sheer (completely, totally) se associa a ‘scale’, ‘number’ e ‘force’.

3. Reproduzir fórmulas ou expressões fixas ao escrever determinados aspectos do seu trabalho, promovendo: familiarização com a retórica do artigo científico escrito em língua inglesa e diminuição da sobrecarga cognitiva sentida pelo escritor no momento de transferências das idéias para um texto escrito (no caso, um texto em língua estrangeira).

Entre as ferramentas/ambientes existentes que adotam esse tipo de abordagem, descreveremos, a seguir, três em especial: o ambiente AMADEUS (Amiable Article Development for User Support), a ferramenta SciPo (Scientific Portuguese) (Seção 2.4) e a ferramenta SciPo-Farmácia (Seção 2.5), cujas respectivas experiências de utilização comprovaram que a aceitação desse tipo de ferramenta está fortemente ligada ao fato de possuírem um córpus com textos da área de pesquisa do escritor, a partir do qual trechos de textos bem-elaborados podem ser reutilizados na escrita de um texto similar (no caso, artigo científico ou tese).

O AMADEUS (Caldeira *et al*, 1992; De Oliveira *et al*, 1992; Aluisio & Oliveira, 1995; Aluisio & Oliveira, 1996; Aluisio & Gantenbein, 1997a; Aluisio & Gantenbein, 1997b; Aluisio *et al*, 2001) é um ambiente computacional de auxílio e ensino da escrita acadêmica em inglês, voltado para escritores não-nativos, que sofrem influência negativa da língua materna ao escreverem em inglês. É composto por quatro ferramentas inter-relacionadas: Ferramenta de Referência, Ferramenta de Suporte, Ferramenta de Crítica e Ferramenta Tutorial. As três primeiras já foram implementadas e a quarta somente projetada[5]. A tabela 2.1 ilustra a relação entre o tipo de conhecimento da língua inglesa e do gênero científico que o usuário possui e o suporte que as ferramentas desse ambiente podem lhe proporcionar.

[5] 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 Experimental. Já a Ferramenta de Crítica foi desenvolvida durante o trabalho de mestrado de Silva (1999) e trabalha com o domínio específico da comunidade de HCI (Human-Computer Interaction).

	Boa experiência com a escrita acadêmica	**Alguma experiência com a escrita acadêmica**	**Nenhuma experiência com a escrita acadêmica**
Bom Domínio do Inglês	***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***

Tabela 2.1: Adequação do tipo de ferramenta quanto ao conhecimento que o usuário possui (Barros, 2000[6] *apud* Feltrim, 2004).

O AMADEUS foi fundamentado nos estudos reportados em Fontana *et al* (1993), a respeito de alunos brasileiros que realizaram sua pós-graduação no exterior, cujos resultados mostraram que algumas deficiências na escrita dos sujeitos de pesquisa estavam relacionadas ao mau uso ou omissão de expressões mais/menos convencionais que desempenham funções específicas no texto científico. Uma solução para esse problema, e que se constitui na estratégia central desse sistema, é a reutilização de expressões de textos reais, bem estruturadas e categorizadas de acordo com suas metas retóricas e inserções nos diferentes componentes[7] da estrutura esquemática de um tipo de texto para que, no caso de dúvida no momento da produção escrita, o escritor tenha uma base de bons exemplos à qual ele possa recorrer. Além de diminuir a interferência negativa da língua materna na produção textual, a utilização de expressões contextualizadas pode auxiliar também na familiarização com construções sintáticas e semânticas na língua não-nativa, na reutilização de marcadores discursos (Paizan, 2001), verbos e tempos verbais, tornando o novo texto mais adequado em termos de sua organização lexical, sintática, semântica e retórica.

A seguir, cada uma das três ferramentas existentes no ambiente AMADEUS será descrita, com a indicação de seu contexto de uso e de seus respectivos tipos de recursos lingüísticos.

[6] BARROS, R. C. Modelagem de usuários para sistemas de auxílio à escrita técnica. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo – USP, 2000.

[7] Componente da estrutura esquemática: elementos que necessariamente precisam estar contidos em um artigo científico, por exemplo, em um resumo é indispensável a presença do contexto, da lacuna e da conclusão. As estratégias são os elementos que caracterizam os componentes e podem se apresentar de diferentes maneiras, por exemplo: um contexto de um resumo pode ser composto por uma declaração de relevância do tópico, por uma familiarização de termos ou ainda pela introdução da pesquisa que está sendo realizada a partir da grande área na qual se encontra inserida.

2.4.1. Ferramenta de Referência

Essa ferramenta tem como objetivo servir de referência para um escritor não-nativo, pois disponibiliza ao usuário uma base de expressões-padrão categorizadas de acordo com os componentes da estrutura esquemática e as estratégias retóricas que um artigo científico pode possuir. Essas expressões são apresentadas com lacunas de modo que o escritor possa preenchê-las com o material factual de sua pesquisa. Exemplos de expressões com lacunas a serem preenchidas são apresentados a seguir:

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 revealed the irreversible changes that can occur during ...
c) Previous reports on related work. Several papers have reported measurements aimed at obtaining evidence for, and insight into, ... processes in ...
d) What is lacking in the field. Although significant advances have been made in the understanding 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 ... available and the requirement for a ...
f) What the present work does. The purpose of the work reported here was to study the influence of ... on the ...
g) 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.
h) Layout or Outline of the paper The organisation (outline) of the (this) paper is as follows. In Section II we describe ... 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 2.1.: Exemplos de expressões-padrão para a escrita de introduções (Aluísio, 1995)

O acesso a essas expressões pode ser feito de três maneiras: (1) por palavras-chave da

área de pesquisa (terminologia da área) e/ou do gênero de texto (*report, paper, model, section, aim, results, objective, purpose*, etc.); (2) por componentes e sub-componentes de um artigo científico; e/ou (3) por funções retóricas de um texto científico, como comparações, definições e exemplificações, ou seja, estratégias genéricas que aparecem nas várias seções do artigo, mas que durante a busca se restringem à seção em foco.

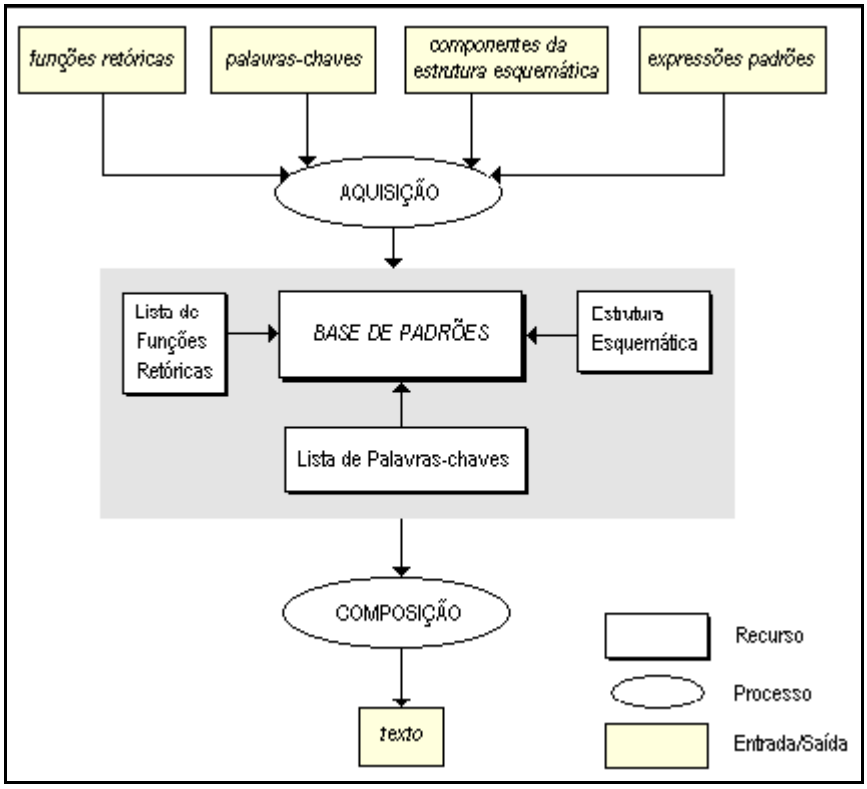


Figura 2.2.: Arquitetura da Ferramenta de Referência (Aluísio, 1995 *apud* Feltrim, 2004)

Conforme pode ser observado na Figura 2.2 há dois processos principais: composição textual e aquisição (inserção de novo material lingüístico à ferramenta, isto é, de novas expressões-padrão, funções retóricas, componentes e subcomponentes e palavras-chave). Esse processo de aquisição permite ao usuário personalizar a ferramenta com o material lingüístico que desejar. Também são apresentados nessa Figura 2.2 os tipos de recursos disponibilizados ao usuário por essa ferramenta: (1) uma base de exemplos de expressões-padrão, chamada de *Base de Padrões*; (2) Lista de componentes e subcomponentes de artigos científicos proposta por Deyes (1982); (3) lista de funções retóricas (definições, exemplificações, etc.) e (4) lista de palavras-chave do gênero (report, paper, model, etc.) que, por meio de acesso à Base de Padrões, retornará todas as sentenças contendo a palavra-chave escolhida.

Testes com essa Ferramenta de Referência foram realizados junto a alunos de pós-graduação e, entre os bons resultados podem ser apontados uma familiarização com a

estrutura retórica de artigos científicos, uma melhoria na organização do texto, bem como um auxílio na superação do bloqueio inicial na escrita de um texto em língua estrangeira. Porém, conforme relata Fontana *et al* (1993), esse desempenho positivo da ferramenta apenas é notado em usuários que já possuíam boa recepção da língua inglesa e alguma experiência com escrita científica, pois somente aqueles que tinham esse tipo de perfil conseguiram reempregar adequadamente as expressões nos contextos corretos. Diante de tal fato, os pesquisadores do AMADEUS notaram a necessidade de elaborar uma outra ferramenta computacional que auxiliasse um usuário menos experiente a adquirir informações relativas aos componentes da estrutura esquemática e estratégias retóricas esperadas para um artigo científico. E é esse o tipo de auxílio proposto pela Ferramenta de Suporte, descrita a seguir.

2.4.2. Ferramenta de Suporte

Com o objetivo de diminuir o problema da falta de coesão e de coerência em textos escritos em inglês por não-nativos, Aluísio (1995) propôs, por meio de uma abordagem baseada em casos, a construção de uma ferramenta de Suporte. A Ferramenta de Suporte trabalha em cooperação com o usuário que, embora tenha um conhecimento razoável da língua inglesa, não possui muita experiência em escrever textos científicos, auxiliando-o, portanto, a garantir coesão e coerência em pequenos trechos de textos (no caso, a escrita de introduções de artigos curtos (*letters*) da área de Física Experimental). Essa ferramenta exige uma análise de córpus bastante detalhada e trabalhosa, pois tem como funcionalidade apresentar várias formas de realizações lingüísticas de componentes de uma dada estrutura esquemática, isto é, para cada estrutura esquemática existem diferentes estratégias retóricas. E, segundo o suporte proposto por essa ferramenta, quando o usuário adota estratégias retóricas adequadas e tais trechos são justapostos de acordo com os componentes de uma estrutura esquemática apropriada, esse usuário acaba conseguindo redigir trechos de texto mais coesos, que podem culminar em um texto coerente.

Para tal, Aluísio (1995) utilizou o *Raciocínio Baseado em Casos* (RBC) (Mantaras e Plaza, 1997[8] *apud* Feltrim, 2004), e desenvolveu uma abordagem composta por dois mecanismos: o primeiro fornece um texto real estruturado para que o escritor veja como os mecanismos de coesão são expressos adequadamente na língua estrangeira e o segundo, por

[8] MANTARAS, R.L. & PLAZA, E. Case-based reasoning: An overview. *AI Communications Journal*, 10(1), p. 21-29, 1997.

sua vez, adapta a estrutura instanciada pelo escritor as suas necessidades. Essa abordagem proposta pode, segundo Feltrim (2004:47) ser resumida em três princípios, a saber:

1. Pré-compilação do conhecimento do gênero em um esquema detalhado que é utilizado 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, isto é, trechos de textos e não de conteúdo científico, ajudando a aumentar a fluência[9] dos textos e servindo de fonte de conhecimento na fase de adaptação;
3. Utilização de raciocínio baseado em casos, RBC, como modelo.

Esse princípio determina as fases de interação e os recursos utilizados que são:

- Uma Base de Casos gerada pela 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 redigido;
- Regras de Revisão utilizadas na adaptação interativa de um caso autêntico para outras necessidades.

A implementação dessa abordagem é feita na Ferramenta de Suporte, que atualmente está construída na plataforma *Windows*. A Figura 2.3 apresenta os processos e recursos contidos nessa ferramenta.

[9] Fluência, neste trabalho, está relacionada a uma melhor estruturação e adequação da informação de um texto.

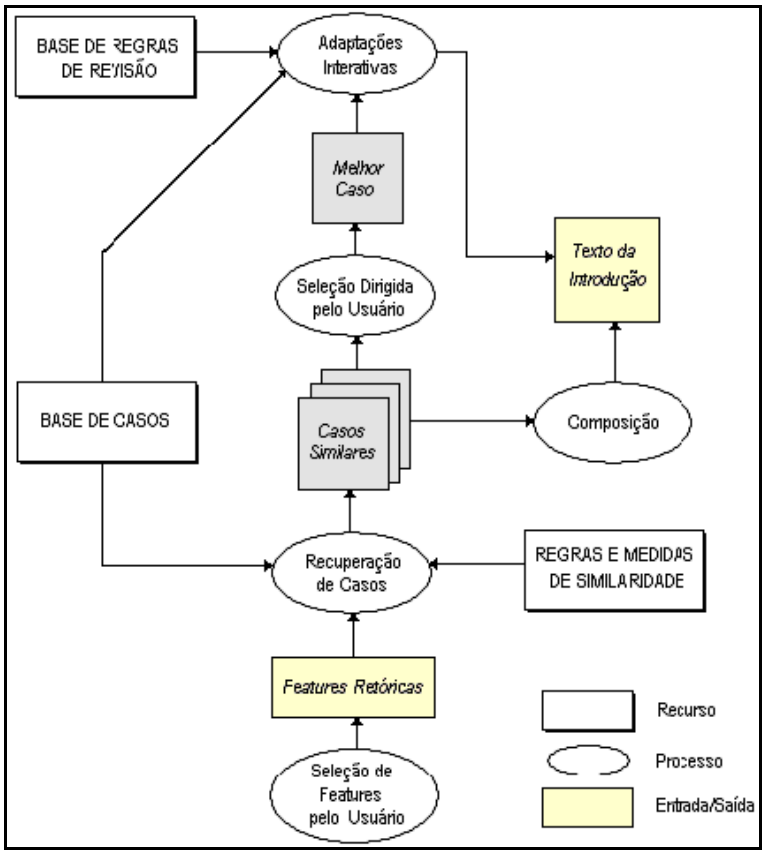


Figura 2.3.: Arquitetura da Ferramenta de Suporte (Aluísio, 1995 *apud* Feltrim, 2004)

Os processos apresentados pela Figura 2.3, nos possibilitam observar o modo como o usuário pode interagir com a Ferramenta de Suporte. Por meio do processo "Seleção de *Features* pelo Usuário", o mesmo elabora sua requisição, que servirá como entrada para o processo de "Recuperação de Casos", que é realizado pela ferramenta. Esse processo recolhe as *Features* retóricas fornecidas pela requisição do usuário, as estruturas retóricas de cada caso da "Base de Casos" e as Regras de Medida de Similaridade, devolvendo ao usuário os casos mais próximos da requisição por ele feita. Diante de todos os casos recuperados, o usuário escolhe o melhor caso pelo processo de "Seleção Dirigida pelo Usuário", e assim pode iniciar as "Adaptações Interativas". Portanto, munido dessas adaptações, que podem ou não ser realizadas, e dos recursos lingüísticos específicos da sua área (expressões-padrão, marcadores discursivos, etc.), o usuário pode dar início à escrita de sua introdução.

Quanto a sua avaliação, a Ferramenta de Suporte foi testada por um número pequeno de usuários reais (8). Já uma limitação que pode ser apontada nas duas ferramentas – Referência e Suporte - é a falta de oferecimento de *feedback* para as escolhas feitas pelos usuários. Essa limitação levou ao desenvolvimento da Ferramenta de Crítica, que será apresentada na seguinte seção.

2.4.3. Ferramenta de Crítica

A ferramenta de Crítica incorporada ao ambiente AMADEUS trabalha em colaboração com o usuário, fornecendo críticas para que a estrutura de seu texto seja adequada a um dado propósito e público-alvo. A abordagem do sistema de críticas incorporado ao ambiente teve como base o modelo proposto por Fischer *et al* (1991[10] *apud* Aluísio, 1995). Segundo esse modelo, um sistema de críticas é composto por dois agentes - um computador e seu usuário - os quais trabalham em colaboração dentro de um processo cíclico. Nessa colaboração, ambos contribuem com seus respectivos conhecimentos para promover a solução de um dado problema. A tarefa básica desse sistema é o reconhecimento e a indicação de deficiências no texto produzido, gerando uma crítica. Com as sugestões dadas pelo sistema, o usuário pode corrigir o problema ou procurar obter explicações adicionais.

Essa ferramenta foi implementada para auxiliar na escrita de introduções de artigos da área de HCI, da qual foram compilados 51 textos da seção "Introdução" de artigos publicados na HCI'96[11]. Esses artigos tiveram suas estratégias retóricas anotadas segundo modelo proposto por Aluísio (1995; Aluísio & Oliveira Jr., 1996), assim como uma estrutura de componentes (estruturas esquemáticas) específicos a cada tipo de artigo (experimental, teórico, reportando um sistema, uma experiência e uma metodologia). A junção dessas duas informações foi chamada de "Estrutura Dual" (Silva, Pelizzoni & Aluísio, 1998; Silva, 1999). A Figura 2.4 ilustra o modo de funcionamento dessa Ferramenta de Crítica.

[10] FISCHER, G. et all. Critics: an emerging approach to knowledge-based human-computer interaction. In *J. Man-Machine Studies*, 35, p. 695-721, 1991.
[11] Conference on Human Factors in Computing Systems realizada em 1996.

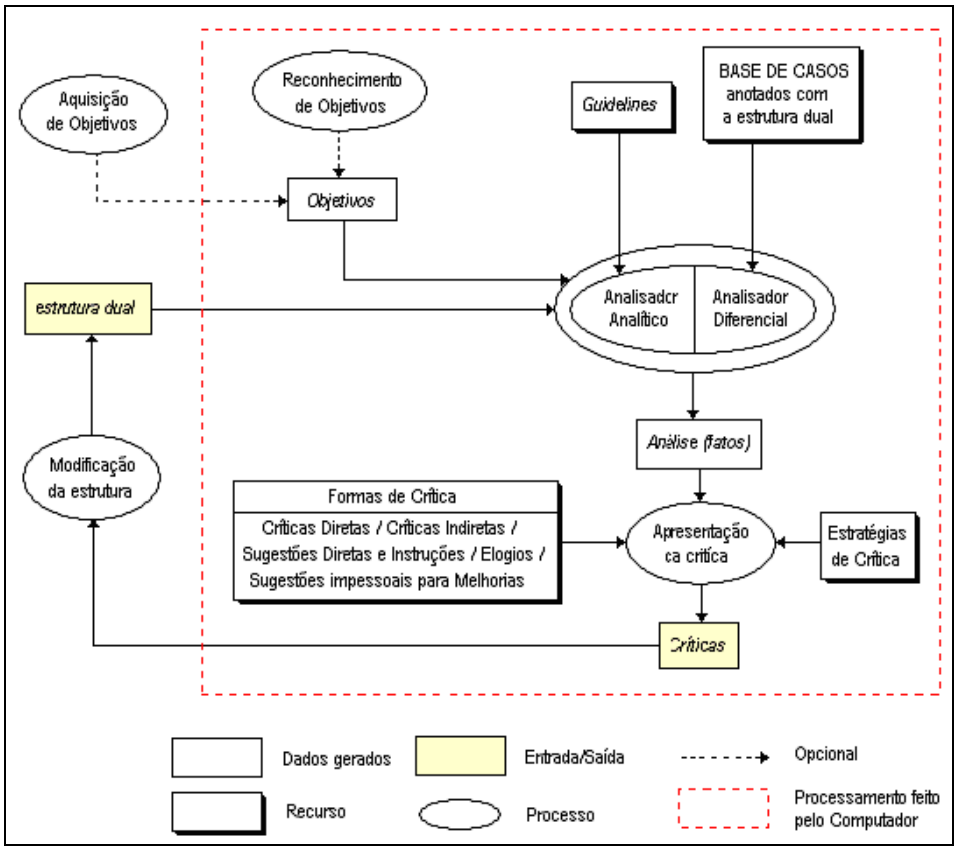


Figura 2.4.: Arquitetura da Ferramenta de Crítica (Silva, 1999[12] *apud* Feltrim, 2004).

Conforme a Figura 2.4, o processo de crítica tem seu início no momento em que o usuário apresenta um texto para essa ferramenta, que para criticá-lo, precisa obter o(s) objetivo(s) do usuário. Esse(s) objetivo(s) pode(m) ser obtido(s) pelo reconhecimento das escolhas utilizadas no texto (*Reconhecimento de Objetivos*) ou por informações explícitas fornecidas pelo usuário (*Aquisição de Objetivos)*. Depois de definir os objetivos, são utilizados dois tipos de processos para se avaliar o texto submetido, o *Analítico* e o *Diferencial*. O primeiro analisador checa o conteúdo do texto (quais componentes da estrutura esquemática devem estar presentes na Introdução), ao passo que o segundo analisa a organização textual (a ordem mais provável que essas estruturas poderem aparecer no texto, quais podem ser opcionais ou ainda quais podem aparecer mais de uma vez).

Como saída desses dois analisadores, vemos os *fatos* sobre as diferenças encontradas entre o texto do usuário, o caso recuperado da Base de Casos e a análise do texto em relação as *guidelines* (as regras heurísticas utilizadas pelo *Analisador Analítico*, as quais ajudam o

[12] SILVA, M.H.B. *A Abordagem de Críticas para a Construção de Sistemas de Aprendizado da Escrita Técnica. Dissertação*. 1999. 130f. Dissertação (Mestrado em Ciências da Computação), Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo . ICMC-USP, 1999.

texto do escritor a entrar em conformidade com as expectativas da comunidade discursiva[13] da qual faz parte). Esses *fatos* são reportados ao usuário na forma de elogios, críticas diretas/indiretas, sugestões e instruções. De posse dessas informações, o usuário pode iniciar uma nova versão de seu texto e reiniciar o processo de críticas.

Para avaliar a Ferramenta de Crítica foram realizadas simulações com três classes de usuários: principiante, intermediário e especialista. Como resultado observou-se que os usuários dessas três classes se beneficiaram da ferramenta, pois através da interação com as críticas fornecidas pelo sistema, produziram estruturas melhores do que as produzidas inicialmente. Também foi concluído nessas avaliações, que aspectos como (1) extensibilidade das *guidelines* de Casos e Estruturas de novas seções ou de novos tipos de artigos, (2) personalização e (3) portabilidade para uma nova comunidade de pesquisa são possíveis de serem implementados, porém a um alto custo. Maiores detalhes de avaliação podem ser encontrados em Silva (1999).

2.4.4. Ferramenta Tutorial

A Ferramenta Tutorial é um módulo idealizado para o ambiente AMADEUS, mas que ainda se encontra em fase de projeto. Ela difere das outras ferramentas apresentadas por focar no auxílio a usuários completamente inexperientes na escrita científica em língua inglesa. Segundo Feltrim (2004), sua interação será baseada no esquema tutor/aprendiz, pois esse tipo de sistema possui mais conhecimento que o seu usuário.

A partir de experiências bem-sucedidas do ambiente AMADEUS, novos sistemas foram desenvolvidos no NILC ao longo dos últimos anos, com introdução de inovações, como um sistema de auxílio à escrita em português de teses e dissertações na área de computação. Esse último sistema, o *Scientific Portuguese* - SciPo[14] (Feltrim *et al*, 2003, Feltrim, 2004) foi construído nos moldes da ferramenta de Crítica do AMADEUS. Também foi desenvolvida um ferramenta de auxílio à escrita de artigos em inglês no domínio das Ciências Farmacêuticas, o SciPo-Farmácia, a qual é composta pela mesma interface (*look-and-feel*) apresentada no SciPo e pelas funcionalidades contidas nas ferramentas de Suporte e de Referência do AMADEUS. Mais detalhes sobre essas ferramentas, a seguir.

[13] Adotamos neste trabalho a noção de comunidade discursiva de Swales (1990:21). Mais detalhes ver seção 3.4.3, do Capítulo 3.
[14] Endereço eletrônico do SciPo: http://www.nilc.icmc.usp.br/~scipo/

2.5 SciPo – Scientific Portuguese

O sistema SciPo é um conjunto integrado de recursos e ferramentas, cujo objetivo é oferecer suporte na estruturação e redação de Resumos e Introduções em português, em especial de teses e dissertações do domínio da Ciência da Computação. Esse tipo de auxílio fornece uma lista de componentes de estrutura esquemática e de suas respectivas estratégias retóricas para a construção da estrutura textual de um resumo ou de uma introdução que será criticada e, posteriormente, utilizada para a recuperação de casos que sejam autênticos e similares à estrutura escolhida pelo usuário.

Implementado como um ambiente Web, o SciPo contempla tanto uma composição *top-down* de um texto (partindo do planejamento estrutural para a escrita propriamente dita), quanto uma composição *bottom-up* (partindo de um rascunho já escrito). Essas formas diferentes de apoio à estruturação são acessadas no SciPo pela "Seleção da Estrutura" e "Crítica Automática", respectivamente. Apesar de possuírem o mesmo fim, a análise e crítica da estrutura do texto apresentam formas de interação distintas.

Na "Seleção da Estrutura", o usuário inicia seu texto por meio da escolha dos componentes e estratégias que irão compor a estrutura de seu texto. Esses elementos são apresentados pelo próprio sistema e deverão refletir o tipo de informação que o usuário pretende incluir em seu texto. A Figura 2.5 ilustra essa etapa descrita.

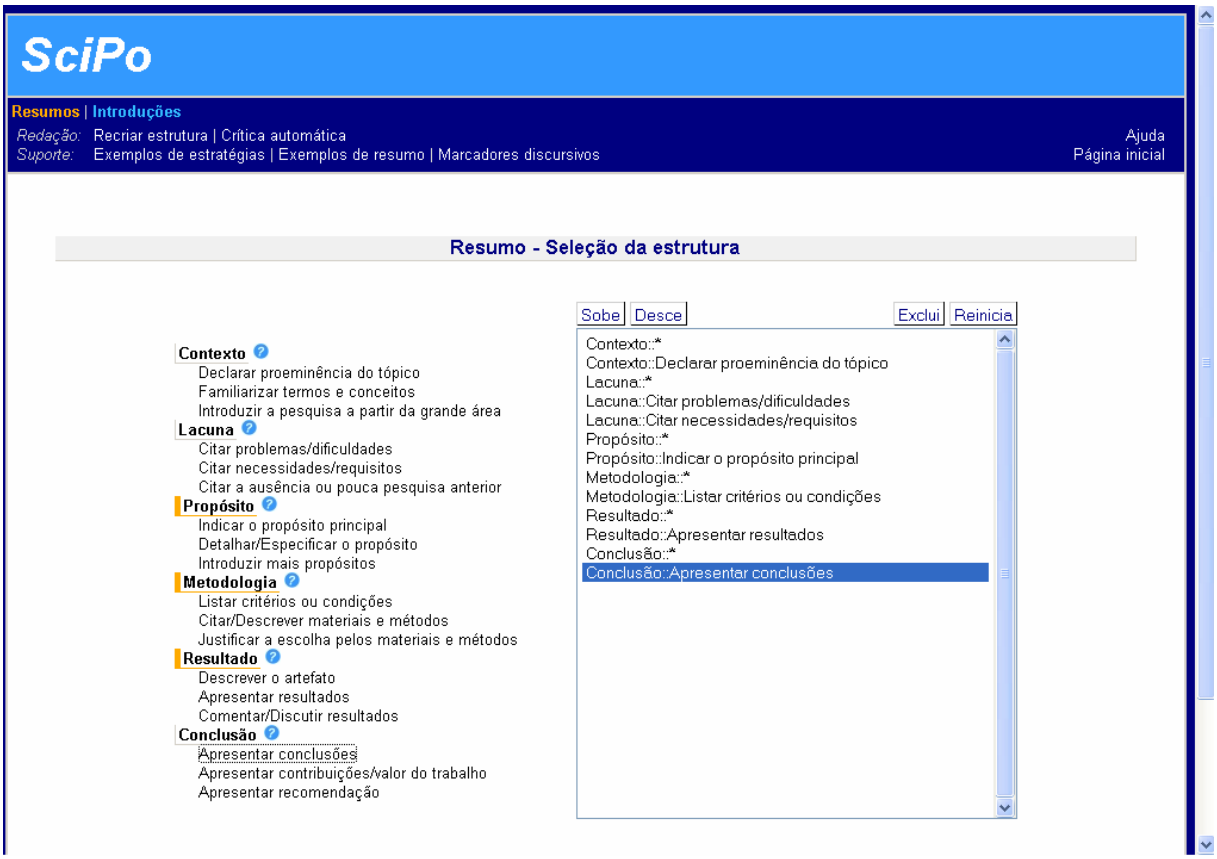


Figura 2.5.: Tela do ambiente SciPo com seleção da estrutura de um texto, no caso, a estrutura de um resumo.

Na “Crítica Automática”, o usuário submete ao sistema um texto pronto (escrito). A seguir, um classificador automático irá atribuir a cada sentença do texto submetido um rótulo correspondente a sua categoria (no caso de um resumo, por exemplo, poderão ser adicionados os rótulos de Contexto, Lacuna, Propósito, Metodologia, Resultado e Conclusão). A partir de então, o texto passa a ser representado no sistema por sua estrutura. Vale dizer que esse classificador automático possui uma precisão média de 70% e que, portanto, é possível haver uma classificação equivocada de algumas sentenças, fato que pode ser corrigido pelo usuário se necessário.

Dependendo da estrutura do texto, o sistema de críticas automáticas poderá emitir críticas e/ou sugestões. As críticas são as mudanças necessárias que o usuário precisará fazer em seu texto de modo que a estrutura do mesmo esteja em acordo com a estrutura mínima aceita. Já as sugestões são dicas que visam uma melhoria dos componentes de uma estrutura esquemática, as quais o usuário pode ou não aceitar.

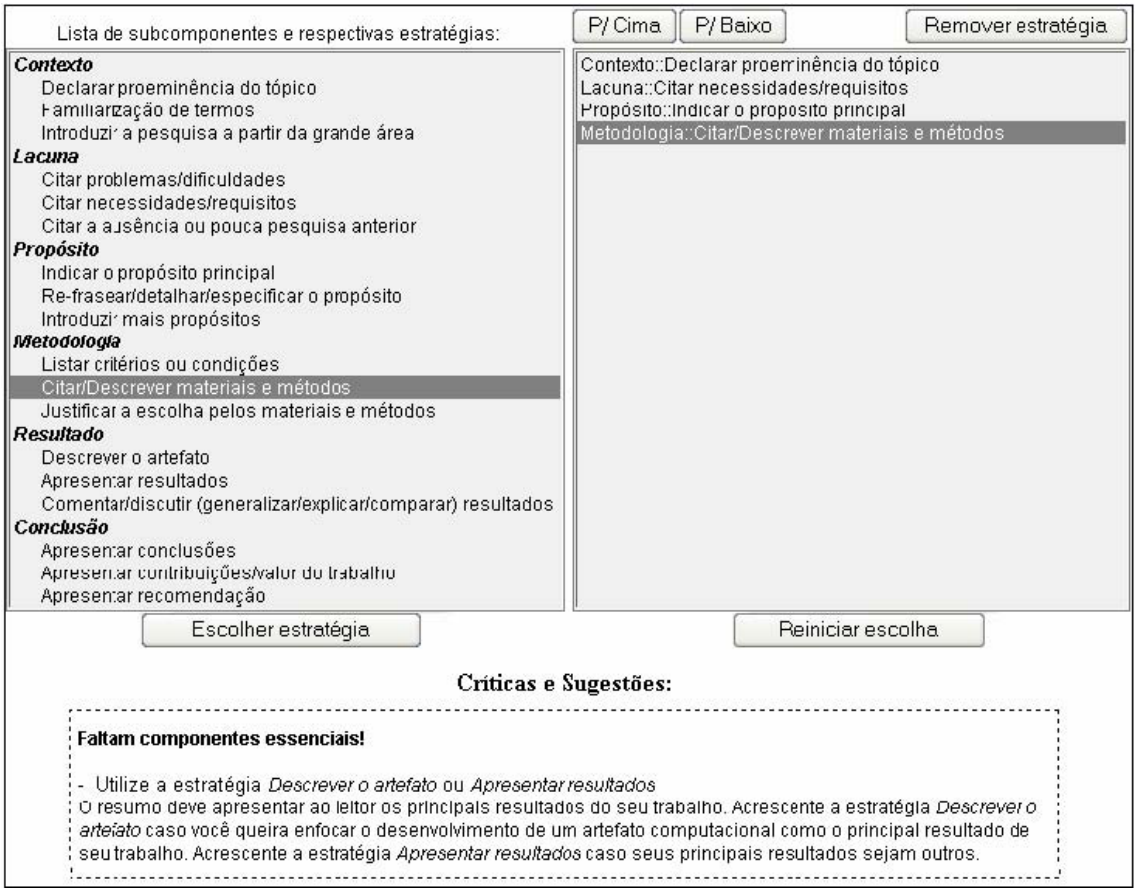


Figura 2.6: Tela de crítica, Feltrim *et al* (2003).

Terminada a revisão da estrutura, o usuário pode iniciar a redação de seu texto. Para iniciar esse processo basta que o usuário clique na opção "Iniciar Redação", que a página de edição irá apresentar a estrutura anteriormente selecionada no formato de um formulário a ser preenchido com o seu texto. Se o usuário acessou o SciPo pelo sistema de "Crítica Automática", o formulário será preenchido com o texto original submetido à análise.

Além desse apoio na estruturação do texto, o sistema SciPo também possui outros recursos que podem auxiliar na escrita acadêmica. O primeiro deles é uma base de exemplos autênticos de teses e dissertações em Ciência da Computação (anotados e comentados), que podem ser acessados em qualquer momento da interação com o ambiente. Além dessa anotação dos componentes da estrutura esquemática e das estratégias retóricas, existem também exemplos de expressões-padrão e de marcadores discursivos (organizados por função que desempenham no texto), isto é, partes do texto que podem ser reutilizadas na escrita de diferentes textos.

Em poucas palavras, pode-se dizer que marcador discursivo são palavras que sinalizam relações entre as idéias do texto, por exemplo, a palavra "portanto" é um marcador discursivo que sinaliza conclusão, a palavra "também" sinaliza, por sua vez, adição de idéias. Para se ter

acesso aos marcadores discursivos do SciPo, basta selecionar no "Menu Principal" o ícone "Marcadores Discursivos".

O sistema também disponibiliza acesso a um revisor ortográfico e gramatical durante a fase de edição do texto, que é acionado pelo ícone "Revisar Texto". Durante a fase de edição dos textos pode ser feita uma reutilização das expressões-padrão contidas na base de exemplos. Quando um exemplo de resumo, por exemplo, é visualizado durante essa fase de edição, as expressões-padrão nele contidas aparecem finalizadas com o símbolo [>], que se trata de um *link* para a transferência automática da expressão-padrão correspondente para a janela de edição. Há também no formulário de edição um contador de palavras que auxilia no controle do balanceamento do texto, isto é, no controle do equilíbrio entre a quantidade de texto redigido em cada entrada do formulário; e a geração de documento .RTF a partir do texto redigido no formulário, possibilitando que o texto gerado seja impresso, por exemplo, ou submetido a um outro tipo de editor de textos para a verificação da ortografia e da gramática, se desejável.

Uma descrição mais detalhada das funcionalidades contidas neste ambiente podem ser acessadas no *link* "Ajuda", que fornecerá o documento "Descrição da Interface do Sistema SciPo". A seguir será apresentada a arquitetura do ambiente SciPo, com os processos e recursos nele contidos.

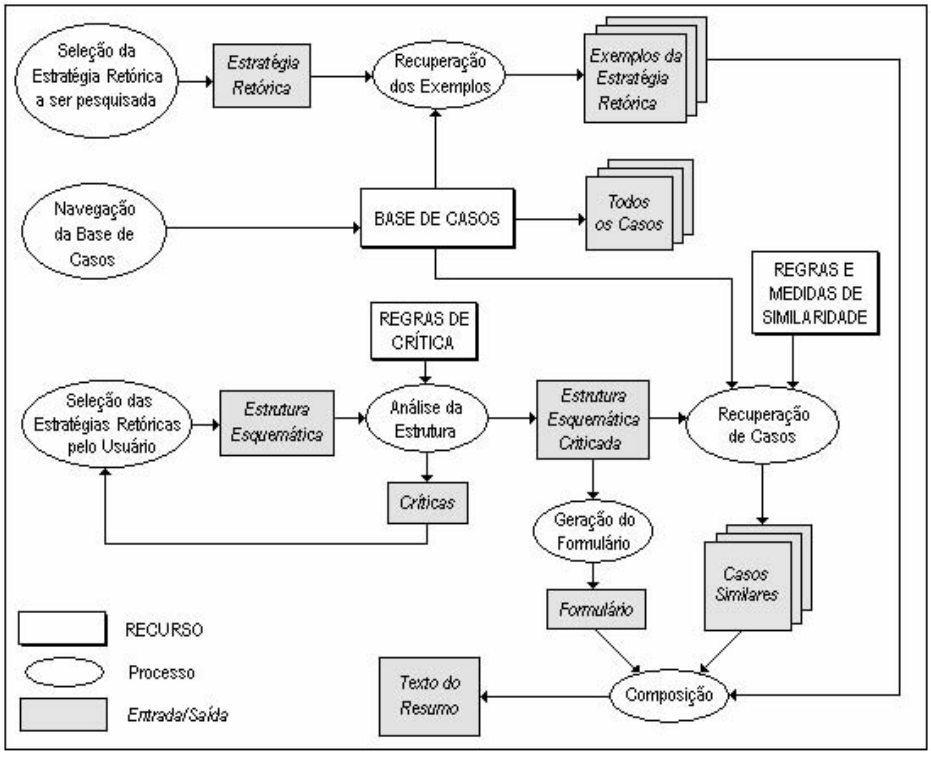


Figura 2.7.: Arquitetura do sistema SciPo (Feltrim *et al*, 2003)

Como podem ser observados na Figura 2.7, os processos da arquitetura representam as formas com as quais o usuário pode interagir com o sistema. Uma opção possível é por meio da navegação pela "Base de Casos", visualização de todos os textos anotados, seguida de uma seleção de uma dada estratégia retórica, a fim de que o sistema recupere todas as ocorrências dessa estratégia na "Base de Casos".

O usuário poderá também montar os componentes de uma estrutura esquemática. As mesmas serão submetidas ao processo de "Análise da Estrutura", que com base nas "Regras de Crítica", o sistema analisará a estrutura e retornará ao usuário as críticas e sugestões necessárias.

A seguir, o usuário acatará as críticas e poderá aceitar ou não as sugestões, modificará a estrutura de seu texto e o submeterá novamente ao sistema. Esse ciclo de revisão só terminará quando a estrutura que organizará o texto a ser redigido estiver satisfatória ao ambiente, isto é, até que nenhuma crítica relativa ao conteúdo ou ordem das estruturas seja acionada. No processo de "Recuperação de Casos" há o retorno ao usuário dos casos estruturais mais próximos da requisição por ele realizada. Para isso, o sistema considera a requisição feita pelo usuário, posteriormente analisada pelo sistema, as quais resultaram numa seleção de estratégias retóricas que serão comparadas com as estratégias retóricas de cada caso contido na "Base de Casos" por meio de Regras e Medidas de Similaridade.

A partir dos componentes da estrutura esquemática construída pelo usuário também é gerado um formulário para a inserção de texto na estrutura gerada, no qual o escritor poderá consultar novamente os casos estruturais similares retornados ou então pesquisar mais exemplos de estratégias retóricas para compor seu texto.

Para avaliar o ambiente SciPo, foram realizados dois experimentos com usuários reais (estudantes da graduação e pós-graduação do curso de Ciências da Computação do ICMC, USP- São Carlos), a fim de verificar as duas abordagens de auxílio utilizadas: o processo *top-down*, que parte do planejamento estrutural para a escrita propriamente dita e o processo *bottom-up*, em que se submete um texto já escrito à análise automática da estrutura. A escrita de Resumos foi o foco dos dois experimentos. De modo geral, conforme aponta Feltrim (2004: 122-125), o SciPo se mostrou eficiente, no sentido de que conseguiu guiar o escritor na composição de seus resumos informativos, cujos componentes retóricos seguem padrões ditados pelo gênero acadêmico. No entanto, a autora observa que não há garantia de uma boa produção textual quando os usuários (estudantes) têm pouco ou nenhum conhecimento sobre o gênero acadêmico.

2.6 SciPo-Farmácia

O SciPo-Farmácia é um conjunto de ferramentas computacionais construídas para auxiliar na escrita de artigos científicos em inglês na área de Ciências Farmacêuticas. Foi construída como uma aplicação Web, com a mesma interface (*look and feel*) do SciPo, porém com funcionalidades mais simples do que as apresentadas por esse outro conjunto de ferramentas e recursos. A Figura 2.8 mostra a interface do SciPo-Farmácia durante a etapa de visualização de exemplos de estratégias retóricas que possuem a função de apresentarem o propósito do estudo com a metodologia.

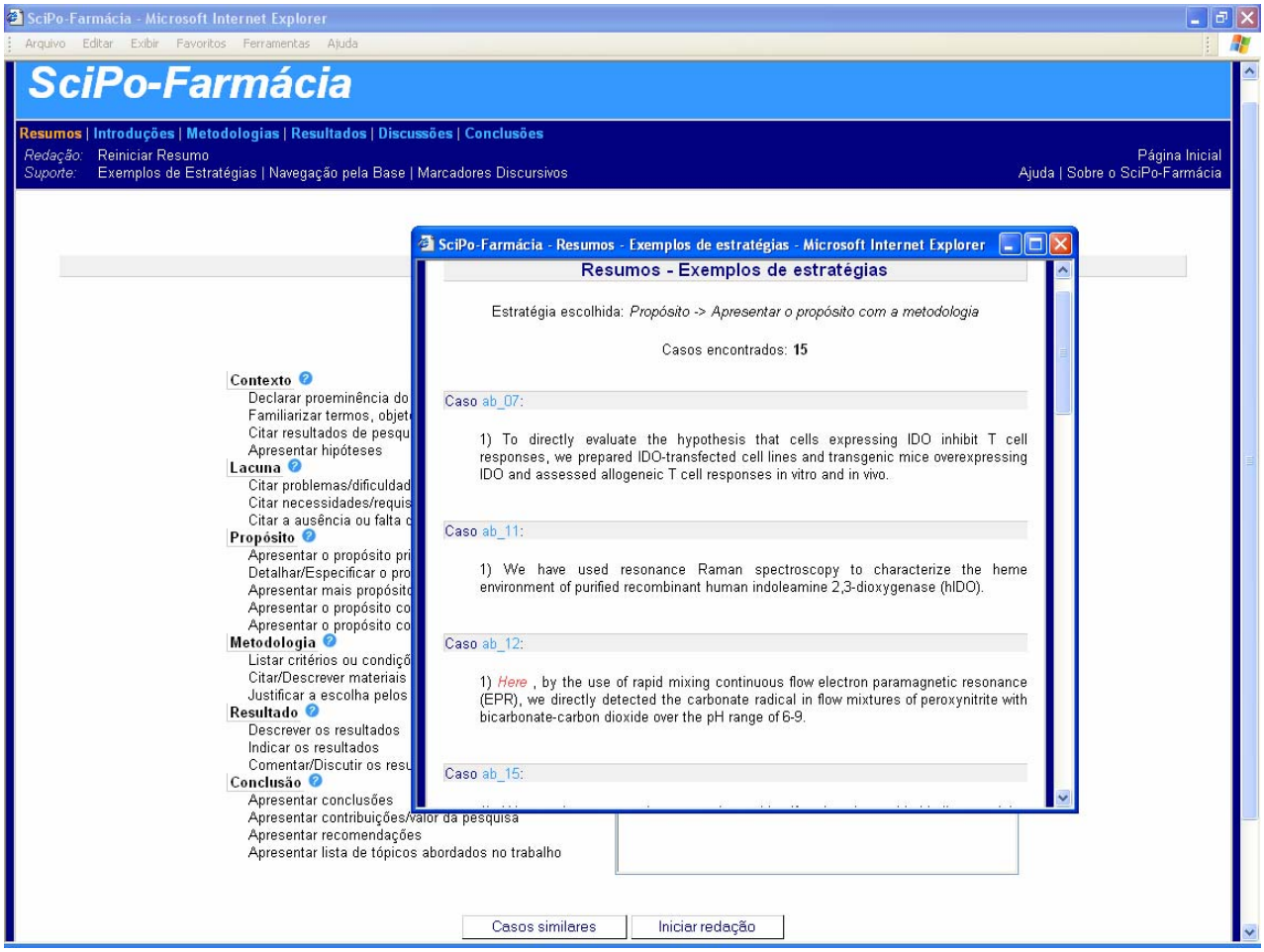


Figura 2.8.: Visualização de exemplos da estratégia retórica que apresenta a descrição do propósito e da metodologia de trabalho juntos.

Apenas as ferramentas de Referência e de Suporte originalmente desenvolvidas no ambiente AMADEUS - mas que também foram adaptadas para o SciPo foram novamente escolhidas para fazerem parte do SciPo-Farmácia. Isso porque: (1) são menos exigentes em

termos de recursos lingüísticos necessários à sua implementação e (2) possuem abordagem implementacional que pode ser extensível para outras áreas.

Vale dizer que o SciPo-Farmácia é voltado para o auxílio na escrita de todas as seções que compõem um artigo científico (Resumo, Introdução, Metodologia, Resultado, e Conclusão), tendo como língua-alvo o inglês. Para tal suporte, o SciPo-Farmácia utiliza um córpus de artigos científicos em inglês publicados na área de Ciências Farmacêuticas. Esse córpus foi analisado e anotado por especialistas tanto na escrita acadêmica quanto na área de Farmácia. Cada texto foi rotulado de acordo com os componentes da estrutura esquemática e estratégias retóricas apresentadas pelos modelos de Weissberg & Buker (1990) e de Swales (1990), perfazendo um total de 43 exemplos de resumos, 39 de Introduções, 26 de Resultados, 11 de Discussões e 22 de Conclusões. Como pode ser observado, não houve a construção de uma base com exemplos de Metodologias, por uma questão de tempo suficiente para a conclusão do projeto. Assim, um dos objetivos pontuais de nossa pesquisa foi a implementação da seção "Metodologia", única seção ainda não confeccionada no SciPo-Farmácia, quando da sua disponibilização pública em 2004.

Com o SciPo-Farmácia o usuário pode: (a) navegar e buscar em sua base de casos textos autênticos com todas as ocorrências de dada(s) estratégia(s) retórica(s) e/ou componentes da estrutura esquemática; (b) pode receber suporte para criar um *outline* (esboço, croqui) como ponto de partida na escrita de um novo artigo; e (c) pode buscar na base de casos textos autênticos, cuja estrutura textual seja semelhante à elaborada no *outline*. A Figura 2.9 mostra esses tipos de funcionalidades citadas, bem como os recursos e processos envolvidos em cada uma delas.

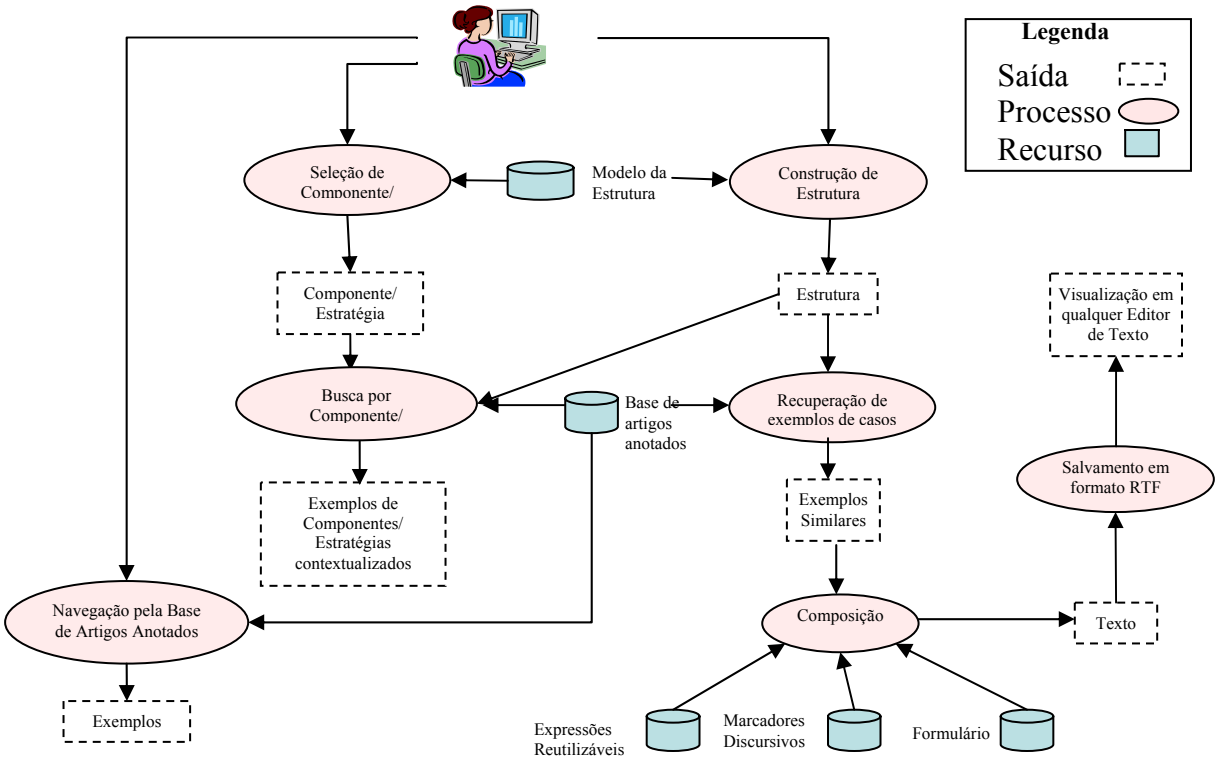


Figura 2.9: Arquitetura do SciPo-Farmácia, adaptada de Aluísio *et al* (2005)

Como pode ser observado na Figura 2.9, o SciPo-Farmácia permite ao usuário três tipos de interações distintas. Na primeira delas, identificada pelo primeiro fluxo da esquerda, o usuário pode simplesmente navegar pela Base de Casos, composta por artigos anotados quanto a sua estruturação de componentes e estratégias retóricas, para obter exemplos de textos com esse tipo de informação identificada. Após essa seleção, são retornados exemplos dos componentes ou estratégias escolhidos. Vale dizer que são retornados diferentes exemplos de componentes ou estratégias fora de seu contexto de uso, isto é, são apresentados em forma de lista e fora da seção do artigo a que pertencem, como pode ser visto na Figura 2.10. Consistiria, então, em uma navegação com o objetivo de conhecer o ambiente sem compromisso com uma escrita imediata, ou seja, visa apenas conhecer as estruturas que compõem os artigos da área de Ciências Farmacêuticas, bem como o modo de funcionamento e os recursos disponibilizados pelo ambiente.

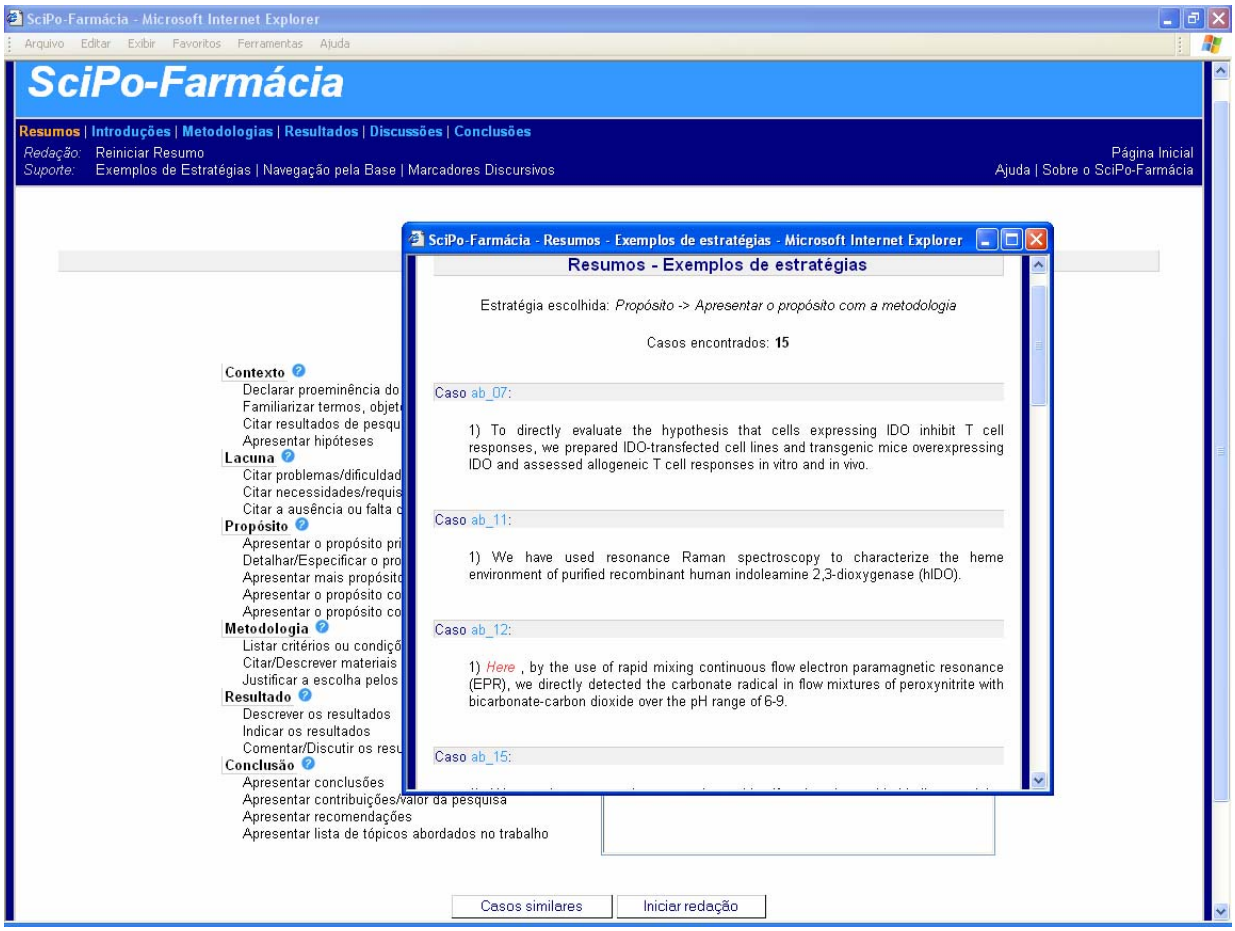


Figura 2.10: Visualização de exemplos da estratégia "Apresentar propósito com a metodologia". Note que os casos ab_07, ab_11 e assim por diante, correspondem aos resumos (contexto) de onde foram retirados os respectivos exemplos.

Em um segundo tipo de interação possível, representado pelo segundo fluxo da esquerda para direita, o usuário escolhe uma estrutura com componentes esquemáticos e estratégias retóricas para compor uma das seções de seu artigo científico. A seguir, ele pede para o ambiente retornar exemplos contidos na base de textos autênticos que possuam o mesmo tipo de estrutura (componentes da estrutura esquemática e estratégias retóricas) que ele selecionou a priori, o qual retorna ao usuário exemplos de estruturas iguais ou o mais semelhantes possível, conforme ilustrado na Figura 2.11. Notamos que nesse tipo de interação, os exemplos já são retornados com o seu contexto de uso, ou seja, aparecem juntos do texto do qual fazem parte.

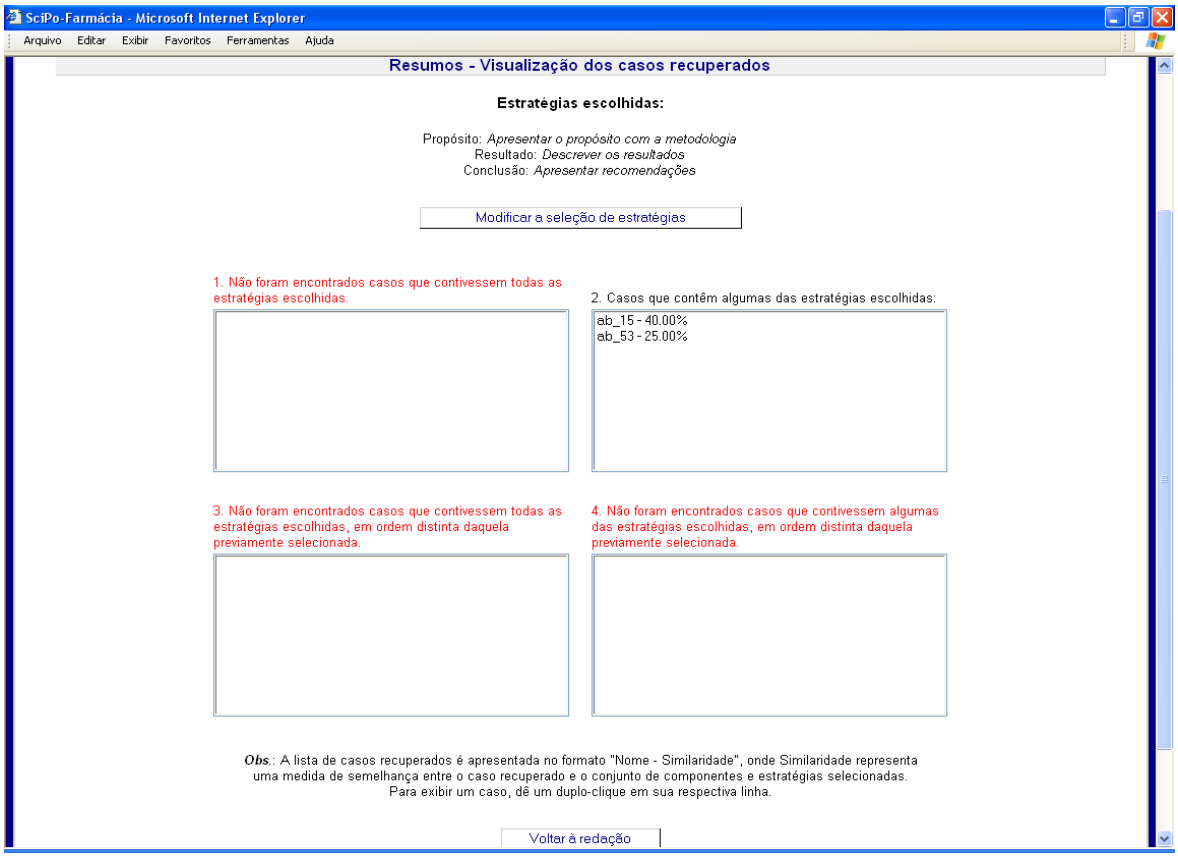


Figura 2.11: Tela com a recuperação de Casos similares à requisição feita. No primeiro quadro da esquerda não foram encontrados casos que contenham toda a requisição feita, levando-se em conta a ordem em que aparecem organizados na requisição. Já no segundo quadro superior, vemos que foram retornados dois casos que contém parte da requisição feita, levando-se também em consideração a ordem. Já nos quadros inferiores, nota-se que não foram encontrados casos que contivessem todas as requisições feitas ou partes delas, mas que estivessem organizadas de forma distinta da previamente selecionada.

No terceiro e último tipo de interação presente na arquitetura do SciPo-Farmácia, o objetivo consiste na escrita propriamente dita de um artigo. Para tanto, o usuário constrói a estrutura textual de uma seção de seu artigo científico (componentes de estrutura esquemática e estratégias retóricas), e com um clique na tela pede para que o ambiente retorne exemplos similares a essa estrutura. Se desejar, esse usuário pode voltar na etapa de elaboração das estruturas e alterar seu modelo escolhido. O próximo passo consiste na composição da seção de artigo em si, etapa que consiste no preenchimento, em um formulário de composição, da estrutura escolhida com trechos de texto referentes à pesquisa do usuário. Nessa mesma etapa de composição, o usuário poderá ter acesso aos marcadores discursivos (organizados em listas e por função que desempenham no texto) e às expressões reutilizáveis, como as expressões-padrão, para poder preencher seu texto. Vale dizer, que esses dois recursos há pouco citados, estão disponibilizados para uso desde o primeiro momento de interação com a ferramenta, sendo, portanto, acessíveis em qualquer etapa/momento de interação com o ambiente. Por

fim, depois de terminada a escrita do texto, o mesmo é salvo em formato RTF, o qual possibilita que o texto seja aberto e editado em qualquer tipo de editor de texto.

Para avaliar o SciPo-Farmácia, foram realizados alguns experimentos e um deles, em específico, com 7 alunos de um curso de pós-graduação em Ciências Farmacêuticas inscritos em um curso de escrita científica. Esse experimento foi dividido em duas etapas: na primeira, os alunos tiveram que escrever um *abstract* sobre suas respectivas pesquisas. O *abstract* deveria ter um tamanho limite de duzentas palavras, e deveria ser produzido sem qualquer tipo de auxílio (dicionário, gramática ou dos professores do curso), com um intervalo de tempo de uma hora para finalizá-lo. Os mesmos alunos também receberam um questionário sobre o conhecimento de inglês (vocabulário e gramática) e de escrita científica que possuíam. Depois dessa primeira tarefa, os alunos foram introduzidos ao SciPo-Farmácia e treinados durante duas semanas a utilizá-lo de maneira adequada, a fim de que ficasse claro para eles os componentes retóricos e a ordem/lógica que obedecem esses componentes dentro de um resumo. Vale salientar que de acordo com Swales (1990) e Weissberg & Buker (1990), os principais componentes retóricos de um resumo são Contexto, Propósito, Metodologia, Resultados Principais e Conclusão. Segundo Schuster *et al* (2005), os principais componentes retóricos contidos em resumo na área de Ciências Farmacêuticas são Propósito, Resultados Principais e Conclusão.

Na segunda tarefa, realizada logo após as duas semanas de ambientação com o SciPo-Farmácia, os estudantes tiveram que escrever um outro *abstract*, com a ajuda apenas do SciPo-Farmácia. A seguir, esses *abstracts* produzidos foram avaliados por três especialistas, obedecendo a uma dada ordem. Primeiramente, um especialista em escrita científica focou na identificação dos componentes das estruturas esquemáticas e estratégias retóricas contidos, na forma de organização em que apareciam no texto e no balanceamento desses elementos. Posteriormente, um lingüista computacional focou na análise de erros gramaticais e de uso da língua inglesa. Por fim, um especialista do domínio das Ciências Farmacêuticas avaliou o conteúdo científico dos textos, bem como a adequação no uso de termos técnicos dessa área.

Ao final desse e dos outros experimentos realizados com o SciPo-Farmácia, os especialistas que com ele trabalharam puderam observar que os estudantes que escreveram seus *abstracts* com o auxílio dessa ferramenta computacional obtiveram progresso considerável quanto à utilização apropriada dos elementos responsáveis pela boa organização retórica dos resumos, conforme são apontados por Schuster *et al* (2005). Tais autores ainda dividem essa contribuição em dois pontos: 1) relativo ao fato dos estudantes terem aprendido a utilizar adequadamente os componentes retóricos de um resumo, bem como a organizá-los

textualmente em uma ordem lógica o mais adequada possível a um resumo e 2) o fato do nível de autoconfiança desses alunos também ter aumentado com o uso do SciPo-Farmácia. O mesmo pôde auxiliar nesse sentido, uma vez que ofereceu a esses usuários um conjunto de resumos adequados retórica e estruturalmente quanto às convenções existentes sobre escrita científica na área de Farmácia, e apresentados de forma que essas convenções pudessem ser identificadas de maneira rápida no texto, como por exemplo, a utilização de marcadores discursivos em uma dada sentença. No SciPo-Farmácia, assim como no SciPo, os marcadores discursivos além de aparecerem em forma de lista com as respectivas funções que podem exercer dentro de uma dada sentença, eles também aparecem destacados, em vermelho, nas seções de artigos científicos da área de Farmácia, que constituem a base de casos dessa ferramenta computacional. O tipo de interação que o SciPo-Farmácia promove com os outros elementos lingüísticos, como os componentes retóricos (estrutura esquemática e estratégias retóricas que serão detalhadas no Capítulo 3), o vocabulário da área de Ciências Farmacêuticas, as construções e expressões lingüísticas adequadas e pertinentes por exemplo, também contribuem para um maior conforto desses usuários que têm contato com o vocabulário da língua estrangeira, bem como com as construções e expressões pertinentes à área do conhecimento da qual participa. Esse conforto se reflete na própria produção escrita por meio de uma melhora na organização, estrutura e até mesmo conteúdo do texto, pois "desfocam a atenção que antes era despendida à estrutura fixa da língua inglesa e agora passa para o conteúdo científico" (Schuster *et al,* 2005).[15]

A seguir, são apresentadas duas versões de resumos escritas por um mesmo autor da experiência relatada acima com o SciPo-Farmácia. O primeiro texto foi escrito na primeira fase da experiência e o segundo, por sua vez, na segunda fase, na qual o aluno pôde utilizar apenas o SciPo-Farmácia como auxílio. Em negrito estão os componentes de estrutura esquemática contida em cada resumo. Esses componentes esquemáticos são responsáveis por indicar a função retórica de cada sentença do resumo. No caso de outras seções de um artigo científico, os componentes de estruturas esquemáticas são outros, uma vez que as seções possuem diferentes funções; logo, possuirão diferentes estruturas. Essas estruturas estão destacadas nas duas versões dos resumos para ficar mais claro como se deu a variação da estruturação retórica do texto nas duas fases.

[15] "(…) their level of confidence rose when they used SciPo-Farmácia, which enabled them to focus on the content of their abstracts and not remain, fixed on the use of the English language" (Schuster *et al,* 2005).

CONTEXTO - Aqueous two-phase systems (ATPS) are widely used to extract biomolecules, such as enzymes, antibodies, amino acids and other molecules. ATPS is formed by two components (two polymers or one polymer and salts), in this case the polymer is polyethylene glycol (PEG) and the salt is citrate. This extraction system is able to purify biomelecules into one phase, usually this phase is formed by PEG, but many variables (molar mass of PEG, concentration of PEG, concentration of citrate and pH) may be studied enough. There are many applications to ATPS, such as biotechnological and pharmaceutical industries. Protease is the most important group of enzymes, that represent 60% of enzymes world sale.
METODOLOGIA - The protease used in this work was produced by Clostridium perfringens. An experimental design (24) was used to evaluate the variables influences. Statistical design of experiments is an important tool used widely to evaluate the significant effects of variables in ATPS. The best results of protease extraction and purification was obtained with molar mass of PEG 10,000 (g/mol) and citrate concentration 8% (w/w).
RESULTADOS - Under their conditions, the protease yield was above 100% and purification factor was 3.32.
CONCLUSÃO - Therefore, the ATPS was suitable for extract protease from C. perfringens.

Figura 2.12 – Resumo escrito por um aluno da pós-graduação em Ciências Farmacêuticas durante a primeira fase de uma experiência de utilização do SciPo-Farmácia na escrita de um *abstract*. O aluno compôs o texto antes de conhecer esse ambiente computacional sem o auxílio de dicionários, gramáticas e professores, com o limite de 1 hora para seu término.

Nessa primeira versão do resumo nota-se que os componentes esquemáticos Contexto e Metodologia foram desenvolvidos adequadamente, isto é, a função do texto contido em cada um deles corresponde ao tipo de função retórica que as sentenças correspondentes aos mesmos devem exercer: a primeira estrutura de contextualizar a pesquisa e a segunda, por sua vez, de relatar os métodos, processos e materiais utilizados. No entanto, nota-se que o Propósito do estudo está ausente na estrutura, apesar de ser um elemento bastante essencial a um resumo da área de Ciências Farmacêuticas, conforme citado anteriormente. Além disso, o balanceamento entre os componentes da estrutura esquemática utilizada não está adequado, visto que grande parte das palavras do resumo aparece distribuída no Contexto e na Metodologia, enquanto que os Resultados e a Conclusão são constituídos apenas por uma sentença.

CONTEXTO - Experiments in many laboratories have been limited by the availability of the enzyme, and because the enzyme purification is very difficult.
LACUNA - Although experimental design for optimization is a strategy to overcome the purification process more simply, statistical analysis exists to facilitate this process.
PROPÓSITO - This goal of work was the purification of protease from Clostridium perfringens fermentation broth by aqueous two-phase system (PEG/citrate) using experimental design.
METODOLOGIA - An statistical design of experiments was used to evaluate the effects of variables (molar mass of PEG, PEG concentration, pH and citrate concentration). The factorial design was 2*4 with 4 central points.
RESULTADOS - The increase in the purification factor of protease in the top phase was dependent on the molar mass of PEG and concentration of citrate, these variables were significant to $p<0.05$, i.e., 95% of confidence. The target products concentrated in the top phase for all the systems evaluated. The purification factor was 3.32-fold using molar mass of PEG (10000 g/mol) and 12% (w/w) of citrate, with maximal recoveries approaching 100%.
CONCLUSÃO - This finding has implications for the bioprocessing industry, as a simple purification process which is likely to cost very little to implement in most purification facilities, has the potential to recovery biomolecules, such as protease from Clostridium perfringens.

Figura 2.13 - Resumo escrito por um aluno da pós-graduação em Ciências Farmacêuticas durante a segunda fase da experiência de utilização do SciPo-Farmácia na escrita de um *abstract*. O aluno compôs o texto depois de duas semanas de familiarização com esse ambiente computacional.

Nessa segunda versão do resumo, nota-se que dois componentes da estrutura esquemática foram adicionados à versão original: a Lacuna e o Propósito. Isso contribuiu para que mais informações referentes à pesquisa relatada fossem trazidas ao texto, principalmente no que diz respeito ao propósito da pesquisa, descrito no componente esquemático de mesmo nome. Esse tipo de informação trazida pelo Propósito é muito importante em um resumo da área de Ciências Farmacêuticas, uma vez que segundo estudos realizados, esse elemento é recorrente na maioria dos resumos existentes na área em foco e também porque é através dele que o leitor identifica o motivo de se realizar/ter realizado o estudo descrito. Vale ainda dizer que o acréscimo dessas duas estruturas fez com que o resumo ganhasse uma organização estrutural e retórica semelhante ao modelo tido como ideal e sugerido pelos pesquisadores de escrita científica, Swales (1990) e Weissberg e Buker (1990).

Nota-se, ainda, que há um certo equilíbrio quanto ao conteúdo de cada componente de estrutura esquemática do resumo, ou seja, essas estruturas são escritas com quantidades equivalentes de texto e, conseqüentemente, de informação. Característica essa, que segundo a literatura especializada em escrita científica, é um dos indício de se tratar de um texto adequado às convenções exigidas pela comunidade acadêmica em geral. Vale ressaltar, que essas características de adequação podem variar conforme a área da comunidade acadêmica

para a qual o resumo ou artigo científico está sendo escrito, mas que todas as áreas contêm componentes esquemáticos que pertencem a esse modelo ideal.

2.7 Considerações Finais

Neste capítulo, foi apresentada uma revisão sobre ferramentas de auxílio à escrita existentes na literatura especializada. Para tanto, foram apresentadas descrições das principais características de cada um dos sistemas estudados que visam ao pré-processamento do texto, de modo a destacar os níveis de auxílio que podem propiciar ao usuário e de que modo esse tipo de auxílio é efetuado. Vale ressaltar que as ferramentas apresentadas apóiam a escrita do primeiro rascunho e inspiraram a motivação principal desta pesquisa.

A justificativa de nossa escolha se dá por primarmos o auxílio a escritores que precisam produzir seus textos de maneira mais confortável, deixando-os menos pressionados e angustiados no momento de produção. E para isso, comprovou-se por meio de experimentos realizados com as ferramentas computacionais estudadas, que a exposição a bons textos da área na qual se precisa escrever, anotados quanto aos componentes e subcomponentes retóricos, as expressões reutilizáveis, as colocações e aos marcadores discursivos pertinentes e a apresentação de todos esses elementos em seu contexto de uso, só vêem a contribuir para uma boa produção escrita desse aluno.

Esta revisão permitiu também avaliar que tipo de abordagem de auxílio à escrita seria a mais adequada aos objetivos propostos por esse trabalho. Para isso, considerou-se o nível de auxílio proporcionado por cada ferramenta, o tipo de categorização dos componentes e subcomponentes retóricos adotados por cada sistema, e o custo/benefício de implementação da cada uma delas, sempre tendo em mente o público-alvo de nosso projeto.

Após essa avaliação, foi feita a escolha da abordagem e do nível de auxílio que se pretende disponibilizar com o CECARL. O segundo passo de nosso projeto foi consultar a literatura sobre: Lingüística de Córpus, pois o córpus constitui o núcleo do tipo de ferramenta que se pretende gerar com nossa proposta; o que os principais teóricos sobre gêneros discursivos dizem a respeito do artigo científico - gênero textual que será abordado no CECARL -; as particularidades de composição textual de artigos científicos, como os componentes da estrutura esquemática que o moldam, as expressões mais recorrentes, o modo como os marcadores discursivos poderiam ser consultados de modo a promover uma familiarização com o uso adequado dos mesmos. Temas esses que serão apresentados no

próximo capítulo com mais detalhes, e os quais nos guiaram pelo viés teórico na elaboração de nossa proposta.

3. Fundamentação Teórica

3.1 Considerações Iniciais

Esse capítulo está dividido em quatro blocos. O primeiro traz considerações históricas, conceituais e de aplicação da Lingüística de Córpus e de seu objeto de estudo, os córpus, que:

> (...) podem ser definidos como uma coleção de dados lingüísticos (sejam eles textos ou partes de textos escritos ou a transcrição de fala) de uma determinada língua, escolhidos segundo um determinado critério, representando uma amostra desta língua ou uma variedade lingüística. (Berber-Sardinha, 2004)

O segundo cita as abordagens existentes para a investigação do uso da língua em condições reais de ocorrência, demonstradas pelas metodologias e pressupostos da Lingüística de Córpus, compartilhadas por este estudo.

O terceiro apresenta um breve histórico sobre gênero e concepções sobre esse conceito à luz de alguns pesquisadores. Com Aristóteles, pretendemos investigar as origens dos estudos de gênero, observando que pontos ali apresentados encontram pertinência ainda hoje. Bakhtin foi incluso devido à importância de suas conclusões e de suas reflexões a respeito de gênero e de comunidade discursiva. A leitura de Swales foi fundamental devido ao seu já bastante conhecido trabalho com gêneros, principalmente com artigos científicos, o que colabora para o desenvolvimento de uma prática pedagógica que busca desenvolver no aluno a consciência de que através da linguagem escrita ou oral, compartilhada pelos membros de uma sociedade ou grupo social, é possível realizar e negociar seus objetivos comunicativos. As questões de gênero sob as perspectivas de Biber e Marcuschi também foram consultadas.

O quarto procura delinear o artigo científico, indicando eventos que lhe são peculiares em sua composição, como os componentes de sua estrutura esquemática e as estratégias retóricas, bem como as expressões formulaicas, os marcadores discursivos e os termos específicos de uma área que se apresentam, em geral, em textos do gênero científico; uma vez que essas escolhas lingüísticas se dão em função da necessidade social dentro de um contexto sociocultural. As concordâncias e o tema sobre rubricas para avaliação também são tratados nesse último bloco.

3.2.1 Lingüística de Córpus: breve histórico

Atualmente, com o crescente uso de métodos estatísticos que utilizam um grande volume de textos para a extração de dados e informações nas mais diversas áreas da Lingüística e do Processamento de Língua Natural (PLN), a Lingüística de Córpus se encontra em grande evidência, embora córpus venham sendo utilizados em pesquisas há muitos anos. Na Antiguidade e na Idade Média, por exemplo, já eram produzidos córpus de citações da Bíblia para a pregação, a fim de se reproduzir com exatidão os trechos bíblicos desejados (Berber-Sardinha, 2000a: 02).

No entanto, além do apelo natural da lingüística chomskyana[1], para a qual o córpus nunca poderia ser uma ferramenta útil para os lingüistas - pois estes deveriam perseguir em suas pesquisas a modelagem da competência e não do desempenho lingüístico - uma crescente quantidade de críticas feitas ao processamento manual de córpus contribuiu para o desaquecimento de sua utilização na década de 50. Uma das críticas mais contundentes era contra o processamento manual de córpus gigantescos, uma vez que comprometia o valor do estudo, pois o trabalho humano em tarefas repetitivas e extensas como as que estão presentes nesses tipos de investigações propiciam uma ocorrência natural de erros. Segundo Berber-Sardinha (2000a) os trabalhos de Thorndike (em 1921 realizou o levantamento das palavras mais freqüentes em um córpus de 4,5 milhões de palavras da língua inglesa) e Käding (em 1897 coletou manualmente um córpus de 11 milhões de palavras do alemão), por exemplo, foram alvos dessas críticas. Abercrombie (1963[2] *apud* Berber-Sardinha, 2000a) era outro crítico avesso à abordagem baseada em córpus e a resumiu como um conjunto de *pseudo-procedures*, pois imaginava que fazer uma busca por um córpus composto por milhões e milhões de palavras utilizando apenas os olhos, era uma tarefa que consumiria muito tempo, abriria margem para surgimento de erros, sem mencionar o fato de ser custosa (requerer grupos grandes de lingüistas para analisar os dados). Notamos, portanto, que o trabalho com córpus requeria habilidades de processamento de dados que não estavam disponíveis na época, fator que contribuiu para o impacto imediato e profundo das críticas realizadas, deixando a Lingüística de Córpus abandonada, mas não totalmente, durante algum tempo.

[1] Surgida na década de 1950, com o trabalho *Syntactic Structures* de Noam Chomsky. Pode-se resumir através das seguintes características as principais diferenças entre a Lingüística de Córpus e a Lingüística Chomskyana: a primeira possui seu foco no desempenho lingüístico, visa a descrição lingüística e defende uma visão mais empirista da pesquisa científica; a segunda possui seu foco na competência lingüística, visa aos universais lingüísticos e possui uma visão racionalista da pesquisa científica (Leech, 1992:107, tradução Berber-Sardinha 2000).

[2] ABERCROMBIE, D. *Studies in phonetics and linguistics,* London: Oxford University Press, 1963.

Essa situação mudou apenas na década de 60, quando o microcomputador surgiu como uma ferramenta que pôde alterar não somente a maneira de se pesquisar a linguagem, mas também a maneira como ela podia ser enxergada:

> O desenvolvimento do computador com memória poderosa seria para a Lingüística o que o desenvolvimento do microscópio com lentes poderosas foi para a biologia: uma oportunidade não somente de ampliar nosso conhecimento, mas de transformá-lo. (Hoey,1993 tradução de Berber-Sardinha, 2000a)

Desde então, com o advento do computador, o estudo baseado em córpus deixou de receber críticas quanto à imprecisão, porque adquiriu uma segurança notável no processamento extensivo e organizado de dados lingüísticos (McEnery & Wilson, 1996). O primeiro impacto mais notável da adoção dos computadores foi a capacidade de armazenamento de grandes quantidades de linguagem natural (textos escritos, transcrições de conversação, etc.) aumentando, portanto, o campo de visão do lingüista acerca da linguagem.

Um segundo impacto que assegurou a adoção do computador enquanto ferramenta de auxílio na investigação lingüística foi sua capacidade de processar automaticamente quantidades inimagináveis de informação a partir de um córpus. Dessa maneira, uma variedade de ferramentas computacionais (concordanciadores, extratores de palavras-chaves, testes estatísticos, etiquetadores morfossintáticos, etc.) [3] pôde ser utilizada, em combinação ou não, de modo a permitir a identificação de semelhanças e diferenças entre os componentes lingüísticos de um dado córpus eletrônico. Portanto, observa-se que a história da Lingüística de Córpus está intimamente ligada à disponibilidade de córpus eletrônicos, isto é, tratáveis por computadores. Mas o que é um córpus? E ainda, um córpus eletrônico, ou seja, tratável por computador?

A seguir, serão apresentadas considerações sobre a noção de córpus.

3.2.2 A noção de Córpus

Existem na literatura várias definições de córpus e algumas das mais conhecidas são apresentadas a seguir. Segundo Atkins, Clear e Ostler (1992:1), córpus pode ser definido como:

[3] Concordanciadores (WebCorp, disponível em http://www.webcorp.org.uk/ e Concordanciador do projeto Lacio-Web, disponível em http://www.nilc.icmc.usp.br/lacioweb/macmorpho.php). Extrator de Lista de palavras-chaves (Ferramenta KeyWord da suíte WordSmith Tools). Teste Estatístico: estatística *Kappa*, Parser PALAVRAS, Eckhard Bick (Bick 2000), disponível em http://visl.sdu.dk/visl/pt/parsing/automatic.

> (...) um subconjunto de uma biblioteca eletrônica de textos, construída conforme critérios específicos necessários a um determinado propósito, como por exemplo, o Cobuild Corpus e o Longman/Lancaster Corpus[4].

Para McEnery & Wilson (1996:21), "em princípio, qualquer coleção composta por mais de um texto pode ser chamada de córpus: o termo corpus em Latim corresponde a corpo, portanto um córpus pode ser definido como qualquer corpo de texto[5].

Outra definição aceita na literatura é proposta por Kennedy (1998:1), na qual um córpus é "um corpo de texto escrito ou de fala transcrita que pode servir como base para análise e descrição lingüística".

Mas o termo 'córpus' quando utilizado no contexto da Lingüística de Córpus tende a apresentar, frequentemente, conotações mais específicas, tais como a Amostragem e Representatividade de um córpus, o seu Tamanho, o Formato Computável e o Padrão de Referência (McEnery & Wilson, 1996:21). Portanto, uma definição que segundo Berber-Sardinha (2000a) seria a mais adequada, pois traz em si as principais características modernas há pouco citadas, foi criada por Sanchez e será a adotada por esta pesquisa, uma vez que corrobora a noção de córpus adotada para o projeto no qual este estudo está inserido:

> Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise. (Sanchez, 1995: 8[6], apud Berber-Sardinha, 2000a)

Esta definição é, segundo Berber-Sardinha (2000a), mais completa que as anteriormente apresentadas, porque incorpora vários pontos importantes da Lingüística de Córpus, como:

1) A origem dos dados, que devem ser autênticos. Textos autênticos são textos naturais, isto é, aqueles que existem naturalmente na linguagem, e que não foram produzidos com o objetivo de fazerem parte de um dado córpus. Além disso, incluída nessa idéia de 'natural' estão somente os textos produzidos por seres humanos. Dessa forma, ficam excluídos os textos

[4] Texto Original: " (…) a subset of an ELT, built according to explicit design criteria for a specific purpose, e.g. The Cobuild Corpus, the Oxford Pilot Corpus". (tradução minha)
[5] "In principle, any collection of more than one text can be called a corpus: the term 'corpus' is simply the Latin for 'body', hence a corpus may be defined as any body of text". (tradução minha)
[6] SANCHEZ, A. Definición e historia de los corpus. In: SANCHEZ, A. *et al*. (Org.) *CUMBRE – Corpus Lingüístico del Español Contemporáneo*. Madrid: SGEL, pp. 7-24, 1995.

gerados por programas que possuem essa finalidade, a linguagem de programação de computadores e a linguagem de notação matemática ou científica.

2) O propósito de um córpus, que deve ter a finalidade de ser um objeto de estudo lingüístico. A escolha dos textos que constituem um córpus que será submetido à análise e a observações não é aleatória, isto é, nem todo conjunto de textos é considerado um córpus. A reunião dos textos será delimitada de acordo com os objetivos da investigação, podendo ser o córpus constituído de textos jornalísticos, ou científicos, por exemplo. Tognini-Bonelli também chama a atenção para esse fato do córpus ter de servir a um propósito específico do pesquisador, notando que ao se determinar a função de um córpus subjaz a relação entre a metodologia escolhida e sua amostragem. Um córpus, segundo essa autora, "precisa ser justificado em termos lingüísticos" e seus textos devem ser "selecionados segundo critério explícito de modo a capturar as regularidades de uma língua, variedade lingüística ou sublíngua". (Tognini-Bonelli, 2001:55)[7]

3) A composição de um córpus, cujo conteúdo deve ser criteriosamente escolhido. Ou seja, os princípios da escolha de textos devem sempre seguir as condições de naturalidade e autenticidade, como também devem obedecer a um conjunto de regras estabelecidas pelos seus criadores de modo que o córpus coletado corresponda às características que se deseja dele.

4) O formato computável do córpus: A coleta, armazenamento e manipulação de um córpus de pesquisa precisa, hoje, levar em consideração as técnicas oferecidas e, também, a existência de uma comunidade científica que preza pelas condições de comunicabilidade do trabalho científico. A idéia do laboratório hermético, do 'calabouço' de produção científica totalmente fechado ao mundo até a apresentação dos resultados é algo arcaico e incoerente. A ciência se dá, também, através da colaboração, que, não raro, é muito útil ao pesquisador. Por essa razão, a construção de um córpus em formato eletrônico e computável, que permita não apenas a disponibilização do mesmo através de sites ou outros meios de comunicação eletrônica, mas também a manipulação de dados por ferramentais específicos, é uma tendência que não pode ser ignorada. Outra vantagem desse padrão de formatação é a possibilidade de armazenamento de córpus num computador e, conseqüentemente ter sua

[7] Texto original: "(...) needs to be justified in linguistics terms. (…) The texts are selected according to explicit criteria in order to capture regularities of a language, a language variety or a sub-language". (tradução minha)

existência prolongada, isto é, não se corre o risco de perder o material através de deterioração com o passar do tempo. Além de sua maior facilidade no manuseio, esse tipo de formato facilita a disponibilização de córpus para outras pesquisas/pesquisadores.

5) A representatividade ou o balanceamento: Segundo Biber, a noção de representatividade se refere "ao tamanho que uma amostra deve conter para abranger toda a variabilidade de uma população" (Biber, 1993:243)[8]. Tognini-Bonelli sintetiza a questão da representatividade de um córpus, dizendo:

> Assim parece haver um acordo geral entre os pesquisadores que escolheram trabalhar com um córpus de que este deveria ser representativo de certa população e de que as afirmações geradas a partir da análise do córpus serão amplamente aplicáveis a uma amostra maior ou ao todo da linguagem. [9] (Tognini-Bonelli, 2001:57[10] apud Berber-Sardinha, 2004)

Assim sendo, tradicionalmente, tem-se a tendência de se ver um córpus como um conjunto representativo de uma variedade lingüística ou mesmo de um idioma. No entanto, essa questão da representatividade divide pesquisadores que trabalham com córpus em dois grupos: aqueles que acreditam que a representatividade é alcançada com um balanceamento na quantidade de gêneros, assuntos e tipos textuais de um córpus e aqueles que acham que a representatividade se consegue com um grande volume de dados, ou seja, quanto maior, melhor. Na primeira linha, são encontrados pesquisadores como John Sinclair e na segunda, Sir Randolf Quirk (Church & Mercer, 1993:17-19). Ocorre que em certas pesquisas, como a construção de um dicionário de língua geral, por exemplo, é necessário realmente um grande volume de dados, pois é preciso encontrar a maior quantidade de significados de uma dada palavra. Já em outras pesquisas, como por exemplo, a terminológica, essa necessidade de um grande volume de dados não se justifica, mas sim a de um bom balanceamento em termos de assuntos/tópicos e gêneros de uma dada área de pesquisa. Em suma, pode se dizer que essa divergência se resume ao questionamento quanto à extensão de um córpus ser representativa (grande) ou balanceada (equilibrada). A decisão, no entanto, de construir um córpus balanceado ou representativo vai depender do propósito a que o córpus se destina.

[8] "to the extent which a sample includes the full range of variability in a population". (tradução minha)
[9] Texto original: "Thus there seems to be general agreement among scholars who choose to work on a corpus that this should be representative of a certain population and that the statements derived from the analysis of the corpus will be largely applicable to a larger or to the language as a whole". (tradução minha)
[10] TOGNINI-BONELLI, E. *Corpus Linguistics at Work*. Amsterdam/Atlanta, GA: John Benjamins, 2001.

3.2.3 Usos de córpus

O estudo com córpus, conforme pode ser observado em McEnery & Wilson (1996:88), pode trazer contribuições para muitas áreas de pesquisa, como: a pesquisa com textos da Fala, a Análise do Discurso, nos estudos lexicais, em estudos sobre a Gramática, a Semântica, a Pragmática, a Sociolingüística, a Estilística, o Ensino de Línguas, a Lingüística Histórica, a Dialetologia, a Psicolingüística, os Estudos Culturais, a Psicologia Social, etc; sem esquecer de incluir nessa lista, é claro, a área de tecnologia da linguagem, para a qual o uso de córpus possibilita o desenvolvimento de sistemas de tradução automática, corretores ortográficos, gramaticais e estilísticos, ferramentas de auxílio à escrita, sumarizadores textuais, entre outros.

Partington (1998:2[11] *apud* Jacobi-Blaszkowski, 2000) também traz um panorama das principais áreas de análise lingüística que utilizam computadores e córpus. Entre elas estão:

- a produção de material didático;
- os estudos de estilística e de autoria que têm como objetivo identificar as características distintivas de um determinado escritor (chamada Lingüística Forense);
- os estudos diacrônicos ou históricos que comparam a língua de diferentes períodos com o objetivo de obter informação sobre mudanças lingüísticas como é o caso do Projeto Tycho Brahe[12], cujo objetivo principal é modelar a relação entre prosódia e sintaxe na mudança lingüística que deu origem ao Português Europeu Moderno a partir do Português Clássico;
- os estudos de análise textual que descrevem fenômenos lingüísticos que vão além da oração (Stubbs, 1996[13] *apud* Jacobi-Blaszkowski, 2000);
- os estudos sobre a língua falada, como, por exemplo, o de Tognini-Bonelli (1993 *apud* Jacobi-Blaszkowski, 2000), que focaliza a forma como o falante organiza o seu discurso;
- os estudos de tradução (Gavioli & Mansfield, 1990[14] *apud* Jacobi-Blaszkowski, 2000);
- os estudos de registro que utilizam córpus para comparar variedades de uma mesma língua. Biber (Biber, Conrad and Reppen, 1998) é um dos pesquisadores que mais se destaca

[11] PARTINGTON, A.Patterns and Meanings – Using corpora for English Language Research and Teaching. Amsterdam/Philadelphia: John Benjamins, 1998.
[12] http://www.ime.usp.br/~tycho/
[13] STUBBS, M. *Text and corpus analysis*. Oxford: Blackwell, 1996.
[14] GAVIOLI, L & MANSFIELD, G. *The PIXI corpora: bookshop encounters in English and Italian*. CLUEB, Bologna, Italy, 1990.

nessa área e Nakamura (1993[15] *apud* Jacobi-Blaszkowski, 2000); e Nakamura e Sinclair (1995[16] *apud* Jacobi-Blaszkowski, 2000), que descrevem métodos para classificar semi-automaticamente textos segundo sua tipologia; e assim por diante.

Se focarmos na área de Ensino de Inglês como língua estrangeira, por exemplo, – um dos objetivos subjacentes ao nosso estudo - podemos perceber que amostras de linguagem derivadas de córpus têm se tornado cada vez mais importantes no ensino-aprendizagem de línguas, pois os córpus têm servido como fontes ricas de língua-alvo utilizada por falantes nativos em diferentes contextos/situações da vida real, as quais têm sido tratadas com insucesso, na maioria das vezes, por materiais de ensino-aprendizagem, por exemplo, de inglês como língua estrangeira (*EFL - English as a Foreign Language*). Leech (1997[17] *apud* Berber-Sardinha, 2004) afirma que um córpus "habilita o aprendiz/estudante a explorar, investigar, generalizar, testar hipóteses (…) [ele é] fonte de aprendizado lingüístico*"*.[18] Aston (1997[19] *apud* Berber-Sardinha, 2004) também defende esse ponto de vista ao dizer que um córpus

> (...) oferece aos professores e aprendizes uma grande variedade de material que pode ser utilizado com a finalidade de aprendizagem de língua. Os propósitos do ensino de língua podem ser melhorados pelo acesso a córpus de textos em língua estrangeira (...).[20]

Poderíamos dizer que esse sucesso do córpus na área de ensino-aprendizagem de línguas pode ser justificado, por exemplo, pelos seguintes fatores:

1. As regras derivadas de dados lingüísticos reais de falantes nativos do inglês podem melhorar a competência comunicativa de estudantes de inglês como língua estrangeira, uma vez que nesse caso a linguagem em uso é considerada para se reconhecer padrões de uso e não o contrário: procurar na linguagem em uso padrões idealizados por uma linguagem ideal.

---

15 NAKAMURA, J. Statistical methods and large corpora: A new tool for describing text types. In BAKER, M., FRANCIS, G. & TOGNINI-BONELLI, E. (eds) Text and Technology. Amsterdam: John Benjamin. 313-332, 1993.

16 NAKAMURA, J. & SINCLAIR, J. The world of woman in the Bank of English. *Journal of Literary and Linguistic Computing,* v. 2, 1995.

17 LEECH, G. Teaching and language corpora: a convergence. In: WICHMANN, A.; Fligelstone, S.; Mc ENERY, T.; and KNOWLES, G. (eds.). Teaching and language corpora. London: Longman, p. 1-23, 1997.

18 Texto Original: "enables the learner/student to explore, to investigate, to generalize, to test hypotheses (…) [It is] a linguistic learning resource."

19 ASTON, G. Enriching the learning environment: corpora in ELT. In WICHMANN, et al (eds) Teaching and language corpora, 1997.

20 "(…) offers teachers and learners an enormous range of material which might be used for language-learning purposes. The purposes of language pedagogy may best be served by access to córpus of foreign language texts (…)" (Aston, 1997). (tradução minha)

2. Pode habilitar os professores a cultivarem em seus alunos um espírito observador e autodidata sobre a língua estrangeira em estudo, por meio da exploração de exemplos relevantes de textos contidos em córpus.

3. O córpus também serviu/serve para mudar o papel de alunos e professores: os professores não precisam ser apenas professores, facilitadores e gerenciadores do processo de ensino-aprendizagem, mas também aprendizes e pesquisadores. Por sua vez, os estudantes não precisam apenas ser aprendizes, pois podem se tornar pesquisadores e professores.

Em Ide and Brew (2000), a **reusabilidade** (característica de um córpus ser usado em mais de um projeto de pesquisa e por mais de um grupo de pesquisadores) e a **extensibilidade** (isto é, a capacidade de córpus serem melhorados em várias direções, por exemplo, com a provisão de um nível a mais de análise lingüística) são colocadas como dois aspectos a serem considerados em projetos de córpus, principalmente nos projetos de grandes córpus.

Projetos de grandes córpus, como o *British National Corpus*[21] (BNC), para a variante britânica do inglês, e o *American National Corpus*[22] (ANC), para a americana, contribuem para a descrição da língua inglesa e a construção de recursos, tais como dicionários e gramáticas. Eles contribuem, também, para o desenvolvimento de ferramentas para o Processamento de Língua Natural (PLN), como lematizadores[23], etiquetadores morfossintáticos[24], sintáticos[25] e anotadores de co-referência[26] que, por sua vez, dão suporte para a própria construção das anotações lingüísticas desses grandes recursos. Isto porque, para se progredir de maneira rápida e confiável na compreensão da história das línguas, por exemplo, é necessário que córpus de estudo estejam anotados, e que se tenha um arcabouço de ferramentas simples, como contadores de freqüência[27] e concordanciadores, mas também as mais elaboradas, como geradores de n-gramas[28], de colocações[29] e acesso a léxicos[30], para elencar neologismos ou palavras que caíram em desuso.

---

21 http://www.natcorp.ox.ac.uk/
22 http://americannationalcorpus.org/
23 Lematizador: é uma ferramenta informatizada que auxilia a marcação no texto da forma canônica, não flexionada, da palavra (aquela que ocorre normalmente na entrada de um dicionário convencional).
24 Etiquetador Morfossintático: é uma ferramenta informatizada que detecta automaticamente as informações morfológicas e sintáticas de todas as palavras de um córpus.
25 Etiquetador Sintático: é uma ferramenta informatizada que detecta automaticamente as informações sintáticas de todas as palavras de um córpus
26 Anotador de co-referência: é uma ferramenta informatizada que faz a detecção automática da ocorrência de múltiplos substantivos (ou nomes) de dado discurso que se referem a uma mesma entidade, objeto ou evento.
27 Contador de freqüência: é uma ferramenta informatizada que calcula a freqüência de todas as palavras do córpus escolhido.
28 N-grama: é uma seqüência de cadeias de caracteres de comprimento n. Exs: uva (unigrama); de lado (bi-grama); lado a lado (tri-grama), etc.

O que pode ser menos visível numa primeira análise é que esses grandes projetos impulsionam também o desenvolvimento de formatos de padrões de anotação e codificação, como o atual XCES[31] (Ide *et al*, 2000), que utiliza XML[32] como linguagem de codificação, bem como de ferramentas computacionais aceitas internacionalmente para a manipulação de córpus.

Em relação à anotação, são basicamente dois os níveis de representação de informações presentes em um córpus: a anotação estrutural e a anotação lingüística. A **anotação estrutural** compreende a marcação de dados externos e internos dos textos. Como dados externos, entendemos a documentação do córpus na forma de um cabeçalho que inclui dados bibliográficos comuns, dados de catalogação como tamanho do arquivo, tipo da autoria, resumo do texto (se houver), e uma tipologia textual - por exemplo, a tipologia quadripartida utilizada no Projeto Lácio-Web, que trata do gênero, tipo textual, meio de distribuição e domínio de um texto (veja mais detalhes desta tipologia em http://www.nilc.icmc.usp.br/lacioweb/tipologia.htm). Como dados internos, temos a **anotação de segmentação** do texto cru que cuida da: a) marcação da estrutura geral – capítulos, parágrafos, títulos e subtítulos, notas de rodapé e elementos gráficos como tabelas e figuras; e b) marcação da estrutura de subparágrafos – elementos que são de interesse lingüístico, tais como sentenças, citações, palavras, abreviações, nomes, referências, datas e palavras em negrito, isto é, destacadas. No processo de codificação dos dados são utilizados dois elementos: um elemento chamado cabeçalho (dados externos) e outro chamado corpo (texto cru mais anotação de segmentação).

---

29 Colocações: são combinações fixas ou semi-fixas, constituídas por substantivo + substantivo (ex: credit card, quality control), substantivo + adjetivo (ex: nursing home, silent movie), substantivo como sujeito + verbo (ex: ariver flows, a volcano erupts) ou verbo + substantivo como objeto (pay a visit), verbo + advérbio (pay dearly) e adjetivo + advérbio (deeply hurt). Definição retirada de http://www.cadernos.ufsc.br/download/9/pdf/Stella-Cadernos9.pdf.

30 Léxico: O léxico de uma língua engloba o conjunto de signos por meio dos quais o homem não só se expressa, se comunica, mas também cria novos conhecimentos e/ou assimila conhecimentos que outros homens criaram, não só na sua civilização, mas também em outras civilizações. Definição retirada de http://www.ime.usp.br/~is/educar2002/dicionarios/dicionarios.html.

[31] http://www.cs.vassar.edu/XCES/

[32] XML (Extensible Markup Language), em português: Linguagem de Marcação Estendida. O XML permite que você crie os seus próprios conjuntos de elementos de marcação. É uma maneira simples e padrão de delimitar os dados do texto. Informações retiradas de http://www.webtutoriais.com/open.php?cut=1670.

A Figura 3.1 mostra o cabeçalho de um texto do córpus global do Projeto PLN-BR[33]. Importa notar que tal cabeçalho segue as recomendações do padrão XCES, cujos esquemas estão disponíveis em http://www.xces.org/schema/2003/, *link* que pode ser visto logo no início do cabeçalho. O cabeçalho do padrão XCES é formado por quatro elementos principais, todos opcionais como mostram as linhas pontilhadas na Figura 3.2: 1) <fileDesc> que contém informações sobre o texto codificado (distribuição, fonte, etc.); 2) <encodingDesc> que contém informações sobre a maneira como o texto foi codificado; 3) <profileDesc> que contém informações sobre vários aspectos do texto (língua usada, classificação do texto segundo a sua tipologia, os participantes de um texto falado e sua situação, anotações, etc.); e 4) <revisionDesc> que resume o histórico de revisão (cabeçalho, segmentação e lingüística) de um documento. Importante citar, também, que um cabeçalho em XML como este, embora seja altamente legível por humanos, é para ser processado por programas computacionais, que reconhecerão os campos do cabeçalho, mostrando para um consulente somente os que são adequados em uma dada situação. A forma com que tais informações são apresentadas aos consulentes pode variar.

A seguir é apresentado um cabeçalho de um texto do córpus denominado PLN-BR CATEG do Projeto PLN-BR que pertence ao gênero informativo, subgênero jornalístico, tipo de texto notícia, meio de distribuição jornal e nenhum domínio ou subdomínio inserido. As informações sobre a tipologia do texto são fornecidas no campo <catRef> do cabeçalho. As palavras-chaves de tal texto são EUA, Férias, Parque, Passeio, Atração, Orlando, Montanha-Russa, Universal Orlando, Simulador, como mostra o campo <keywords>. Como pode ser notada, a anotação dessas informações são realizadas em XML, pois as informações referentes ao texto são apresentadas entre os sinais "<" e ">".

[33] *PLN-BR: Recursos e Ferramentas para a recuperação de Informações em Bases Textuais em Português do Brasil*[33], que tem duração de 2 anos a partir de 2006 e é financiado pelo CNPq/CTInfo (#550388/2005-2). Vinculado a ele, estão sete subprojetos: (1) Construção, Manutenção e disponibilização de Córpus (NILC/ Universidade de São Paulo (USP), campus de São Carlos); (2) Anotação de Córpus (Universidade do Vale do Rio dos Sinos - UNISINOS); (3) Glosagem da Wordnet.Br e sua indexação à WordNet de Princeton (Universidade Estadual Paulista (UNESP), campus de Araraquara); (4) Aprendizagem Automática de Informações Lexicais (Pontifícia Universidade Católica do Rio de Janeiro (PUC-RJ); (5) Sumarização Automática e recuperação da Informação Textual (Universidade Federal de São Carlos (UFSCar); (6) categorização de Textos (Pontifícia Universidade Católica do Rio Grande do Sul (PUC-RS); (7) Representação do Conhecimento Textual (Universidade Presbiteriana Mackenzie).

<?xml version="1.0" encoding="UTF-8" ?>
<cesHeader xmlns="http://www.xces.org/schema/2003"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.xces.org/schema/2003" version="1.0.4">
<fileDesc>
<titleStmt>
<title>2000jul_9275</title>
<respStmt>
<respType>Criação do Header</respType>
<respName type="person">Kleber Infante</respName>
</respStmt>
<respStmt>
<respType>Criação do Header</respType>
<respName type="person">Marcelo Muniz</respName>
</respStmt>
</titleStmt>
<extent>
<wordCount>377</wordCount>
<byteCount units="bytes">4494.0</byteCount>
<extNote>2</extNote>
</extent>
<publicationStmt>
<pubAddress>Av. Trabalhador São-carlense, 400 - Centro, Caixa Postal: 668 - CEP:
13560-970 - São Carlos - SP</pubAddress>
<telephone>+55 16 33739663</telephone>
<eAddress type="www">http://www.nilc.icmc.usp.br</eAddress>
<pubDate>2006</pubDate>
</publicationStmt>
<sourceDesc>
<biblStruct>
<monogr>
<title>Filme 3D produz queda inocente de 122 m</title>
<title>Simulador põe visitante dentro do mundo do Homem-Aranha; montanha-
russa arremessa ao céu</title>
<author>DA ENVIADA ESPECIAL A ORLANDO</author>
<respStmt>
<respType>crédito</respType>
<respName type="institution">DA ENVIADA ESPECIAL A ORLANDO</respName>
</respStmt>
<imprint>
<pubPlace>Folha de São Paulo</pubPlace>
<publisher type="org">Empresa Folha da Manhã S.A.</publisher>
<pubDate>03/07/2000</pubDate>
<pubAddress>São Paulo</pubAddress>
</imprint>
<biblNote>TURISMO</biblNote>
<biblScope type="PP">G16</biblScope>
</monogr>
</biblStruct>
</sourceDesc>
</fileDesc>
<encodingDesc>

<projectDesc>**O projeto Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil (PLN-BR) - CNPq/CTInfo #550388/2005-2 - está subdividido em 7 subprojetos relativamente autônomos, mas que compartilham o mesmo ponto de partida - qual seja, o tratamento da informação mobilizada em um mesmo córpus do português do Brasil - e tem por objetivo geral a construção de um espaço interinstitucional de interação e intercâmbio de práticas de análise e investigação lingüístico-computacional acerca da representação e da recuperação de informação de natureza semântica e pragmático-discursiva veiculada por enunciados produzidos em português brasileiro. O projeto vincula pesquisadores da Universidade de São Paulo (USP), campus de São Carlos; da Universidade Federal de São Carlos (UFSCar); da Universidade Estadual Paulista (UNESP), campus de Araraquara; à Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS); da Pontifícia Universidade Católica do Rio de Janeiro (PUCRJ); da Universidade do Vale do Rio dos Sinos (UNISINOS); e da Universidade Presbiteriana Mackenzie.**</projectDesc>
<samplingDecl>**PLN-BR CATEG é o córpus formado por textos do jornal Folha de São Paulo que podem ser acessados integralmente na Web por usuários que detenham senha específica de acesso. Foi criado exclusivamente como recurso de testes em software sem fins comerciais de recuperação de informação, de categorização, de classificação e de agrupamento de textos. Ele é uma amostra aleatória estratificada e proporcional à distribuição do córpus global do projeto PLN-BR (chamado de PLN-BR FULL) com relação aos textos dos cadernos do jornal. Ele é formado por 30% dos textos do córpus PLN-BR FULL, o que equivale a 30.000 textos, e possui somente notícias e reportagens para as quais a Folha de São Paulo possui direitos de republicação. Este córpus contém o córpus PLB-BR GOLD, também criado no escopo do projeto PLN-BR. O córpus PLN-BR FULL, por sua vez, é formado por 103,080 mil textos do jornal Folha de São Paulo, compondo um ano construído a partir do ano de 1994 (toma um mês aleatório até o ano de 2005). A classificação em notícias e reportagens foi feita de forma automática usando-se um classificador de tipos de textos treinado com os 40 tipos de textos do Projeto Lácio-Web (http://www.nilc.icmc.usp.br/lacioweb/) no córpus montado para o projeto de doutorado de Rachel Aires que foi defendido no ICMC-USP em 2005 sob orientação da Profa. Sandra Aluísio (mais informação sobre o classificador em http://www.nilc.icmc.usp.br/nilc/projects/linguarudo.html).**</samplingDecl>
</encodingDesc>
- <profileDesc>
- <textClass>
<catRef target="**genero.8 genero.8.18 genero.8.18.10 distribuicao.12 tipotextual.35**" />
- <keywords>
<keyTerm>**EUA**</keyTerm>
<keyTerm>**FÉRIAS**</keyTerm>
<keyTerm>**PARQUE**</keyTerm>
<keyTerm>**PASSEIO**</keyTerm>
<keyTerm>**ATRAÇÃO**</keyTerm>
<keyTerm>**ORLANDO**</keyTerm>
<keyTerm>**MONTANHA-RUSSA**</keyTerm>
<keyTerm>**UNIVERSAL ORLANDO**</keyTerm>
<keyTerm>**SIMULADOR**</keyTerm>
</keywords>
</textClass>
- <annotations>
<annotation type="**logical**" ann.loc="**TURISMO_2000_29416-logical.xml**">**Logical markup**</annotation>

```
<annotation type="s" ann.loc="TURISMO_2000_29416-s.xml">Sentence
   boundaries</annotation>
<annotation type="content" ann.loc="TURISMO_2000_29416.txt">Document
   content</annotation>
   </annotations>
   </profileDesc>
   </cesHeader>
```

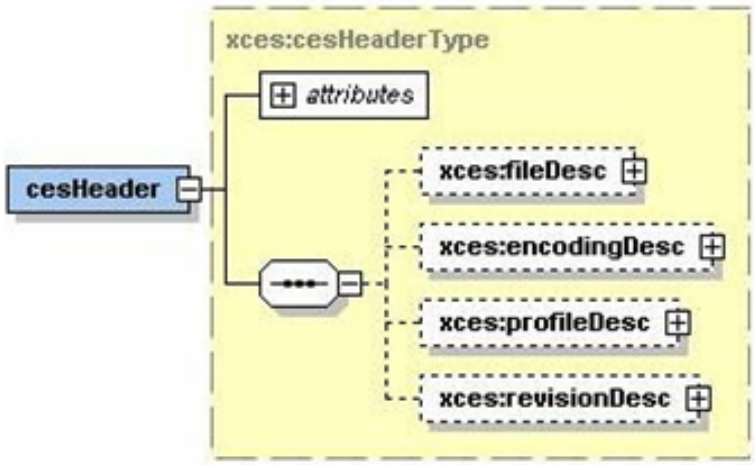


Figura 3.2. Os quatro elementos de um cabeçalho segundo o padrão XCES.

A **anotação lingüística** pode se dar em qualquer nível (morfológico, morfossintático, sintático, semântico, retórico, etc.) e pode ser inserida por três formas: manualmente (por lingüistas), automaticamente (por ferramentas de PLN) ou semi-automaticamente (correção manual da saída de outras ferramentas). Essa última forma de anotação tem provado ser, segundo experiências relatadas sobre anotação de córpus (por exemplo, projeto Lacio-Web), a mais eficiente, pois revisar é mais rápido e gera dados mais corretos do que anotar um córpus pela primeira vez.

3.2.4 Status da Lingüística de Córpus: abordagem, metodologia ou disciplina

Na literatura, há certa convergência quanto às finalidades e objetivos da Lingüística de Córpus (Sinclair, 1991; McEnery & Wilson, 1996; Biber *et al*, 1998; Berber-Sardinha, 2004), cujo principal papel seria o de ser meio/instrumento pelo qual se torna possível a investigação da estrutura da linguagem. Essa convergência pode ser representada, por exemplo, pela acepção de Berber-Sardinha:

> A Lingüística de Córpus ocupa-se da coleta e exploração de córpus, ou conjuntos de dados lingüísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador. (Berber-Sardinha, 2004: 3)

Por outro lado, nessa mesma literatura, pode ser apontada uma constante divergência entre os estudiosos da área quanto ao *status* da Lingüística de Córpus: seria ela uma disciplina, uma metodologia ou uma abordagem?

Afirma-se que a Lingüística de Córpus não é uma disciplina tal qual a sociolingüística ou a semântica, pois seu objeto de pesquisa não é delimitado como em outras áreas (Berber-Sardinha, 2004:35). Além disso, a Lingüística de Córpus não se dedica a um assunto definido, mas a vários fenômenos comumente enfocados em outras áreas, como o léxico e a sintaxe (Leech, 1992:106[34] *apud* Berber-Sardinha, 2004). Seria seguro afirmar que a Lingüística de Córpus é, então, uma metodologia que as outras áreas podem utilizar? Também não, pois a Lingüística de Córpus não se resume apenas a uma metodologia disponível para outras áreas (McEnery & Wilson, 1996:1; Berber-Sardinha, 2000a: 355; Tognini-Bonelli, 2001:1[35] *apud* Berber-Sardinha, 2004). Essa, por exemplo, possibilita a investigação de comportamentos lexicais (estudo das colocações), pesquisas típicas realizadas por lingüistas de córpus como John Sinclair, que não encontram espaço em outras disciplinas:

> Ela possui caráter essencialmente ascendente e tem como doutrina a não categorização a priori (*trust the text* é o seu lema). Por isso, exemplifica com precisão a prática empirista e situa-se como o pólo mais distante das abordagens racionalistas. Aliás, foi por isso mesmo que uma das maiores correntes de pesquisa em córpus surgiu. (Berber-Sardinha, 2000a)

Uma outra razão pela qual a Lingüística de Córpus não pode ser considerada metodologia é o fato de seus praticantes produzirem conhecimento novo:

> Embora o escopo da Lingüística de Córpus possa ser definido em termos do que as pessoas fazem com córpus, seria um engano assumir que Lingüística de Córpus é somente um meio mais rápido de descrever como a linguagem funciona (…). A análise de um córpus pode revelar, e freqüentemente revela, fatos a respeito de uma língua que nunca se pensou em procurar. (Kennedy, 1998: 9, tradução de Berber-Sardinha, 2000a)

[34] LEECH, G. Corpora and theories of linguistic performance. In SVARTVIK, J. (Org.). *Directions in Corpus Linguistics*. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991, p. 105-127. Berlin, New York: De Gruyter, 1992.
[35] TOGNINI-BONELLI, E. *Corpus Linguistics at Work*. Amsterdam/Atlanta, GA: John Benjamins, 2001.

Uma terceira possibilidade que se apresenta é a de que a Lingüística de Córpus não seja nem disciplina nem metodologia. Segundo Hoey, "Lingüística de Córpus não é um ramo da lingüística, mas a rota para a lingüística" (Hoey, 1997[36], tradução de Berber-Sardinha, 2004). Segundo essa definição, a Lingüística de Córpus seria uma perspectiva, isto é, uma maneira de se chegar à linguagem. Essa definição faz alusão ao conceito de teoria lingüística enquanto 'janela' que molda como enxergamos a linguagem (Pike, 1972[37] *apud* Berber-Sardinha, 2000a). Dessa forma, a Lingüística de Córpus não seria apenas um instrumental, mas sim uma abordagem. De modo similar, Leech (1992:106) a define "não somente como uma nova metodologia emergente para o estudo da linguagem, mas uma nova empreitada de pesquisa e, na verdade, uma nova abordagem filosófica". Daí a preferência de alguns influentes lingüistas de córpus pelo termo *Corpus-Based Approach*. Esse termo, porém, é utilizado com algumas ressalvas por Biber *et al* (1998:8), pois para eles, essa abordagem pode complementar outras abordagens já existentes, e (...) não [ser adotada] como a única abordagem correta".

A seguir serão apresentadas as abordagens de Investigação Lingüísticas que utilizam córpus como objeto de estudo.

3.3 Abordagens para a Investigação Lingüística

A investigação lingüística, que utiliza córpus como objeto de seu estudo, está situada dentro de um quadro teórico que se divide em duas vertentes: a primeira delas de natureza comprobatória, na qual os dados levantados podem servir como exemplo para uma teoria previamente elaborada pelo pesquisador; e a segunda, de natureza exploratória, na qual o pesquisador procura padrões ou distinções entre os dados lingüísticos para servirem como base na formulação de generalizações, a fim de se chegar a uma teoria lingüística. Para Tognini-Bonelli (2001) e Hunston (2002 *apud* Possamai, 2004)[38], as abordagem de estudo com córpus são respectivamente conhecidas como *Corpus-Based* e *Corpus-Driven*. Na primeira abordagem podem ser citados trabalhos de Aarts (1991 *apud* Kauffman, 2005)[39] e Leech (1991)[40]; já no segundo, os de Sinclair (1991) e Tognini-Bonelli (2001).

---

[36]HOEY, M . From concordance to text structure: New uses for computer corpora. In: LEWANDOSWKA-TOMASZCZYK, B. & MELIA, P.J. (org.). PALC'97 – Practical Applications in Language Corpora. Lodz: Lodz University Press, 1997.

[37] PIKE, K.L. Towards a theory of the strcture of human behavior. In BREND, R.M. (org). *Kenneth L Pike: selected writings*. Hague, Mouton, pp. 106-16, 1972.

[38] HUNSTON, S. Corpora in Applied Linguistics. Cambridge: Cambridge University Press, 2002.

[39] AARTS, B. Intuition-based and observation-based grammars. In: AIJIMER, K.; ALTENBERG, B. (orgs.). *English Corpus Linguistics* – Studies in Honour of Jan Svartvik. London: Longman, 1991.

[40] LEECH, G. The state of the art in corpus linguistics. In AIJIMER, K.; ALTENBERG, B. (org.). *English Corpus Linguistics* – Studies in Honour of Jan Svartvik. London: Longman, 1991.

Mais detalhes sobre essas abordagens de estudo com córpus são apresentados a seguir, com a exposição de críticas feitas à abordagem C*orpus-Based*. Há também, em contrapartida, a apresentação de estudos sobre gêneros na perspectiva do ESP – *English for Specific Purposes*, que utilizaram com sucesso esse tipo de abordagem.

3.3.1 Abordagem Dirigida por Córpus (*Corpus-Driven Approach*)

Nessa abordagem de investigação da linguagem, a afirmação teórica é formulada após a extração dos dados do córpus de estudo. Conseqüentemente, esse córpus será responsável pela teoria e gerará uma mudança qualitativa na descrição da língua (Tognini-Bonelli, 2001:11).

Ainda nessa abordagem, as afirmações teóricas irão refletir e se basear nas evidências geradas pelo córpus. Nesse caso, pode-se supor que os padrões recorrentes e a freqüência das palavras no córpus fornecerão as evidências necessárias para categorias lingüísticas, conforme cita a autora, "(...) o caminho metodológico geral é claro: observação conduz a hipóteses, que conduzem a generalização, que por sua vez conduz à unificação em afirmação teórica". (Tognini-Bonelli, 2001:85)[41]

A abordagem dirigida por córpus pode ser utilizada em nossa pesquisa para avaliar se nosso córpus em estudo, o Córpus Met[42], é balanceado ou não. Para isso, podem ser extraídas informações estatísticas relativas a: (1) quantidade de exemplos de cada estratégia retórica contida no córpus, e a (2) quantidade de textos contida em cada subárea das Ciências Farmacêuticas, segundo uma árvore de domínios elaborada para esta área de pesquisa (mais detalhes ver Capítulo 4).

3.3.2 Abordagem Baseada em Córpus (*Corpus-Based Approach*)

Tognini-Bonelli (2001:65) menciona que as teorias lingüísticas:

> (...) são o resultado de reflexão de um pesquisador depois de ter incorporado uma grande quantidade de experiência com língua e linguagem e de ter testado as implicações e conseqüências com referência à intuição de falantes competentes ou nativos.[43]

[41] "(...) the general methodological path is clear: observation leads to hypothesis leads to generalization leads to unification in theoretical statement."[41] (Tognini-Bonelli, 2001:85) (tradução minha)

[42] Córpus Met: córpus construído com 30 seções metodologias de artigos científicos em inglês da área de Farmácia. Foi utilizado na implementação da seção "Metodologia" do SciPo-Farmácia. Mais detalhes sobre esse córpus, ver Capítulo 4.

[43] Texto original: "(...) are the result of reflection by a scholar after absorbing a great deal of experience of language and languages, and testing the implications and consequences with reference to the intuition of competent or native speakers". (tradução minha)

Dessa forma, a utilização de um córpus para se fazer a verificação de teorias pode possibilitar a descoberta de padrões e variações contemplados pelas mesmas, fazendo-se, assim, distinção daqueles que não o são. Os lingüistas costumam utilizar esse tipo de abordagem, por exemplo, para realizar análise de modelos de língua, pois o córpus poderá indicar onde pequenos ajustes precisam ser feitos em tais modelos, como também pode servir de fonte de informação quantitativa para a pesquisa.

Ainda segundo essa autora, existe um modo de solucionar o problema de se obter dados que não são contemplados pela teoria em estudo, o qual envolve três etapas: (1) isolar os dados problemáticos da teoria para que nunca haja uma confrontação entre eles; (2) utilizar princípios de simplificação e padronização de dados a fim de encontrar uma teoria organizada e clara na desordem dos dados, reduzindo-os a categorias ordenadas; e (3) construir os dados em um sistema de possibilidades abstratas, nas quais a dimensão probabilística adicionada à evidência do córpus não afetaria realmente as escolhas paradigmáticas abstratas disponíveis no sistema em qualquer tempo.

A abordagem baseada em córpus é utilizada nesta dissertação para se realizar uma análise qualitativa dos dados provenientes do córpus em estudo, pertencentes à área de Ciências Farmacêuticas. Para tal análise, optamos pelo modelo de organização retórica de textos científicos proposto por Swales (1990), mais precisamente da seção "Metodologia", a fim de se anotar e observar se esse modelo foi adequado para identificar a organização dessa seção escolhida. Ao final dessa tarefa, constatamos que as categorias puderam ser observadas e anotadas, mas que também foi preciso adicionar mais uma categoria ao modelo proposto. Para mais informações, ver Capítulo 4.

3.3.3 A abordagem baseada em córpus, o ensino de língua estrangeira e o gênero textual

No contexto científico, o conhecimento é construído a partir de leituras, da realização de experimentos científicos, da elaboração de textos e de sua conseqüente publicação. Nesse sentido, é estritamente necessário que os membros dessa comunidade tenham/desenvolvam consciência de como se organiza a prática de investigação científica e mais ainda, de como se dá o processo da construção lingüístico-discursiva dos relatos de pesquisa, uma vez que "a pesquisa não pode ser considerada completa até que esteja disponível à comunidade de pesquisa mais ampla". (Swales, 1990: 94)

Assim, a realização de pesquisas e suas conseqüentes publicações podem ser consideradas essenciais, o que tem provocado um interesse cada vez maior pela consulta a periódicos como fontes de informação e meios de disseminação de pesquisas. La Porte (1998:

4) destaca que no princípio, ou seja, no momento em que os periódicos começaram a serem concebidos, eles eram utilizados exclusivamente para facilitar o intercâmbio de pesquisa entre os cientistas e que, em anos mais recentes, eles passaram a servir também a um outro propósito: o de dar prestígio àqueles trabalhos que possuem muitas publicações. Quanto maior for o número de publicações nos periódicos de maior prestígio, maior o sucesso na obtenção de bolsas, emprego e reputação. Nesse sentido, o artigo científico é visto como o principal meio de veiculação de informação e materialização de pesquisa.

Halliday & Martin (1993:124) defendem a idéia de se ter razões práticas para analisar textos científicos e que dentre as mesmas a mais óbvia seria a educacional, pois existem estudantes que possuem dificuldades lingüísticas ao lerem e/ou escreverem textos científicos em língua inglesa, segundo relatórios de pesquisa (Idem: ibidem). Perceberam, então, a necessidade de que pesquisadores e professores entendam como a linguagem desses textos é organizada, no sentido de proporcionar aos alunos uma visão mais detalhada do material que está sendo lido/produzido. Essa condição é devida ao fato de que há certas características no modo como o sentido e o conteúdo são trabalhados nos textos, as quais podem se tornar problemáticos para os aprendizes, independentemente do seu conhecimento prévio a respeito do assunto abordado. Essas características envolvem questões lexicais (relacionadas às escolhas referentes ao vocabulário), questões discursivas (referentes à composição do texto e suas estruturas retóricas), ou ainda questões referentes à ideologia, às crenças e aos valores que constituem o contexto cultural do texto, ou seja, a área de conhecimento na/para a qual o texto será produzido ou está sendo lido. Conforme argumenta Halliday & Martin (1993:167), para entender como o discurso científico é manifestado no texto, é preciso prestar atenção à maneira como essa linguagem é estruturada, isto é, como cientistas transpõem a linguagem técnica em textos científicos, e quais os recursos de linguagem utilizados por esses escritores para ajustar o discurso contido em seus textos para o de seu público-alvo.

Devido a esse fato, o ensino-aprendizagem de língua estrangeira na comunidade universitária se caracteriza por uma abordagem voltada para a leitura e produção escrita estritamente associadas ao contexto de atuação do aluno de graduação ou pós-graduação. Em outras palavras, o ensino-aprendizagem de língua estrangeira possui uma finalidade ou propósito específico (Jordan, 1997: 5), isto é, o aluno trabalha com as questões, os problemas, as metodologias de sua área e com os gêneros textuais utilizados dentro do seu ambiente científico. Essa abordagem é conhecida como *English for Specific Purposes (*ESP) e estudiosos sobre gêneros como, por exemplo, Swales (1990) e Bhatia (1993) fazem parte dessa corrente, tendo produzido muitos de seus trabalhos voltados para o ensino de ESP.

Podemos dizer que essa abordagem difere do ensino de inglês geral, isto é, sem propósitos específicos, pelos seguintes fatores: 1) visa avaliar e atender a(s) necessidade(s) específica(s) do(s) aluno(s); 2) tem seu conteúdo programático relacionado a determinada atividade ou disciplina; 3) centra o ensino-aprendizagem na linguagem específica empregada pela área ou disciplina escolhida como alvo; 4) prima por uma ou poucas habilidades lingüísticas e 5) não segue uma metodologia pré-estabelecida, ou seja, as necessidades e particularidades de aprendizado do(s) aluno(s) é que guiarão a metodologia e o material a serem utilizados.

É comum interpretar o ESP como abordagem que foca apenas na gramática e no léxico de um dado tipo de texto. Em contrapartida, Skulstad (1999 *apud* Tavares, 2004)[44] defende a idéia de que uma abordagem centrada na análise de gênero deve, antes de tudo, desenvolver no aluno a *consciência do gênero*. Desse modo, o aluno poderá se tornar consciente sobre as maneiras pelas quais os discursos são utilizados/organizados para realizar objetivos específicos, além de serem expostos aos padrões discursivos e às convenções de uma variedade de gênero dentro de sua área específica de atuação. Em outras palavras, de acordo com Santos (1996: 18):

> Levantar uma manifestação textual (oral ou escrita) como um gênero, então, consiste em levantar as características sócio-culturais e lingüísticas que regulam a forma, o conteúdo e as escolhas léxico-gramaticais que o compõem e que são desempenhadas por uma comunidade discursiva específica, identificada e descrita.

Assim, se voltarmos para a questão apontada no início dessa seção, que diz que há interesse crescente no acesso a periódicos, poderíamos apontar que os mesmos poderiam também se constituir enquanto fonte de material/consulta para esse tipo de ensino. E mais ainda, que se os mesmos periódicos já estivessem, por exemplo, com seus componentes esquemáticos, sua estruturação retórica, seus marcadores discursivos destacados, os alunos e professores de ESP teriam um material rico para compor suas atividades e, indo mais além, por que não obter todo esse material organizado no formato de uma ferramenta de auxílio à escrita científica? Esse, portanto, é um dos objetivos de nosso trabalho: possibilitar que um material gerado de linguagem em uso (córpus) e direcionado a propósitos específicos (escrita dentro de um gênero científico em uma dada área do conhecimento) possa ser utilizado em sala de aula por professores, ou até mesmo ser produzido pelos alunos e posteriormente

[44]SKULSTAD , A. S. Genre awareness in ESP teaching: issues and implications. *International Journal of Applied Linguistics*, London, v. 9, n. 2, 1999.

utilizados pelos mesmos, auxiliando, portanto, o ensino-aprendizado de inglês em um contexto específico de uso: a escrita de textos do gênero científico, em uma dada área do conhecimento.

3.4 Concepções sobre o conceito de Gênero

A seguir, serão brevemente apresentadas perspectivas sobre o conceito de gênero presentes na Lingüística de Córpus. Na Seção 3.4 é feita uma introdução aos estudos de gênero que consideramos os de maior visibilidade e relevância para a área e, consequentemente, para nosso trabalho. Para tal, realizamos um percurso de ordem cronológica que abrange os estudos de Aristóteles, Mikhail Bakhtin, Swales, Biber e Marcuschi.

3.4.1 Breve Histórico sobre gênero

O fato de sermos usuários de uma língua e do conhecimento de mundo por ela descrito permite-nos, intuitivamente, fazer distinções das diferentes realizações de textos ou discursos que auxiliam na organização de nossas atividades de comunicação de todos os dias. Essa habilidade do ser humano de diferenciar, agrupar por semelhanças ou denominar o conhecimento é intrínseca e se reflete, segundo Ciapuscio (1994[45] *apud* Possamai, 2004), em diferentes áreas do saber. Nas humanidades, por exemplo, mais especificamente em relação à distinção dos diferentes tipos de textos ou discursos, a obra de Aristóteles - *Arte Retórica e a Arte Poética* - é considerada um dos primórdios desse tipo de reflexão e do surgimento do conceito de gênero, que a define.

Para esse filósofo grego, há três gêneros que determinam três tipos retóricos: o judiciário, o deliberativo e o demonstrativo, os quais são variações de uma comunicação oral, destinada a uma audiência específica e historicamente situada (Aristóteles, 384-322 a.C.). Nessa obra, os tipos retóricos seriam variações de como proceder de acordo com o assunto, o público-alvo e a finalidade e obter, conseqüentemente, o resultado desejado. Para o autor, essa maneira de comunicar objetivando um fim é compreendida como a arte de persuadir, "(...) a faculdade de ver teoricamente o que, em cada caso, pode ser capaz de persuasão. A retórica parece ser capaz de, por assim dizer, no concernente uma dada questão, descobrir o que é próprio para persuadir". (Aristóteles, 384-322 a.C.)

[45] CIAPUSCIO, G.E. Tipos textuales. Buenos Aires: Oficina de Publicaciones; 1994.

Uma aproximação que poderíamos fazer entre o conceito de gênero proposto por Aristóteles e o nosso trabalho é referente ao caráter de persuasão contido na retórica e essa ser uma das características primordiais do artigo científico, um gênero que surgiu no século XVII, muito tempo depois de Aristóteles. O texto de um artigo objetiva, na maioria das vezes, validar uma pesquisa e torná-la conhecida na comunidade científica. E para tal, o pesquisador precisa adaptar o conhecimento científico empírico adquirido para uma produção textual científica que corresponda às expectativas da comunidade acadêmica da qual participa e, só então, possa ter seu trabalho aceito e divulgado conforme desejado. Ou seja, para atingir o seu objetivo de publicação, o pesquisador necessitará levar em consideração o seu público-alvo, conhecer as idiossincrasias existentes nas diferentes áreas do conhecimento, como também os esquemas rígidos de estruturação e de escolha de conteúdos desse gênero textual, a fim de só então poder argumentar com propriedade e poder adquirir as condições adequadas de produção e aceitação do texto produzido.

As reflexões sobre gêneros ganham um diferente ponto de vista a partir do ensaio produzido pelo filósofo russo Mikhail Bakhtin, *O Problema dos Gêneros do Discurso*, contido no livro *A Estética da Criação Verbal*, de 1953. Segundo Todorov, que escreveu a introdução deste livro, esse texto seria o início ou o plano de um livro consagrado aos gêneros do discurso, uma espécie de síntese das reflexões lingüísticas de Bakhtin na década de 20. Algumas questões relativas a gênero tratadas por ele e importantes ao nosso trabalho são apresentadas a seguir.

3.4.2 O conceito de gênero sob a perspectiva de Bakhtin

Do ponto de vista da Lingüística, segundo Eggins e Martin (1997: 236 *apud* Kauffman, 2005)[46], Bakhtin pode ser considerado o responsável pela ampliação do conceito de gênero para além da classificação tradicional herdada de Aristóteles. Para podermos compreender a noção de gênero discursivo por ele proposta, é essencial que se entenda língua como um processo que envolve maneiras múltiplas de realização. Para Bakhtin, o ser humano utiliza a língua em quaisquer esferas de atividades humanas que desempenha e, a partir dos interesses e propósitos específicos contidos em cada uma dessas atividades, realiza os enunciados[47] lingüísticos (escritos ou falados) de diferentes maneiras denominadas *gêneros*.

[46] EGGINS, S.; MARTIN, J. R. Genres and registers of discourse. In: VAN DIJK, T.A. (Ed.). *Discourse as Structure and Process* – Discourse Studies: A Multidisciplinary Introduction, v. 1. London, Thousand Oaks, New Delhi: Sage Publications, 1997.

[47] O(s) enunciado(s) seria(m) a realização da língua, em toda e qualquer esfera das diversas atividades humanas, a partir dos interesses e propósitos específicos existentes em cada uma delas (Bakhtin, 1997).

Segundo a visão bakhtiniana, os enunciados se originam nas diferentes esferas sociais e estão estritamente relacionados aos diferentes tipos de intercâmbios sociais. Assim, para esse autor, as *condições de produção* do discurso é que modelam a existência dos gêneros, de acordo com as funções que se deseja expressar:

> Uma dada função (científica, técnica, ideológica, oficial, cotidiana) e dadas condições específicas para cada esfera de comunicação verbal geram um dado gênero, ou seja, um dado tipo de enunciado, relativamente estável do ponto de vista temático, composicional e estilístico. O enunciado reflete as condições específicas e as finalidades de cada uma dessas esferas, não só por seu conteúdo (temático) e por seu estilo verbal, ou seja, seleção operada nos recursos da língua – recursos lexicais, fraseológicos e gramaticais -, mas também e, sobretudo, por sua construção composicional. (Bakhtin, 1997:279)

Ainda segundo esse autor, cada uma dessas esferas de atividade sócio-discursivas (cotidianas ou especializadas) desenvolve tipos de enunciados relativamente estáveis e diversos, que passam a ser comumente associados a cada uma delas, formando-se, assim, os *gêneros do discurso* (Bakhtin 1997: 301). Ou seja, os gêneros são marcados pela predominância de blocos seqüenciais, que constituem o texto como um todo. Essa constatação de relativa estabilidade é interessante para o nosso estudo, pois confirma por um viés teórico o pressuposto que queremos demonstrar na prática: que existem expressões textuais no artigo científico que são freqüentes e sistemáticas e que poderiam, portanto, ser tratadas do ponto de vista computacional. Esse tipo de tratamento favoreceria a colaboração das mesmas, por exemplo, enquanto recursos lingüísticos em ferramentas de suporte à escrita: ora como base de casos de estruturas retóricas, componentes da estrutura esquemática e marcadores discursivos, ora como córpus de treinamento de categorizadores de componentes esquemáticos (para mais detalhes ver Capítulo 4 – Etapa E2).

Um outro ponto do texto de Bakhtin interessante para nosso estudo afirma que o domínio de um dado gênero discursivo se dá pela vivência das situações de comunicação e também pelo contato com os diferentes gêneros que surgem na vida cotidiana. Segundo o autor, esse tipo de domínio discursivo é determinado pela capacidade do indivíduo de prever as regras de conduta e pela seleção vocabular e de estrutura de composição que estão sendo utilizadas no contexto de produção da comunidade da qual o texto (oral ou escrito) a ser produzido deverá fazer parte. Em outras palavras, esse tipo de conhecimento permite a um indivíduo prever as relações de sentido e de comportamento necessárias. E ainda, quanto mais competente e experiente for o indivíduo, isto é, quanto mais conhecimento ele possuir sobre um gênero, mais proficiente ele será na diferenciação de determinados gêneros e mais

facilidade terá para reconhecer as estruturas, por exemplo, lexicais, retóricas, sintáticas e de sentido que o compõe. Essa parte da teoria de Bakhtin é importante para o nosso trabalho, uma vez que ao conhecer o funcionamento de um texto científico, o pesquisador-escritor poderá adaptar sua produção textual ao contexto do qual faz parte. Em outras palavras, conhecendo as idiossincrasias existentes nas diferentes áreas do conhecimento como também os esquemas relativamente estáveis de estruturação e de escolha de conteúdos desse tipo de gênero, o escritor poderá adquirir condições adequadas para a produção de textos que correspondam às expectativas da comunidade acadêmica.

A Estilística é outro fato observado por Bakhtin, e que também é interessante para nosso estudo. Para ele, o estilo está unido ao enunciado e aos gêneros (formas típicas do enunciado), uma vez que o enunciado é individual, ele pode refletir essa individualidade, sendo alguns gêneros mais propícios para essa reflexão do que outros. Porém, é o estilo lingüístico que está indissociavelmente ligado ao gênero:

> (...) o estilo lingüístico ou funcional nada mais é senão o estilo de um gênero peculiar a uma dada esfera da atividade e da comunicação humana. Cada esfera conhece seus gêneros, apropriados à sua especificidade, aos quais correspondem determinados estilos. (...) o estilo é indissociavelmente vinculado a unidades temáticas determinadas e, o que é particularmente importante, a unidades composicionais: tipo de estruturação e conclusão de um todo, tipo de relação entre o locutor e os outros parceiros da comunicação verbal. (Bakhtin, 1997: 284)

Relacionando com nossa pesquisa, podemos dizer que as expressões formulaicas, os marcadores discursivos e as os componentes da estrutura esquemática e as estratégias retóricas fazem parte da especificidade do gênero discursivo artigo científico e do estilo que essa esfera de atuação (no caso, a grande comunidade científica internacional e, em um segundo plano, não menos importante, a comunidade internacional de uma dada área específica) condiciona ou à qual os autores estão condicionados.

Importante ainda é ressaltar que a noção de gêneros do discurso nesse texto de Bakhtin se presta a uma reflexão muito mais ampla do que as questões até então apresentadas neste trabalho. *Os gêneros do Discurso* é elaborado por uma teoria sobre linguagem (baseada nas noções de enunciado e gênero) completa e inovadora, ao trazer à luz uma reflexão diferente sobre gêneros, principalmente com as noções de esfera de atividade e relativa estabilidade dos gêneros e, claro, também com a consideração de um interlocutor como requisito fundamental na atividade de comunicação (Bakhtin, 1997:324).

3.4.3 O conceito de gênero sob a perspectiva de Swales

John M. Swales é um pesquisador que se destaca, há mais de dezesseis anos, na tradição norte-americana de Lingüística Aplicada direcionada aos estudos de gêneros textuais. Embora haja algumas ressalvas quanto ao caráter generalizador de seu modelo formal/funcional de linguagem acadêmica, ainda assim, permanece como fonte confiável para a realização de estudos atuais no Ensino de Inglês com Propósitos Específicos (*ESP*). Podemos dizer que o principal tributo que se deve a esse pesquisador é o abandono das extensas discussões gramaticais que não davam conta de ensinar a modalidade escrita de língua inglesa para estrangeiros e a busca por uma análise textual mais global (Augusto-Navarro: 2002), ou seja, ele partiu de investigações baseadas em textos de pesquisas científicas em inglês, para compor seus estudos sobre as estruturas de gêneros subjacentes aos textos científicos de diferentes áreas do conhecimento.

Antes de se apresentar o conceito de gênero sob a óptica de Swales, temos necessariamente que considerar sua origem: a comunidade discursiva da qual o gênero provém. O conceito de comunidade discursiva adotado por Swales é o mesmo definido por Herzberg, com a ressalva de ser utilizado enquanto o centro de um conjunto de idéias que endossam os seus pressupostos de ensino de língua:

> (...) o discurso opera dentro de convenções definidas por comunidades, sejam elas disciplinas acadêmicas ou grupos sociais. (...) O uso da língua em um grupo é uma forma de comportamento social, o discurso ali produzido é um meio de manter e expandir o conhecimento do grupo e de introduzir novos membros no grupo, sendo esse discurso epistêmico ou constitutivo do conhecimento do grupo.[48] (Herzberg, 1986: 1[49] apud Swales, 1990:21)

Assim, segundo Swales (1990), para um dado grupo de indivíduos se constituir enquanto uma comunidade discursiva deve possuir as seguintes características:

(1) Apresentar objetivos comuns: que podem ser públicos, ou seja, de conhecimento de todos os membros, como o contrato de uma escola, ou implícitos. Entretanto, não significa dizer que todos os membros resolveram se afiliar porque possuem os mesmos objetivos, pois as pessoas podem se associar por razões diversas;

[48] Texto original: "(...) discourse operates within conventions defined by communities, be the academic disciplines or social groups. (…) language use in a group is a form of social behavior, that discourse is a means of maintaining and extending the group's knowledge and of initiating new members into the group, and that discourse is epistemic or constitutive of the group's knowledge".

[49] HERZBERG, B. The politics of discourse communities. Paper presented at the CCC Convention, New Orleans, LA, March, 1986.

(2) Possuir mecanismos de intercomunicação entre seus membros: esses mecanismos de intercomunicação (reuniões, atas, e-mails, etc.) podem variar conforme a comunidade discursiva, e parte-se do pressuposto que seu alcance se estenda a toda a comunidade;

(3) Utilizar seus mecanismos de participação para realizar trocas de informações: ou seja, os membros de uma comunidade utilizam os mecanismos de participação existentes em seu contexto discursivo com o objetivo principal de fornecer informações e comentários avaliativos;

(4) Possuir gêneros específicos para a realização da intercomunicação em uma dada comunidade: uma comunidade discursiva possui um ou mais gêneros textuais em acordo com seus objetivos comunicativos, os quais são responsáveis pela comunicação e transmissão de informações entre seus membros;

(5) Possuir um léxico altamente especializado: além de possuir gêneros textuais específicos e em acordo com seus propósitos comunicativos, uma comunidade discursiva também tem um léxico característico, preferencial, que pode incluir jargões, siglas, termos técnicos, etc. Esse léxico é reconhecido pelos membros, como também é o responsável pela efetiva comunicação entre eles e;

(6) Possuir um número alto e oscilante de membros especialistas na área: ter um número de membros, com certo grau de conhecimento da área e proficiência discursiva, que oscila. Em outras palavras, assim como em quaisquer comunidades, novos membros são inseridos e outros são afastados por variados motivos. Sendo assim, a sobrevivência de uma comunidade discursiva depende de um número constante de membros, sejam eles novatos ou experientes. Motta-Roth endossa essa última característica ao afirmar:

> Para se engajar em uma determinada comunidade, um indivíduo aprende os gêneros e as convenções normalmente integradas pelos membros mais experientes do grupo, através de iniciação profissional, em um processo denominado aculturação. (Motta-Roth, 1995: 47-8)

No entanto, essa primeira definição de comunidade discursiva levantou várias críticas e questionamentos, que motivaram Swales (1992; 1993; 1998) a rever sua idéia de comunidade discursiva, mostrando as limitações desse conceito e a amplitude de sua nova visão. O autor cria uma definição mais precisa de comunidade que se adapta à realidade atual, pois não havia considerado, em sua primeira versão, fatores como conflitos que podem existir dentro das comunidades. Os exemplos de comunidades discursivas apresentadas não eram adequados para representar a realidade, pois mostravam comunidades atípicas, validavam grupos já formados e não ofereciam a possibilidade de analisar seus processos de formação.

Swales faz uma redefinição desses seis critérios apresentados acima, justificando que os anteriores manifestam um caráter reducionista, utópico e estático do conceito por eles abordado:

(1) Apresentar objetivos comuns: a comunidade discursiva aceita os objetivos, formula-os ou os estabelece. Esses objetivos podem ser consensuais, mas também podem ser distintos e se relacionar;

(2) Possuir mecanismos de intercomunicação entre seus membros: não houve alterações, pois segundo Swales (1990), podem variar de acordo com a comunidade;

(3) Utilizar seus mecanismos de participação para realizar trocas de informações: acrescenta que uma comunidade discursiva utiliza mecanismos de participação para diferentes propósitos e não apenas para informação e *feedback*.

(4) Possuir gêneros específicos para a realização da intercomunicação em uma dada comunidade: em vez de utilizar um ou mais gêneros para alcançar seus objetivos, uma comunidade discursiva utiliza uma seleção crescente de gêneros no alcance dos mesmos.

(5) Também possuir um léxico altamente especializado: uma comunidade discursiva adquire e continua sempre buscando uma terminologia específica.

(6) Possuir um número alto e oscilante de membros especialistas na área: uma comunidade discursiva tem uma estrutura hierárquica explícita ou implícita que orienta os processos de admissão e de progresso dentro dela.

Nota-se que o conceito de comunidade discursiva foi ampliado, abarcando mais elementos e ficando mais flexível, no sentido de que há assim uma tentativa de "(...) representar um mundo mais complexo e um tanto obscuro" (Swales: 1992). Mas, afinal, o que esse autor entende por gênero?

Para formular sua própria noção de gênero, Swales considerou as noções de gêneros existentes nas áreas do folclore, da literatura, da lingüística e da retórica, a qual deseja que seja aplicada apenas aos propósitos de ensino e aprendizado de língua estrangeira. Ele diz que ainda que a definição que propõe não seja totalmente adequada, ela representou um avanço nas próprias formulações anteriores que havia feito:

> Um gênero compreende uma classe de eventos comunicativos, cujos membros compartilham um conjunto de propósitos comunicativos. Esses propósitos são reconhecidos pelos membros da comunidade discursiva que trabalha com eles e, portanto, constituem a lógica subjacente aos gêneros. Essa lógica modela a estrutura esquemática do discurso influenciando e restringindo a escolha do conteúdo e o estilo. Além do propósito, os exemplares de um gênero exibem vários padrões de similaridade em termos de estrutura, estilo conteúdo e público-alvo. (Swales, 1990: 58)[50]

Nesses termos, o gênero pode ser entendido como um modo de interação de uma dada comunidade discursiva, que possui propósito(s) comunicativo(s) específico(s), os quais determinam os componentes da estrutura esquemática do discurso, restringindo, portanto, as escolhas de conteúdo e estilo. Dessa maneira, textos de um mesmo gênero possuem em comum o propósito comunicativo, a estrutura, o estilo e o público-alvo.

Porém, um evento comunicativo envolve mais do que a comunicação em si, pois abrange: (a) a linguagem, (b) as funções que essa linguagem desempenhará segundo seus usuários e, também, (c) o modo como é produzida e/ou recebida por seus usuários. Em suma, podemos dizer que a comunicação para Swales se daria por meio do gênero, que possui as características do grupo social (contexto social) do qual se origina.

Conforme Santos sintetiza:

> Para Swales, o conceito de gênero privilegia o caráter/propósito comunicativo de uma situação, suas convenções e regras lingüísticas e discursivas compartilhadas pela comunidade discursiva que convive, atua e interage em uma dada situação, dominando gêneros do discurso articulado e intencionado (a quem se destina: público-alvo) por ela mesma. Uma vez configurada as expectativas, uma manifestação genérica pode ser considerada como prototípica pela comunidade geradora. (Santos, 1996:18)

Bhatia (1993:13) traduz a visão de Swales sobre gênero como:

> um evento comunicativo reconhecido e caracterizado por um conjunto de propósito (s) comunicativo (s) identificados e mutuamente compreendidos pelos membros de uma comunidade profissional ou acadêmica na qual ele regularmente ocorre.

[50]Original: "A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. Communicative purpose is both a privileged criterion and one that operates to keep the scope of a genre as here conceived narrowly focused on comparable rhetorical action. In addition to purpose, exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content and intended audience". (Swales, 1990:58)

3.4.4 O conceito de gênero/registro sob a perspectiva de Biber

Biber (1988) utiliza indistintamente os termos **registro** e **gênero** ao definir uma variedade lingüística geral ou específica, como, por exemplo, discurso científico e cartas pessoais, preferindo o emprego da primeira forma em seus trabalhos mais recentes (Biber, 1998; 1999). Para esse autor, um gênero/registro é definido por variáveis situacionais, isto é, não lingüísticas e seus rótulos são empregados corriqueiramente pelos falantes nativos da língua. Segundo Berber-Sardinha (2000b), Biber não faz uma diferenciação formal entre **registro** e **gênero**, ao contrário do que ocorre em outras áreas (e.g. lingüística sistêmico-funcional).

Uma distinção importante que é feita em Biber (1994) faz referência aos termos gênero/registro e **tipo textual** (*text type*). Enquanto **gênero/registro** são definidos por categorias situacionais, o **tipo textual** é definido exclusivamente com base em critérios lingüísticos (Biber, 1994:380[51] *apud* Tognini-Bonelli, 2001:60), chegando-se à descrição de tipos de texto somente em estágios avançados da Análise Multidimensional, quando já se descreveram as dimensões e se mapearam os registros participantes em cada uma.

A Análise Multi-traço e Multidimensional de Variação de Registro (*Multi-feature Multi-dimensional Analysis of Register Variation*), ou simplesmente Análise Multidimensional foi criada por Biber e se propõe a descrever/caracterizar automaticamente uma língua ou um conjunto de tipos textuais existentes em grandes córpus. O nome da abordagem deriva do conceito de **dimensão** de variação. Uma dimensão é um conjunto de traços que subjazem a um córpus, que pode se consistir em uma seleção de textos, de um conjunto de gêneros ou até mesmo de amostras relativas a um idioma inteiro. Conforme é apontado por Berber-Sardinha (2000b), essa abordagem possui várias características, que, em conjunto, a distinguem de outros sistemas analíticos de descrição:

1. Baseia-se em córpus, isto é, ela pretende descrever um grande número de textos autênticos;

2. É em sua essência computacional, pois faz uso de ferramentas automáticas e semi-automáticas para etiquetar as características de interesse;

3. Proporciona a descrição de conjuntos de textos ou registros, em vez de textos individuais;

[51] BIBER, D. Representativeness in corpus design. In ZAMPOLLI, A., CALZOLARI, N., PALMER, N. (eds). *Current Issues in Computational Linguistics:* In Honour of Don Walker. Linguistica Computazionale IX. Pisa e Dordrecht: Giardini e Kluwer Academic Publishers, 1994.

4. Possui um caráter comparativo, porque possibilita o contraste entre os textos ou registros;

5. Como diz seu rótulo, ela é multidimensional, isto é, reconhece a variação entre textos e registros por meio de múltiplos parâmetros;

6. Utiliza-se de um aparato quantitativo de descrição, permitindo a especificação da co-ocorrência dos traços lingüísticos de modo preciso. Porém, essa abordagem não descarta a utilidade de técnicas qualitativas de interpretação, pois as dimensões são etiquetadas conforme a interpretação qualitativa dos fatores;

7. Combina níveis macro e micro de análises, uma vez que a micro-descrição dos traços de cada texto permite a indução dos macro-agrupamentos textuais ou genéricos;

8. Possui caráter cumulativo, o que possibilita empreender uma análise de larga escala em um corpus fazendo-se descrições individuais ao longo do tempo, combinando-se posteriormente as análises para fins comparativos.

9. É flexível, podendo acomodar diversos tipos de traços lingüísticos. Tradicionalmente, tem-se utilizado características lexicais e gramaticais (Biber, 1988[52] *apud* Berber-Sardinha, 2000b), entretanto é possível incluir características de cunho mais discursivo (Pacheco de Oliveira, 1997[53] *apud* Berber-Sardinha, 2000b) e funcionais (Shimazumi: 1998[54] *apud* Berber-Sardinha, 2000b). Além disso, as próprias dimensões não são definitivas, podendo ser modificadas com a inclusão de novas características lingüísticas.

3.4.5 O conceito de gênero sob a perspectiva de Marcuschi

A visão de gênero defendida por Marcuschi decorre da "noção de língua enquanto atividade social, histórica e cognitiva. Privilegia a natureza funcional e interativa e não o aspecto formal e estrutural da língua" (Marcuschi, 2002:22). Assim, segundo esse autor, os "gêneros caracterizam-se muito mais por suas funções comunicativas, cognitivas e institucionais do que por suas peculiaridades lingüísticas e textuais". (Marcuschi, 2000: 19)

Para Marcuschi, que possui as visões sobre tipo e gênero textual semelhantes às apresentadas por Biber (1988) e Swales (1990), essas duas noções podem ser definidas como:

52 BIBER, D. Variation across Speech and Writing. Cambridge: Cambridge University Press, 1988.
53 PACHECO DE OLIVEIRA, L. Variação intercultural na escrita: contrastes multidimensionais em inglês e português. São Paulo, 1997. Tese (Doutorado em Lingüística Aplicada e Ensino de Línguas) – LAEL, PUC-SP, 1997.
54 SHIMAZUMI, M. Investigating EFL writing: A Multidimensional analysis. In Convenção BRAZ-TESOL, 6, Recife, 13-16 de julho, comunicação oral, 1998.

> Usamos a expressão tipo textual para designar uma espécie de seqüência teoricamente definida pela natureza lingüística de sua composição (aspectos lexicais, sintáticos, tempos verbais, relações lógicas). Em geral, os tipos textuais abrangem cerca de meia dúzia de categorias conhecidas como: narração, argumentação, exposição, descrição, injunção. (Marcuschi: 2002:22)

Entre essas categorias apresentadas podemos reconhecer as cinco bases temáticas textuais típicas que dão origem aos tipos textuais, que foram propostas por Werlich (1976[55] *apud* Baldo, 2004): base temática descritiva, base temática narrativa, base temática expositiva, base temática argumentativa e base temática instrutiva.

Sobre a noção de gênero textual, Marcuschi diz:

> Usamos a expressão gênero textual como uma noção propositalmente vaga para referir os textos materializados que encontramos em nossa vida diária e que apresentam características sócio-comunicativas definidas por conteúdos, propriedades funcionais, estilo e composição característica. (...) Alguns exemplos de gêneros textuais seriam: telefonema, sermão, carta comercial, carta pessoal, romance, bilhete, etc... (Marcuschi: 2002:22)

Assim, notamos que, na visão desse autor, enquanto os gêneros textuais são teoricamente ilimitados, os tipos textuais constituem um conjunto pequeno e fechado. Nesse mesmo texto de 2002, também é apresentada a noção de domínio discursivo, uma vez que é esse domínio que determina os discursos produzidos em um dado grupo:

> Usamos a expressão domínio discursivo para designar uma esfera ou instância de produção discursiva ou de atividade humana. Esses domínios não são textos nem discursos, mas propiciam o surgimento de discursos bastante específicos. Do ponto de vista dos domínios, falamos em discurso jurídico, discurso jornalístico, discurso religioso, etc., já que as atividades jurídica, jornalística ou religiosa não abrangem um gênero em particular, mas dão origem a vários deles. Constituem práticas discursivas dentro das quais podemos identificar um conjunto de gêneros textuais que, às vezes, lhe são próprios (em certos casos exclusivos) como práticas ou rotinas comunicativas institucionalizadas. (Marcuschi, 2002: 23-24)

Marcuschi ainda ressalta a importância de não se confundir a noção de texto com a de discurso como se fossem sinônimas nas definições teóricas que faz. Ele diz que apesar das muitas discussões existentes a esse respeito, pode-se dizer que "texto é uma entidade concreta realizada materialmente e corporificada em algum gênero textual" (Marcuschi, 2002:24); em contrapartida tem-se discurso enquanto "aquilo que um texto produz ao se manifestar em alguma instância discursiva. Assim, o discurso se realiza nos textos" (idem:ibidem).

55 WERLICH, E. A text grammar of English. Heidelberg: Quelle and Meyer, 1976.

A seção a seguir traz, à luz da reflexão sobre gêneros feita até aqui, o artigo científico, ainda que sucintamente e sem maiores pretensões.

3.5.2 O Artigo Científico

Para Swales (1990:93), o artigo científico é definido como:

> (...) um texto escrito (embora freqüentemente contenha elementos não-verbais) geralmente limitado a alguns milhares de palavras, no qual é relatada alguma investigação realizada por um autor ou autores. Além disso, em um artigo científico, em geral, o pesquisador relaciona suas descobertas com as de outros, podendo também examinar tópicos teóricos e/ou metodológicos.[56]

Nesse mesmo viés de opinião, Berkenkotter & Huckin (1995:27[57] *apud* Kanoksilapatham, 2005) dizem que o artigo científico é fruto da atividade de pesquisa da comunidade de onde foi gerado, e que apresenta como características a alusão a outras pesquisas ou autores que compartilham ou não do mesmo assunto tratado, uma apresentação de objetivos posteriores à identificação de um problema apontado, a citação de outras pesquisas que possam corroborar na apresentação e a generalização dos resultados obtidos. Nesse sentido, o artigo pode ser considerado como meio responsável pela divulgação de um estudo, reunindo etapas que se estendem desde a contextualização de uma pesquisa até a conclusão do estudo realizado.

Segundo pesquisa realizada por Motta-Roth (1995), entre os gêneros discursivos mais utilizados por pesquisadores na leitura e publicação científicas estão os capítulos de livros e artigos de revistas acadêmicas. Nesse contexto, podemos considerar o artigo científico um dos gêneros mais utilizados no ambiente científico como forma de acesso e de produção de conhecimento científico. Além de muito utilizado, é também um dos textos mais antigos da comunidade acadêmica, pois existe desde o ano de 1665, ano em que apareceu a primeira revista acadêmica, *The Philosophical Transactions of the Royal Society* (Swales, 1990:110; Berkenkotter & Huckin, 1995: 27).

---

[56] "(...) a written text (although often containing non-verbal elements), usually limited to a few thousand words, that reports on some investigtaion carried out by its author or authors. In addition, the Research Article will usually relate the findings within it to those of others, and may also examine issues of theory and/or methodology".

[57] BERKENKOTTER, C. & HUCKIN, T. N. *Genre knowledge in disciplinary communication: cognition/culture/power*. Hillsdale: Lawrence Erlbaum Associates, Publishers, 1995.

De acordo com Ard (1983[58], *apud* Swales, 1990), o artigo científico teve sua origem nas cartas informativas enviadas para essa revista acadêmica, que os cientistas escreviam e trocavam entre si. No momento em que esse periódico começou a proporcionar um local para discussões, os textos acabaram refletindo as novas situações retóricas diferentes da escrita na forma de carta, dando origem, assim, ao artigo científico que gradualmente se tornou distinto das cartas trocadas entre os pesquisadores das quais se originou. Entre os cientistas que colaboraram para estruturar os primeiros artigos científicos estão Robert Boyle e seus companheiros. De acordo com Shapin (1984[59] *apud* Swales, 1990:111), "através da experiência com o fato real, Boyle e seus colegas procuraram transformar reivindicações e especulações em um tipo de conhecimento que fosse mais amplamente aceito". Para tanto, Boyle desenvolveu estratégias retóricas e estilísticas, as quais consistiam, por exemplo, em usar testemunhas para provar que as experiências realmente foram realizadas, e mostrar ilustrações do aparato em questão na ocasião de sua publicação.

Sobre esse assunto, Swales menciona que não é que as notas tomadas no laboratório não pudessem ser publicadas de maneira linear, nem que a primeira versão de um artigo seria totalmente impublicável, mas o que acontece é que, na construção de um artigo, há um processo de:

> (...) crítica técnica e controle social operando tanto no ambiente particular de pesquisa como em um outro mundo imaginado sobre o que os outros cientistas irão pensar (Swales, 1990: 120); pois o artigo publicado é um híbrido com multi-níveis, co-produzido pelos autores e por membros da audiência para o qual é direcionado. (Knorr-Cetina 1981: 106[60] apud Swales, 1990)

Podemos perceber, assim, que entre a pesquisa em si e a escrita de um artigo existem muitos fatores operando, os quais podem torná-la difícil e complicada até mesmo para membros experientes das comunidades científicas.

Tal situação tem suscitado investigações sobre o processo e o produto envolvidos no processo de escrita científica, as quais têm gerado importantes revelações. Se considerarmos apenas os estudos que investigam o produto, ou seja, o texto acabado, obteremos ainda duas correntes de investigação: a primeira com pesquisas centradas no estudo de aspectos gramaticais e estilísticos do discurso científico e a segunda com foco de estudo na organização estrutural dos textos científicos (Ozturk, 2006).

---

[58] ARD, J. The role of the author in science discourse. Paper given at the annual *American Applied Linguistics Meeting*, Minneapolis, Minn, December, 1983.
[59] SHAPIN, S. Pump and circumstance: Robert Boyle's literary technology. *Social Studies of Science*, v. 14, p. 481-520, 1984.
[60] KNORR-CETINA, K.D. *The manufacture of knowledge.* Oxford: Pergamon, 1981.

Entre os fenômenos investigados pela primeira vertente, podem ser citadas pesquisas sobre tempos e aspectos verbais (Hinkel, 2004), o uso e as funções de adjetivos (Soler, 2002), os substantivos (Flowerdew, 2003) e assim por diante. Por sua vez, entre os trabalhos da segunda, pode ser citada a investigação dos componentes da estruturação esquemática das diferentes seções que constituem um texto científico, como a seção "Resumo" (Samraj, 2005; Biasi-Rodrigues, 1998), "Resultados" (Brett, 1994), "Discussões" (Silva, 1999), "Conclusões" (Yang & Allison, 2003), "Metodologia" (Huckin & Olsen, 1991; Oliveira, 2003) e "Introdução" (Swales:1990; Aluísio: 1995; Motta-Roth:1995).

Importante lembrar que não é por ser um texto condicionado por muitos padrões que o artigo científico deixa de ser rico e ter caráter composto por muitas particularidades. Mauranen, por exemplo, dá suporte a essa idéia afirmando que:

> (...) assim como faz sentido falarmos sobre a ciência em geral, também faz sentido falar sobre o gênero da ciência e da comunidade acadêmica como objetos culturais. Assim, podemos falar, por exemplo, do 'artigo científico' como um gênero no mundo da ciência, apesar do fato de disciplinas particulares diferirem de alguma maneira nas suas realizações convencionais. Se insistíssemos que a comunidade de pesquisa de cada disciplina tem seus próprios gêneros, perderíamos uma importante generalização no que diz respeito à atividade científica. (Mauranen,1993: 5[61] apud Mirahayuni, 2002)

Motta-Roth (1999:119-28) a esse respeito, diz que por meio de um texto científico pode-se perceber, por exemplo, as seguintes particularidades/habilidades de um pesquisador quanto a: 1. seleção das referências bibliográficas relevantes ao assunto; 2. reflexão sobre estudos anteriores na área (contextualização); 3. delimitação de um problema ainda não totalmente estudado na área; 4. elaboração de uma abordagem para o exame desse problema; 5. delimitação e análise de um conjunto de dados representativo do universo sobre o qual se quer alcançar generalizações; 6. apresentação e discutição dos resultados da análise dos dados; 7. conclusão, elaborando-se generalizações a partir desses resultados, conectando-as aos estudos prévios dentro da área de conhecimento em questão.

Halliday & Martin (1993:124) sustentam que há razões práticas para analisar textos científicos e que a mais óbvia dessas razões é a educacional. Alguns estudantes teriam dificuldades lingüísticas ao lerem textos científicos em língua inglesa, segundo relatórios de pesquisa (Idem:ibidem). Percebe-se então a necessidade de que pesquisadores e professores dessa área entendam como a linguagem desses textos é organizada, no sentido de

[61] MAURANEN, A. Cultural Differences in Academic Rhetoric: A Textlinguistic Study. Peter Lang, Frankfurt, 1993.

proporcionar aos alunos uma visão mais detalhada do material que está sendo lido. Essa condição deve-se ao fato de que há certas características na maneira como o sentido é organizado e trabalhado nos textos, que podem se tornar problemáticas para os aprendizes, independentemente do seu conhecimento prévio a respeito do assunto abordado. Essas características envolvem questões lexicais (relacionadas aos significados construídos entre as sentenças), questões discursivas (referentes à composição do texto e suas estruturas retóricas), ou ainda questões referentes à ideologia, às crenças e aos valores que constituem o contexto cultural do texto (Idem:ibidem).

Em suma, podemos dizer que a relevância desses tipos de estudo que investigam a organização/composição de um texto científico está no fato de possibilitarem a identificação das peculiaridades discursivas existentes nas diferentes comunidades científicas. Assim sendo, podem contribuir para que pesquisadores em contato com esse tipo de informação atentem para os paradigmas a serem seguidos em suas comunidades, cometendo-se, assim, menos inadequações ao escrever.

A seguir, serão apresentadas essas peculiaridades discursivas referentes à organização/composição de textos científicos, em geral, presentes na maioria dos artigos científicos das diferentes áreas do conhecimento.

3.5.1 Estruturação de Artigos Científicos

O reconhecimento das dificuldades comumente enfrentadas por membros da comunidade acadêmica menos experientes quanto à redação de artigos científicos tem guiado os estudos de pesquisadores sobre descrição e explicitação da organização retórica dos gêneros acadêmicos (Swales: 1990; Nwogu: 1990). Essas pesquisas partem do pressuposto de que para haver uma boa compreensão e produção de um texto do gênero acadêmico, faz-se necessário o conhecimento do padrão rigoroso de estruturação de um texto escrito para essa comunidade, assim como, das características individuais existentes em cada área de pesquisa. Além disso, é importante também que alunos de graduação e pós-graduação possam publicar textos científicos (resumos, resenhas, artigos, livros, dissertações e teses) a fim de produzirem não só conhecimento relevante e originar adequadamente novas formulações, mas também produzir significado sobre um objeto de estudo em uma determinada área e não apenas repetir conceitos pré-estabelecidos (Motta-Roth,1998: 106).

Para que isso ocorra, é interessante que esses membros da comunidade acadêmica adquiram consciência de como um texto científico se organiza em uma língua estrangeira.

Esse conhecimento parece ser possível a partir da produção de uma abordagem, que se proponha a demonstrar como o texto se articula lingüística e discursivamente nos contextos acadêmicos. Isso implica dizer, que professores e alunos precisam tomar consciência das convenções disciplinares a fim de obterem um desempenho mais eficaz na leitura e escrita de textos científicos. Conforme Motta-Roth aponta:

> Ao considerarmos a relação entre conhecimento, linguagem e contexto acadêmico é preciso ter em mente a natureza heterogênea desse universo acadêmico e o fato de a linguagem se articular em tipos de textos associados a atividades humanas que ocorrem em contextos recorrentes. (Motta-Roth 2000: 4).

Conforme já mencionado no início deste capítulo, Swales é tradicionalmente citado quando se trata de estudos de gênero e, em particular, de um deles, o artigo científico. Seu foco e motivação para a pesquisa são bastante pedagógicos, conforme ele mesmo justifica: "O principal objetivo (...) é oferecer uma abordagem para o ensino do inglês científico" (Swales, 1990:1). Segundo esse autor, as investigações sobre a estrutura de artigos científicos iniciaram com os modelos *problem-solution structure* de Stanley (1984[62] *apud* Swales, 1990), *Introduction-Method-Result-Discussion (IMRD)* de Bruce (1983[63] *apud* Swales, 1990), e o modelo *Dogma-Dissonance-Crisis-Search-New* de Hutchin (1977[64] *apud* Swales, 1990), que revisa o modelo de Kinneavy (1971[65] *apud* Swales, 1990), *Dogma-Dissonance-Crisis-Search-New*. Outra tentativa de modelagem dessa estrutura de organização ou formação é de autoria de Hill *et al* (1982)[66]. Nessa é feita uma analogia entre a estrutura de um artigo científico e a estrutura de uma ampulheta:

[62] STANLEY, R.M. The recognition of macrostructure: a pilot study. *Reading in a Foreign Language*, v. 2, p. 156-168, 1984.
[63] BRUCE, N.J. Rhetorical constrains on information structure in medical research report writing. *Paper presented at the ESP in the Arab World Conference*, University of Aston, UK, August, 1983.
[64] HUTCHIN, J. On the structure of scientific texts. In *UEA Papers in Linguistics*, v.5, p.18-39, UK: University of East Anglia, 1977.
[65] KINNEAVY, J. L. *A theory of discourse: the aims of discourse*. Englewood Cliffs, NJ: Prentice-Hall International, 1971.
[66] HILL, S.S., SOPPELSA, B.F. & WEST, G.K. Teaching ESL students to read and write experimental research papers. *TESOL Quarterly*, 16(3), p. 333-347, 1982.

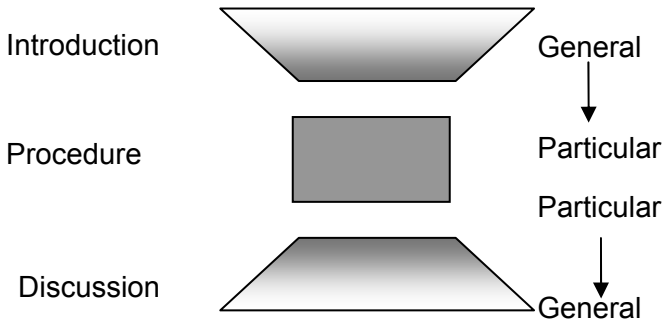


Figura 3.3: Organização geral de um artigo científico (Hill *et al*, 1982[67] *apud* Swales, 1990)

Segundo essa modelagem proposta por Hill *et al* (1982), a estrutura de um artigo científico é constituída por três partes principais: Introdução, Procedimentos e Discussão. Essas, por sua vez, são divididas quanto ao fluxo de informação presente em cada uma delas, que vai do geral-particular para o particular-geral.

Esse tipo de divisão do artigo científico também reflete a organização da pesquisa, ou seja, os passos seguidos para sua realização. Reflete também a possibilidade de identificação mais fácil e precisa dos pontos de interesse do leitor, que não necessariamente obedecem à cronologia de desenvolvimento da pesquisa relatada.

Esse mesmo esquema de organização inspirado em ampulheta também é apresentado por Weissberg & Buker (1990). Nesse esquema, a idéia de organização geral das partes principais de um artigo se repetem de maneira mais detalhada, como por exemplo, a inclusão do componente Resumo, não citado no primeiro modelo. Segundo Weissberg & Buker (1990), o Resumo é apresentado em separado da estrutura global de um artigo por conter informações gerais e específicas de todo o texto, podendo, portanto, ser visto como um texto autocontido.

[67] HILL, S.S.; SOPPELSA, B.F.; WEST, G.K. Teaching ESL students to read and write experimental research papers. *TESOL Quarterly*, v.16, pp. 333-347, 1982.

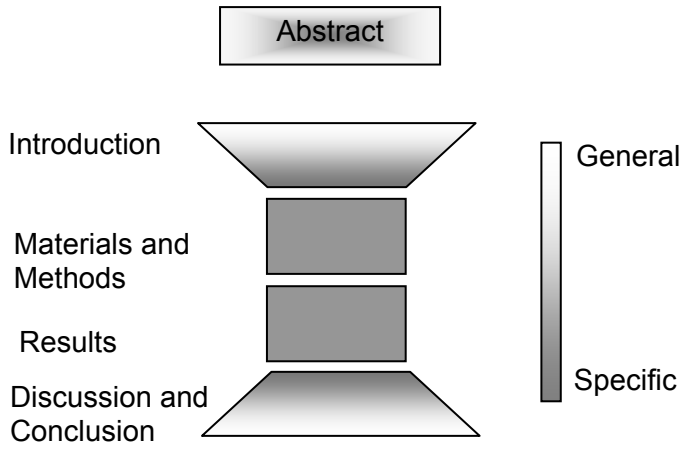


Figura 3.4. Movimento Geral-Específico-Geral da estrutura global de um artigo científico (Weissberg & Buker:1990).

Ao observarmos os dois esquemas de organização para artigos científicos, podemos dizer que, apesar das diferenças de detalhes dos componentes presentes em cada modelo, o objetivo de ambos é o mesmo: apresentar o texto a partir do contexto de pesquisa em que se encontra inserido. Para isso, cada um dos componentes contidos em ambas as estruturas apresentadas possuem propósitos ou papéis retóricos bem definidos. Acrescentamos também que, embora nem todos os artigos científicos sejam, na prática, assim, formalmente divididos, eles contemplam em seu desenvolvimento, aspectos semanticamente relacionados a essa divisão. Aspectos, cujo significado tende a remeter para esse tipo de organização do fluxo de informação geral-específico-geral ilustrado pelos modelos de diagramas em forma de ampulheta. Essa forma de transição da informação entre os componentes é iniciada a partir de um campo ou contexto mais geral do experimento descrito na seção "Introdução". Depois, para um mais específico através da descrição de uma inadequação/lacuna na pesquisa prévia, que originou o presente experimento relatado. As seções "Metodologia" e "Resultados" consistem em uma trajetória limitada e particular do experimento. Na seção "Discussão", a partir das descobertas específicas do estudo realizado, se deduz as implicações ou generalizações mais amplas do mesmo para a grande área do conhecimento a qual a pesquisa realizada pertence.

Podemos também dizer que a relação entre pesquisadores nativos ou não-nativos e os padrões de organização apresentados é que, quando conhecidos, podem permitir um melhor entendimento das características determinantes de um artigo científico, como também, podem contribuir para que se tenha facilitado o processo de escrita para publicação e,

conseqüentemente, se obtenha uma boa comunicação do propósito da pesquisa, objetivo principal de um relato científico.

Em vista disso, neste projeto pretendemos utilizar o Modelo de Swales (1990) – Movimentos e Passos Retóricos – para a criação de um manual de anotação de estruturas retóricas, com procedimentos detalhados e exemplificados, para cada seção textual constitutiva de um artigo científico. O destino de tais manuais é ser utilizado pelo público-alvo deste trabalho, interessado em construir uma ferramenta de suporte à escrita seguindo nossas etapas aqui propostas. Considerado um dos precursores nessa área de investigação, o trabalho de Swales tem servido de aporte teórico para pesquisadores que se dedicam a questões sobre gêneros textuais, principalmente na área de Inglês para Fins Acadêmicos.

O próximo foco dessa investigação, que é apresentado a seguir, visa a descrição de como as informações retóricas são apresentadas em cada um desses movimentos, ou seja, mostrar como as informações retóricas se materializam lingüisticamente em estratégias retóricas por meio de determinadas escolhas léxico-gramaticais representativas desses movimentos.

3.5. Estrutura Esquemática

Na literatura, muito já se discutiu sobre a publicação científica via artigo científico ser um processo documentado e ordenado segundo um esquema clássico e metódico de composição, que visa abranger os objetivos contidos nesse gênero textual (Severino, 1996; Barrass, 1979; Weissberg & Buker, 1990; Swales, 1990). Segundo esses estudos, a estrutura subjacente a um artigo é composta por Introdução, Desenvolvimento e Conclusão, podendo o Desenvolvimento desdobrar-se em Metodologia (ou Materiais e Métodos) e Resultados, ou ainda Metodologia, Resultados e Discussão.

Em linhas gerais, pode-se dizer que essa estrutura deve guiar o leitor, fazendo com que ele siga o fluxo de informação, que obedece ao movimento geral-específico-geral iniciado na Introdução e finalizado na Conclusão. Para tanto, cada um dos componentes dessa estrutura desempenha um papel bem definido, que será discutido a seguir.

De acordo com o ponto de vista semântico de Kintsch & van Dijk (1978[68] *apud* Aluísio, 1995 e Fontana, 1989), a estrutura de um texto é formada por dois níveis, o micro e o macroestrutural. O primeiro sendo caracterizado pela estrutura das sentenças e de suas

[68] KINTSCH, W. & van DIJK, T.A. Toward a model of text comprehension and production. *Psychological Review*, 85, p. 363-394, 1978.

relações, e o segundo pela natureza global, definindo os principais tópicos de discussão do texto. Nos artigos científicos, gênero textual fortemente convencionalizado, a seleção dos principais tópicos (macroestrutura) é dirigida por um esquema formal, ou seja, pelos componentes da estrutura esquemática do discurso ou superestruturas *(*Kintsch & van Dijk, 1978: 366 *apud* Fontana,1989) contidos em artigos científicos. Esses componentes esquemáticos organizam as macro-posições ao longo do texto, definindo qual macro-elemento se ajusta melhor às funções específicas convencionalizadas para cada texto, além de ajudar a identificar seu gênero.

Do ponto de vista da estruturação de componentes textuais em propósitos comunicativos, ou funções retóricas desempenhadas, temos o trabalho de Swales (1981b; 1990), que utiliza o termo *Moves* (Movimentos) para se referir à função retórica contida em cada componente textual. Bhatia (1993) também mantém a utilização desse mesmo termo em seus trabalhos, ao se referir às funções retóricas: "(...), cada movimento também serve uma intenção comunicativa típica que é sempre subserviente ao propósito comunicativo maior do gênero" (Bhatia, 1993:30). Entretanto, é possível encontrarmos na literatura, definições distintas para as mesmas partes que estruturam retoricamente um texto científico, como mostra os termos empregados em alguns estudos sobre gênero textual, como, por exemplo, o de Motta-Roth (1995). Essa autora, por sua vez, utiliza o termo **subfunções**, para definir "uma série de unidades funcionais menores ou atos de fala, como informar ou perguntar, que realiza a intenção do escritor de acordo com as limitações impostas pelo gênero" (Motta-Roth: 1995). Podemos encontrar ainda o uso dos termos **funções** e **subfunções** retóricas de um texto, por exemplo, no trabalho de Santos (1996).

Importante lembrar, que estudos indicam que, apesar das diferenças existentes na organização e na elaboração da estrutura textual científica, devido às diferenças entre as nacionalidades, culturas e áreas do conhecimento, os trabalhos científicos compartilham uma mesma estrutura genérica (Taylor & Tingguang, 1991[69] *apud* Feltrim: 2004) dos movimentos textuais. Acrescentam ainda, que essa estrutura, entretanto, não deve ser vista como um conjunto fixo e rígido de etiquetas para se rotular um texto; em vez disso, deve aceitar variações em sua estrutura (Upton, 2002 e Flowerdew, 2005).

Nesta pesquisa de mestrado, seguindo o trabalho de Kintsch e van Dijk (1978), utilizamos **componentes da estrutura esquemática** para referenciar as partes do texto que desempenham determinados propósitos comunicativos/funções retóricas.

[69] TAYLOR, G. & TINGGUANG, C. Linguistic, cultural and subcultural issues in contrastive discourse analysis: Anglo-american and Chinese scientific texts. *Applied Linguistics* 12(3), p. 319-336, 1991.

Vale lembrar que, para cada seção de um artigo científico há uma determinada estrutura esquemática, que é representada no texto por um formato constituído por diferentes peculiaridades (Smith e Lansman, 1988[70] *apud* Feltrim, 2004). Enquanto a primeira segue um esquema de organização mais rigoroso, a segunda, já se apresenta sob diferentes formas.

Importante também dizer que a proposta de apresentar um modelo, no formato de um manual de anotação, para identificação de componentes esquemáticos e de estratégias retóricas das diferentes seções de artigos científicos em geral, é uma das etapas desta pesquisa. Assim, os Apêndices 1,2,5,6,7 e 8 apresentam, respectivamente, um manual para anotação dos componentes da estrutura esquemática e das estratégias retóricas da seção "Metodologia", "Resumo", "Resultados", "Discussão", "Conclusão" e "Introdução". O modelo escolhido para categorizar pesquisas experimentais e os manuais de anotação visam auxiliar na tarefa de identificação desses dois tipos de informação lingüística em córpus que serão inseridos em futuras ferramentas de auxílio à escrita científica, produzidas com o auxílio deste trabalho.

3.5.3 Estratégias Retóricas

Conforme já mencionado, as seções constitutivas de artigos científicos possuem determinadas estruturas referidas neste trabalho por componentes da estrutura esquemática. A funcionalidade desses componentes é fazer referência às partes do texto, que desempenham determinados propósitos comunicativos ou funções retóricas. Essas estruturas realizam-se lingüisticamente em um texto de diferentes formas, ou seja, por diferentes estratégias retóricas, conforme os componentes esquemáticos nos quais se encontram contidas.

Segundo Swales (1990), essas diferentes realizações/materializações lingüísticas dos movimentos contidos em cada seção de um artigo científico são chamadas de passos *(steps)*.

Vale dizer, que o modelo de Swales serve de base tanto para a identificação dessas estratégias retóricas em nosso córpus de estudo, o córpus Met, como também para a identificação dessas mesmas estratégias em futuros córpus produzidos como o auxílio de nossos manuais.

3.5.4 Expressões Formulaicas

No ensino-aprendizagem de língua estrangeira tem-se a crença de que para se aprender uma língua é essencial que se aprendam suas regras gramaticais. Restando ao vocabulário, apenas o papel de coadjuvante nesse processo, sendo ensinado de maneira segmentada e

[70] SMITH, J.B. & LANSMAN, M. *A Cognitive Basis for A Computer Writing Environment*. Technical Report, University of North Carolina at Chapel Hill, 1988.

artificial. Assim, pode-se deduzir a razão de a comunicação de estudantes submetidos a esse tipo de escola de ensino de língua não ser muito rápida, nem fluente, nem muito efetiva. Isso porque, nesse tipo de processo de ensino-aprendizado, não há uma boa incorporação de regras composicionais e convencionais de língua em uso.

Acreditamos, portanto, que para um aprendiz de língua estrangeira ter um discurso mais rápido e fluente[71], (no nosso caso, o discurso acadêmico), é necessário que tal discurso esteja em acordo com as convenções ditadas pela comunidade, na qual se encontra inserido. Achar que o discurso é, em geral, apenas um conjunto de regras sintáticas e gramaticais, que quando dominadas pelo aprendiz, este apenas escolhe de maneira aleatória os elementos lexicais para compor seu discurso é, entender língua enquanto preenchimento de lacunas. Fato que não é verdadeiro, conforme pode ser demonstrado com a utilização da metodologia baseada em córpus aplicada ao ensino. Frente a isso, julgamos não ser adequada a realização de ensino do léxico de maneira descontextualizada e, no caso das expressões formulaicas, de maneira a isolar seus constituintes, que, a rigor, só têm valor no conjunto da expressão.

Ellis (1997:12) afirma que as fórmulas têm um papel importante não só no desempenho comunicativo, mas também na facilitação do aprendizado de alguns fatores gramaticais. Ellis também afirma, que o aprendizado de L2 envolve diferentes tipos de conhecimento: por um lado, o do aprendiz de L2 que internaliza fórmulas; por outro, o desse mesmo aprendiz que aprende regras (o contexto determina a função da expressão utilizada). Ainda segundo Ellis (1997), ao se estudar uma língua estrangeira, não é suficiente para esse aprendiz instruir-se apenas de itens lexicais que a compõem, mas também do sistema lingüístico que estrutura essa língua. Assim, é interessante que esse aprendiz tenha consciência tanto da sistematização da gramática, quanto da língua como um todo, pois pode vir a perceber também que a língua não é uma mera combinação de palavras, mas que, por exemplo, que seqüências semelhantes em diferentes contextos podem produzir diferentes significados. Mas o que são essas expressões formulaicas ou fórmulas, como preferem alguns autores?

Autores como Wray (2002[72] *apud* Tavares, 2004) e Tagnin (1989), trabalham com o conceito de fórmula. Tagnin (1989:57), do ponto de vista pragmático[73], faz menção às *fórmulas situacionais*, as quais seriam expressões utilizadas em determinadas ocasiões, ou

[71] Fluência, isto é, com estruturação e conteúdo adequados.
[72] WRAY, A. Formulaic Language and the Lexicon. Cambridge: Cambridge University Press, 2002.
73 Levinson (1983:32) afirma que o uso moderno do termo pragmática é atribuído ao filósofo Charles Morris que se ocupava com a ciência dos sinais, ou a semiótica. Para o autor, o termo pragmática inclui tanto aspectos de estrutura lingüística dependentes do contexto como princípios de uso de língua e entendimento que muitas vezes não têm nada, ou têm muito pouco a ver com a estrutura lingüística. Assim, os pragmaticistas estão especificamente interessados na inter-relação da estrutura da língua e os princípios de uso da língua.

seja, em situações que exigem um determinado ritual. Se entendermos ritual enquanto seqüência de atos consagrados pelo uso, podemos denominar tais expressões como fórmulas de rotina. Tagnin ainda reforça a utilidade de tais expressões ao declarar que, na conversação diária, grande parte da nossa fala segue caminhos já trilhados. Em geral, as conversas são destituídas de um caráter mais profundo, desenvolvendo-se de acordo com padrões de pensamento e de expressão verbal pré-concebidos. Esses padrões fazem com que nossa comunicação flua com mais facilidade e eficiência, evitando a necessidade de sermos criativos a todo o instante, o que tornaria a conversação uma prática dificílima. Esta proposta serve para ambos os interlocutores, uma vez que o ouvinte não seria capaz de estar constantemente decodificando seu interlocutor. Em suma, Tagnin estabelece a existência de uma conveniência lingüística indicada pela disponibilidade de um conjunto de expressões às quais podemos recorrer sempre que necessário. Para Wray (2002: 7), a fórmula:

> (...) pode ser uma sentença completa ou um grupo de palavras, ou pode ser uma só palavra, ou pode ser somente parte de uma palavra, - (...), mas precisa sempre ser algo que para o instinto da fala seja uma unidade que não pode ser mais analisada ou decomposta da mesma forma que uma combinação livre pode.

Wray (2002:15) aponta o fato de que, a fórmula prevalece no sistema da linguagem adulta, por conta do processamento do princípio da economia. Essa economia acontece, porque temos acesso a estruturas pré-fabricadas, que utilizamos para expressar nossas idéias, sem que precisemos recorrer a um trabalho de criar um novo enunciado sempre que desejamos expressar uma idéia. Wray também afirma, que as palavras se combinam e obedecem às restrições atribuídas pelo contexto, e pelas regras sociais definidas. Tornando claro que, uma vez mapeados os padrões de distribuição para as palavras, fica perceptível que as combinações não são explicáveis apenas através de ajuntamentos graduais, ou seja, por meio de análise de seqüências lineares de sentenças. Ainda segundo o autor, as palavras, que funcionam em uma seqüência formulaica, produzem um significado que vai além do somatório de significados individuais, pois os falantes não as decodificam isoladamente, mas obtêm um significado advindo do todo que estas representam.

Wray (2002:11) também define o termo **seqüência formulaica** levando em consideração que uma expressão formulaica é também uma seqüência formulaica. Sua proposta para caracterização de uma seqüência formulaica é:

> (...) uma seqüência, contínua ou descontínua, de palavras ou outros elementos, que é, ou parece ser, pré-fabricada: isto é, armazenada e acessada por inteiro da memória na hora do uso, ao invés de ser sujeita a geração ou análise pela gramática da língua.

No entanto, esse autor (Ibid: 44) reconhece a grande dificuldade existente em se encontrar uma única definição capaz de capturar todos os traços relevantes para a identificação de uma fórmula, apesar de outros autores já terem proposto muitas classificações para esse fenômeno lingüístico. Por conta disso, faz algumas considerações sobre uma possível proposta de classificação para as fórmulas. Segundo Wray, as estruturas formulaicas são capazes de conter espaços, que aceitam uma classe aberta de itens, criando uma nova mensagem com pequena criatividade, e trazendo economia e eficácia ao uso da língua. Embora seja necessário lembrar, que há uma infindável capacidade lingüística de forjar novas seqüências formulaicas de todos os tipos, o que dificulta este tipo de classificação. A classificação baseada na prática, por outro lado, não precisa de um arcabouço teórico tão profundo, embora tenha de funcionar em seu propósito inicial. Em dicionário ou ensino de língua, vemos a necessidade de tal abordagem. No entanto, o problema se apresenta na necessidade de decidir o que incluir e o que omitir como seqüência formulaica.

Tão relevante quanto a definição de expressão formulaica é o papel que esse elemento representa dentro do CECARL[74] proposto por este trabalho. Devido as já citadas estratégias de sucesso no ensino-aprendizado de línguas, por meio do reuso de material lingüístico em ferramentas de auxílio à escrita (*cf.* Capítulo 2). A seguir, será apresentada como as expressões formulaicas e outros elementos reutilizáveis de uma língua podem ser trabalhados em seu contexto de uso.

Conforme dito no Capítulo 2, as ferramentas de auxílio à escrita científica baseadas no AMADEUS fazem uso, dentre os recursos lingüísticos reutilizáveis de uma língua, de agrupamentos (*chunks)* de expressões lingüísticas que podem ser (re)utilizadas em contextos distintos dos quais foram coletadas. A reutilização desses pedaços subjaz a idéia de que por meio de uma reorganização desses pedaços textuais, como se fossem peças de LEGO™, um novo texto pode ser produzido. Importante dizer, que essa prática não consiste em plágio, uma vez que sentenças completas não são utilizadas, mas sim apenas partes textuais com informações não factuais, isto é, com informações que não trazem o conteúdo da pesquisa, que descrevem ou representam.

[74] CECARL – Conjunto de Etapas para Criação e Alocação de Recursos Lingüísticos.

Depois de se ter essa experiência com sentenças, o aprendiz pode começar a trabalhar com passagens maiores de textos, repetindo o procedimento de combinar os pedaços, ligando-os agora com elementos conectores (ver seção 3.5.5). A seguir, o aprendiz poderá tentar produzir uma seção completa de um artigo, por exemplo. Para tanto, poderá selecionar os componentes esquemáticos dessa seção, bem como as estratégias retóricas, que realizam lingüisticamente essas estruturas no texto. Para isso, poderá navegar pela base de casos da ferramenta de suporte à escrita, que estiver utilizando. Em seguida, esse aprendiz/autor poderá checar o uso de marcadores discursivos adequados de modo a obter coesão e coerência no texto produzido. Para tal checagem, poderá ser utilizado o ícone "Marcadores Discursivos" contido na ferramenta Scientific Writing, disponibilizada junto de nosso CECARL. Uma outra contribuição interessante dessa mesma ferramenta ao ensino-aprendizagem de escrita científica contextualizada é o auxílio via ícone de "Expressões Formulaicas". Neste item, assim como ocorre no ícone anterior, as expressões formulaicas, coletadas de trabalhos realizados com base em córpus, aparecem organizações sob uma lista de funções as quais podem desempenhar. Mais detalhes sobre esses itens ver seção 4.9.

E por fim, poderá passar para o processo de edição do texto produzido, verificando erros ortográficos, eliminando palavras desnecessárias, checando a consistência das estratégias retóricas selecionadas para compor a seção e a relação existente entre elas. Com o constante uso dessas listas, o usuário do CECARL tenderá a se familiarizar com o uso desses termos e poderá identificar, nos córpus futuramente coletados, como incremento de sua base de casos, identificar novas expressões formulaicas ou até mesmo marcadores discursivos que carregam importantes mensagens textuais. A partir de então, o aprendiz de língua pode começar a "brincar" com seus pedaços de textos, identificando diferentes combinações que aparecem nos textos originais e criando, em seguida, sua própria combinação.

Um exemplo de sentença com partes factuais a serem preenchidas é apresentado na Figura 3.5:

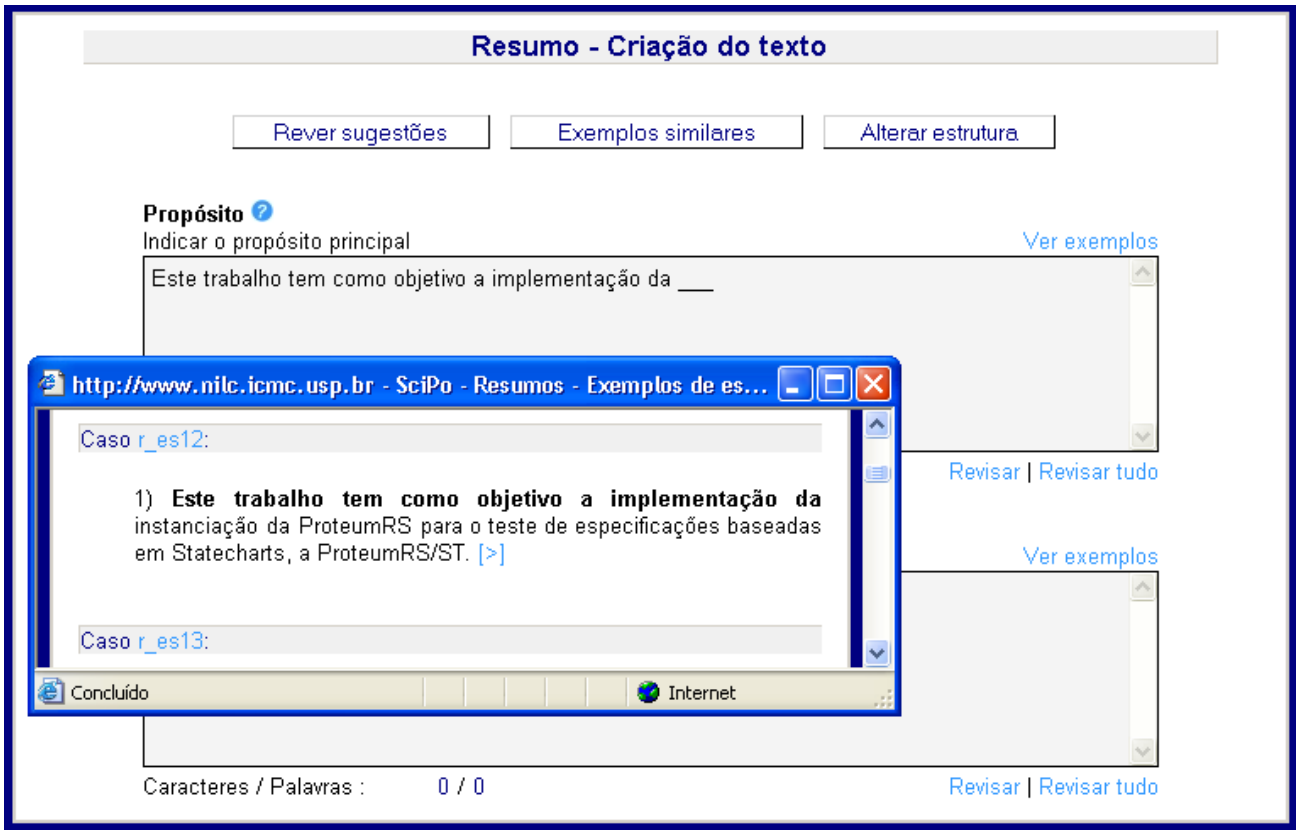


Figura 3.5: Exemplo de reuso de expressões formulaicas do português no SciPo. A navegação pela base de casos da ferramenta SciPo também traz à mostra expressões formulaicas que podem ser reusadas no texto do usuário.

3.5.5 Marcadores Discursivos

Para compor mais um elemento lingüístico descrito em nosso CECARL, foram investigados os Marcadores Discursivos (MDs doravante). Os MDs são uma classe ou categoria de elementos, que se tornou amplamente conhecida, a partir da publicação pioneira ocorrida em 1987 do livro *Discourse Markers* de Deborah Schiffrin. Desde então, tem havido um interesse crescente pela investigação desse assunto[75], assim como da dificuldade também envolvida em tal tarefa, uma vez que os MDs recebem na literatura várias denominações. Isso devido ao fato dessas diferentes denominações refletirem as diferentes abordagens sob às quais os MDs foram examinados. Entre os termos utilizados, podemos encontrar *discourse markers, discourse connectives, discourse operators*, *cue phrases, pesky little particles, metatextual elements, contextualising frames*. Segundo Traugott (1995[76] *apud* Paizan, 2001) outros termos menos freqüentes incluem: *discourse particles, discourse signaling devices, indicating devices, phatic connectives, pragmatic connectives, pragmatic expressions,*

[75] It has been characterized as "a growth industry in linguistics" (Fraser, 1999: 932).
76 TRAUGOTT, E. The role of the development of discourse markers in a theory of grammaticalization. Paper presented at ICHL XII, Manchester, 1995.

pragmatic formatives, pragmatic operators, pragmatic particles, semantic conjuncts, and sentence connectives, discourse connectives, clue words.

As definições para os MDs são igualmente variadas. Redeker (1991[77] *apud* Paizan, 2001), diz que "marcadores discursivos não une apenas sentenças contíguas, mas a sentença *ou* expressão em foco com seu contexto imediato" [78]. Por outro lado, Schiffrin (2001[79] *apud* Paizan, 2001), acredita que "*marcadores discursivos podem conter funções tanto locais quanto globais* (ex.: eles podem ligar significados proposicionais ou, na conversação, determinar a estrutura de troca")[80], incluindo itens como *oh, y'know,* os quais Fraser (1999), Redeker (1991), and Blakemore (2002[81] *apud* Fraser, 2005) não consideram MDs.

Como recorte necessário, limitaremos a discussão sobre MDs aos trabalhos de Fraser (1993;1999; 2005) e Quirk et. al (1985), assim como o fez Paizan (2001), em seu trabalho de produção de um módulo computacional de auxílio a leitura instrumental em inglês. A escolha por Fraser se justifica pela análise clara e sistemática necessária para fundamentar a classificação, que buscamos para os MDs já coletados em nosso córpus Met. Já a escolha por Quirk *et al* (1985), se deve a uma desejável busca exaustiva dos principais MDs existentes no inglês, os quais o fazem nesse trabalho com o auxílio do Córpus SEU (Survey of English Language)[82]. Vale ainda dizer, que o elenco de marcadores discursivos, que selecionamos a partir do nosso córpus Met foi classificado primeiro empiricamente, encontrando, posteriormente, fundamentação teórica nos pressupostos desses autores escolhidos para justificar a classificação das unidades selecionadas.

Em busca de uma sistematização para os MDs coletados em nosso córpus, e para aqueles que ainda o serão, futuramente, em outras coletas, escolhemos Fraser que faz uma análise lingüística de MDs existentes no inglês (veja Seção 3.5.5.1).

---

77 REDEKER, G. Linguistic markers of discourse structure. [review of Discourse Markers by Deborah Schiffrin]. Linguistics, 29(6), p. 1139-72, 1991.

78 discourse markers link not only contiguous sentences, but the current sentence or utterance with its immediate context.

79 SCHIFFRIN, D. Discourse markers: Language, meaning, and context. In: D. SCHIFFRIN, TANNEN, D. & HAMILTON, H. (eds.). The Handbook of Discourse Analysis. Oxford: Basil Blackwell, p. 54-75, 2001.

80 "discourse markers can have both local and global functions (i.e., they may connect propositional meaning or, in conversation, determine the structure of the exchange).

81 BLAKEMORE, D. Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers. CUP, 2002.

82 Córpus que começou a ser compilado por Randolf Quirk e sua equipe, em Londres, a partir de 1953. O SEU foi planejado para ter o tamanho de 1 milhão de palavras. Foi o primeiro a definir um número fixo de textos (200) textos e de palavras (5000) para cada texto, as quais foram analisadas gramaticalmente, com cada ficha recebendo uma categoria gramatical. O conjunto de categorias resultante serviu de base para que se desenvolvessem etiquetadores computadorizados, que fazem a identificação de traços gramaticais, bem como da gramática Conprehensive Grammar of the English Language, de Quirk et al. A transformação completa do SEU em versão computadorizada só ocorreu em 1989, já a sua parte falada havia sido computadorizada antes, sendo conhecida como London-Lund Corpus (Berber-Sardinha, 2000:326).

3.5.5.1 Os marcadores discursivos e o modelo de Fraser (1999)

Segundo Fraser (1996; 2005:1; Schourup,1999[83] *apud* Paizan, 2001), há em toda língua uma classe de expressões lexicais denominada *Pragmatic Markers*. Essas expressões são "pistas linguisticamente codificadas, que sinalizam as intenções comunicativas potenciais do falante" (Schourup,1999:238) e podem ser divididos em quatro tipos:

1. Marcadores Básicos (*Basic Markers*) => modificam o conteúdo proposicional da mensagem: (*I promise* that I will be on time).
2. Marcadores de Comentário (Commentary *Markers*) => acrescentam um comentário sobre o conteúdo proposicional da mensagem e podem ser de diferentes tipos: *Assessment Markers (*We got lost almost immediately. *Fortunately,* a police officer happened by.*), Manner-of-speaking Markers (*A: Mark, you've got to do something. B: *Frankly* Harry, I don't know what to do.*), Evidential Markers (*A: Will he go? B: *Certainly*, he will go.*) e Hearsay Markers (*A: Is the game still on? B: *Reportedly*, the game was postponed because of rain*).*
3. Marcadores Paralelos (Parallel Markers) => sinalizam o acréscimo de um conteúdo proposicional à mensagem e podem ser de dois tipos: *Deference Markers* (*Sir,* you must listen to me) e *Conversational Management Markers* (*Now*, where were we when we were interrupted?).
4. Marcadores Discursivos (Discourse Markers) => seu significado é procedimental e não conceitual, uma vez que não apresentam um conjunto de características semânticas, mas sim como o segmento do qual fazem parte deve ser interpretado com relação ao anterior, especificando uma espécie de roteiro para a interpretação da relação entre os segmentos. (A: I like him. B: *So*, you think you'll ask him out then?).

A seguir, essa última classe de marcadores será tratada com mais detalhes ainda sob o ponto de vista semântico-pragmático de Fraser, partindo-se de sua definição canônica para os elementos contidos nessa classe:

[83] SCHOURUP, L. Discourse Markers. *Língua 107*, p. 227-265, 1999.

> Para uma seqüência de segmentos do discurso S1-S2, cada qual codifica uma mensagem completa. Uma expressão lexical LE funciona como um marcador discursivo se, quando ela ocorre em posição inicial de S2 (S1 – LE + S2), LE indicar que uma relação semântica que ocorre entre S2 e S1 possa ser: a) Elaboração; b) Contraste; c) Inferência ou d) Temporalidade.[84] (Fraser, 2005:4)

Tal definição que tem por base relações que se estabelecem entre segmentos discursivos é ampliada e comentada pelo próprio Fraser nos próximos quatro segmentos a seguir:

1. O primeiro ponto a ser observado é sobre a especificidade da definição que diz que S1 e S2 são segmentos discursivos contíguos. No entanto, como o próprio Fraser (2005) mostra, os segmentos relacionados pelos MDs não precisam estar necessariamente lado-a-lado. Ou seja, o MD pode relacionar S2 com o enunciado imediatamente precedente ou anterior a este (ver ex. 1). Pode ainda relacionar S2 com vários segmentos anteriores (ver ex.2), com o contexto situacional (ver ex. 3) ou ainda com segmentos subseqüentes (ver ex.4).

(ex.1) A: I don't want to go very much. B: John said he would be there. A: *However,* I do have an obligation to be there.
(ex.2) You want to know the truth? *Essentially,* John stayed away. Jane came but didn't participate. And Harry and Susan fought the entire evening.
(ex.3) (on entering the room and finding the computer missing) So, Where'd you put it?
(ex.4) You want to know how my garden grew this summer. *Essentially*, the tomatoes grew well. The broccoli was fair as were the peppers. The eggplant and carrots were terrible.

2. Os MDs, conforme observado nos exemplos que se seguem, necessariamente não introduzem o segmento em que ocorrem, podendo, portanto, estar em posição medial ou final da sentença.

(ex.1) It is freezing outside. I will, *in spite of this*, not wear a coat.
(ex.2) We don't have to go. I will go, *nevertheless*.

[84] For a sequence of discourse segments S1 – S2, each of which encodes a complete message. A lexical expression LE functions as a discourse marker if, when it occurs in S2-initial position (S1 – LE + S2), LE signals that a semantic relationship holds between S2 and S1 which is one of: a) Elaboration; b) Contrast; c) Inference; or d) Temporality (Fraser, 2005:4).

3. Quanto à função sintática, os MDs:

a) Relacionam frases independentes.

(ex.1) We left late. *However*, we arrived at home on time.

b) Duas orações de estruturas distintas, <S1. MD+S2> ou <S1, MD+S2> podem ser conectadas por um MD proveniente da classe das conjunções coordenadas.

(ex.1) Jack played tennis. *And* Mary read a book.

Jack played tennis, *and* Mary read a book.

c) Algumas expressões não funcionam como MDs se não introduzirem uma nova mensagem. Assim, *as a result of* funciona como MD em (ex.1) mas não em (ex.2).

(ex.1) There was considerable flooding. *As a result*, farmers went bankrupt.

(ex.2) *As a result of* considering flooding, farmers went bankrupt.

d) Elementos como *since, because, while* e *unless* apresentam estruturas diferentes: <S1, MD+S2> e <MD+S2, S1>, como em (ex.1) e (ex.2) respectivamente, pois, provenientes da classe das conjunções subordinadas, não podem introduzir uma única sentença, exigindo a presença de uma outra independente anterior, como nos exemplos (ex.1b) e (ex.2b):

(ex.1) Harry will not go, *unless* he is paid an appearance fee.

(ex.2) *While* she is pregnant, Martha will not take a plane.

(ex.1b) U*nless* he is paid an appearance fee.

(ex.2b) A: Harry will not go. B: *Unless* he is paid an appearance fee.

4. Para que uma seqüência seja considerada coerente, as interpretações dos segmentos discursivos S2 e S1 devem ser compatíveis com o MD utilizado. No próximo exemplo, S2 deve ser interpretado como uma promessa, ou pelo menos, uma não ameaça. Assim, em geral, a mensagem que o falante pretende transmitir com o enunciado deve ser considerada em qualquer determinação de coerência.

   (ex.1) I will help you. *Similarly*, I will take care of Martha.

   No próximo exemplo, o MD relaciona mensagens explícitas veiculadas tanto por S1 quanto por S2. Já em (ex.2) e (ex.3) os MDs relacionam a mensagem explícita veiculada por S2 e uma implícita veiculada por S1, havendo uma proposição subentendida em S1 e uma pressuposta em S2, respectivamente.

(ex.1) A: Box up my entire office. B: *So*, he fired you too.

(ex.2) I realize that Jack is sick. *But* you know Jack is not sick.

Depois de ter definido e argumentado por meio de exemplos que um MD é um tipo de expressão lexical, que sinaliza um tipo de relação existente entre segmentos discursivos adjacentes, Fraser (2005:7), analisa a disposição de MDs em diferentes níveis lingüísticos, começando pelo fonológico.

Segundo o autor, não parece haver grandes generalizações possíveis sobre a fonologia associada aos MDs. Eles, normalmente, não são átonos, mas podem o ser, especialmente, quando o MD é monossilábico, como por exemplo, *but, so* e *and*, com os quais a seqüência das sentenças se constituem em S1+DM+S2 e os MDs estão em posição inicial:

a) Child: There was a big puddle. Parent: *So* **-** you had to jump right in?
b) A: John is at home. B: *But* **-** I just saw him at the mall.

E quando se tem ênfase no segundo segmento, o MD é frequentemente precedido de uma pausa: John was hungry – *so* he must have been really grouchy.

Quanto ao nível morfológico, Fraser (2005), também afirma não ter muito o que dizer. Apenas cita que, assim como muitos MDs são monossilábicos (*but, so, and* e *thus*), há aqueles que são polissilábicos (*furthermore, consequently, nevertheless,* e *before*), e aqueles que consistem em uma expressão inteira (*as a consequence; I mean* e *that is to say*).

Quanto ao nível sintático, embora a classe de MDs seja definida funcionalmente como aquela, cujas expressões lexicais sinalizam uma relação existente entre mensagens adjacentes, essas expressões são todas membros de uma, entre cinco categorias sintáticas: conjunções coordenadas, subordinadas, preposições, locuções preposicionais e advérbios.

a) *Conjunções Coordenadas* => and, but, or, nor, so, yet…
b) *Conjunções Subordinadas* => after, although, as, as far as, as if, as long as, assuming that, because, before, but that, directly, except that, given that, granting that, if, in case, in order that, in that, in the event that, inasmuch as, insofar that, like, once, provided that, save that, since, such that, though, unless, until, when(ever), whereas, whereupon, wherever, while…

c) *Advérbios* => anyway, besides, consequently, furthermore, still, however, then…
d) P*reposições* => despite, in spite of, instead of, rather than…
e) L*ocuções Preposicionais* => above all, after all, as a consequence (of that),as a conclusion, as a result (of that), because of that, besides that, by the same token, contrary to that, for example, for that reason, in addition (to that), in any case/event, in comparison (with that), in contrast (to that), in fact, in general, in particular, in that case/instance, instead of that, of course, on that condition, on that basis, on the contrary, on the other hand, on top of it all, in other words, rather than that, regardless of that,…), as quais podem ser agrupadas em três variações: *as formas fixas* (above all, after all, as a conclusion…); a forma PREP+*that* na qual *that* faz referência a S1 (despite that, in spite of that, in addition to that…) e a forma DM+*of this/that* na qual *that* faz referência a S1 (as a result of that, because of that, instead of doing that), rather(than do/that).

E é justamente a categoria sintática de cada MD, que determinará o local de sua ocorrência em S2. Assim, todos os MDs, com exceção de *though*, ocorrem na posição inicial de S2; sendo a posição inicial de S2 a única condição possível de realização das conjunções coordenadas e subordinadas, devido às restrições sintáticas impostas às conjunções. As outras três categorias (preposições, locuções prepositivas e advérbios), têm uma maior extensão de ocorrência sintática, podendo ocorrer em posição final de S2, com outros ocorrendo tanto em posição medial quanto final.

a) A: You must go today. B: *But* I *(*but)* don't want to go *(*but)*.
b) We started late. *However,* we (*however*) arrived on time (*however*).
c) The trip was tiring. *Despite that*, he *(*despite that*) remained cheerful (*despite that*).
d) A: The movie is over. B: *Then* we *(*then)* should head for home (then).

Do ponto de vista semântico, há três questões a serem consideradas:
1. Um elemento estando na função de MD, relaciona dois segmentos discursivos, mas não contribui para o conteúdo proposicional, ou seja, para o valor verdade contido. No exemplo a seguir, nota-se que o MD pode ser retirado, sem que se afete o valor verdade da proposição:
I want to go to the movies tonight. *After all*, it's my birthday.
I want to go to the movies tonight. It's my birthday.

Entretanto, quando retirado o destinatário/interlocutor fica sem pistas para especificar a relação pretendida pelo remetente/locutor dos segmentos. Também é de se duvidar, que todas as relações podem ser reconhecidas com a ausência de um MD.

2. O significado de um MD é procedimental e não conceitual, conforme dito anteriormente, uma vez que esses elementos não apresentam um conjunto de características semânticas, mas sim a expressão de como o segmento do qual fazem parte deve ser interpretado em relação ao adjacente, isto é, funcionam como guias, roteiros para se interpretar a relação existente entre segmentos. O MD *in contrast*, por exemplo, sinaliza abaixo o contraste entre dois indivíduos e o seu peso relativo: John is fat. *In contrast*, Jim is thin.

3. Cada MD possui um significado nuclear procedimental que específica o roteiro de interpretação, que não esgota a interpretação global do enunciado, que por sua vez depende dos contextos lingüístico e situacional: Susan is married. *So*, she is no longer available, I guess.

Continuando suas reflexões a respeito de MDs, sob o ponto de vista semântico, Fraser (2005) traz uma tipologia semântica, que considera básica aos mais de cem MDs existentes na língua inglesa, e produto do reflexo das mesmas em seus respectivos usos. Os MDs podem, segundo esse autor, ser classificados em dois grandes grupos: aqueles que relacionam mensagens (subdividem-se em quatro) e aqueles que relacionam tópicos, composto de apenas uma subdivisão. O autor ainda acrescenta, que sua proposta não se pretende exaustiva, mas se trata de elaboração já contida em Fraser (1999). Para a primeira classe citada, representou o que considerava ser o MD primário de cada uma representando-o em negrito, com seus respectivos membros:

1) *Marcadores Contrastivos* => **but,** alternatively, although, contrariwise, contrary to expectations, conversely, despite (this/that), even so, however, in spite of (this/that), in comparison (with this/that), in contrast (to this/that), instead (of this/that), nevertheless, nonetheless, (this/that point), notwithstanding, on the other hand, on the contrary, rather (than this/that), regardless (of this/that), still, though, whereas, yet

   Sinalizam que a interpretação explícita do S2 contrasta com a interpretação de S1. O conteúdo de S1, por sua vez pode ser explícito, pode ser uma mensagem implícita não esperada ou ainda uma mensagem acarretada, como observado respectivamente nos excertos abaixo:

   We left late. *Nevertheless*, we got there on time.

A: Chris is a happy bachelor. B: Chris is a female.

2) *Marcadores Elaborativos =>* ***and****, above all, also, alternatively, analogously, besides, by the same token, correspondingly, equally, for example, for instance, further*(*more*)*, in addition, in other words, in particular, likewise, more accurately, more importantly, more precisely, more to the point, moreover, on that basis, on top of it all, or, otherwise, rather, similarly*, *that is* (*to say*)

Sinalizam uma relação quase paralela entre S2 e S1. Em todos os casos, os MDs indicam uma relação de equivalência entre as mensagens expressas por S2 e S1, podendo também, aumentar ou refinar a mensagem de S1. No exemplo abaixo, o MD *furthermore* sinaliza que o conteúdo de S2 deve ser considerado como um item a mais da lista especificada no discurso anterior.
The picnic is ruined. The mayonnaise has turned rancid. The beer is warm. *Furthermore*, it's raining.

3) *Marcadores Inferenciais =>* ***so****, after all, all things considered, as a conclusion, as a consequence* (*of this/that*), *as a result* (*of this/that*), *because* (*of this/that*), *consequently, for this/that reason, hence, it follows that, accordingly, in this/that/any case, on this/that condition, on these/those grounds, then, therefore, thus*

Sinalizam que S2 deve ser interpretado como uma conclusão baseada em S1:
It's raining. *Under those conditions*, we should ride our bikes.
There's a fearful storm brewing. *So*, don't go out.

4) *Marcadores Explicativos => because, for this/that reason, since, after all.*

Sinalizam que o S2 fornece um motivo para o conteúdo expresso em S1.
I'm not going to live with you anyway, *since* I can't stand your cooking.

Em Fraser (2005:122) essa quarta classe é substituída pela classe de Marcadores Temporais => ***then****, after, as soon as, before, eventually, finally, first, immediately afterwards, meantime, meanwhile, originally, second, subsequently, when*.

A segunda classe de MDs, formada pelas relações de tópico, envolve apenas um aspecto de gerenciamento do discurso:

> Marcadores de Tópicos => to return to my point of, incidentally, back to my original point, before I forget, by the way, incidentally, just to update you, on a different note, speaking of X, that reminds me, to change to topic, to return to my point, while I think of it, with regards to.

Em Fraser (1993), há ainda uma terceira grande classe para agrupamento de MDs, intitulada Marcadores de Atividade do Discurso, e consiste em MDs que sinalizam a atividade discursiva corrente, relativa a alguma parte precedente do discurso. Essas atividades referem a tipos de discurso, que operam como explicação ou sumarização, por exemplo, e não ao tipo de mensagem (um pedido ou uma promessa) que o falante/escritor transmite na comunicação. Fraser (1993:10-11) alerta ainda que, apesar das sete categorias levantadas, não se trata de uma lista completa.

1. *Esclarecimento*: by way of clarification, to clarify…
2. *Concessão*: admittedly, after all, all in all, all the same, anyhow, anyway, at any rate, besides, for all that, in any case/event, of course, still and all…
3. *Explicação*: by way of explanation, if I may explain, to explain…
4. *Interrupção*: if I may interrupt, to interrupt, not to interrupt…
5. *Repetição*: at the risk of repeating myself, once again, to repeat…
6. *Seqüência*: finally, first, in the first place, lastly, next, on the one/other hand, second, to begin, to conclude, to continue, to start with…
7. *Sumarização*: in general, in summary, overall, so far, summarizing, summing up, thus far, to sum up, at this point…

3.5.5.1 Os marcadores textuais e o modelo de Quirk *et al* (1995)

Conforme dito anteriormente, Quirk *et al* (1985) foi escolhido em nossa busca por um levantamento exaustivo dos principais MDs do inglês por se desejar que o mesmo fosse produto de dados empíricos, como o foi o trabalho apresentado por esses autores.

Quirk *et al* (1985: 631-632), dão a esses elementos a denominação de *Conjuncts* e dizem que sua função é a de "relacionar unidades independentes e de sinalizar como o falante

vê a relação existente entre essas unidades", o que não deixa de ser um tipo de caracterização formal do que foi apresentado até o momento.

Para classificar um dado elemento como *conjunct*, os autores propõem um teste heurístico, compreendido por quatro itens (Quirk *et al*, 1985: 631), deduzidos a partir da seguinte sentença.

She may be unable to attend the meeting. You should nonetheless send her the agenda.

1. Os *Conjuncts* não podem ser o ponto mais importante de uma sentença dividida;
...*It is *nonetheless* that you should send her the agenda.
2. O *Conjunct* nunca é base de uma de uma sentença interrogativa ou negativa alternativa;
…*Should you send he the agenda *nonetheless* or *therefore*?
3. O *Conjunct* nunca é o foco de um *subjunct* evidenciador;
…*You should *only* <nonetheless> send her the agenda.

Aqui, vale fazer um parêntese sobre o termo subjunto *(subjunct)*. O subjunto é uma classificação dada aos sintagmas adverbiais e preposicionais, elaborada por Quirk *et al* (1985), e que tem a função de direcionar o ouvinte/leitor a uma dada interpretação: *This play presents visually a sharp challenge to a discerning audience*. Outra característica do subjunto é o seu papel subordinativo em comparação aos outros elementos da sentença que, segundo esses autores, são sujeito, verbo, complemento, objeto, e adjunto adverbial.

4. O *Conjunct* nunca é parte do escopo da predicação de uma elipse ou pro-forma.
...* If they open all the windows, *then* I'm leaving and so is Bob.

Segundo Quirk *et al* (1985:634), podemos distinguir sete papéis semânticos dos *Conjuncts*:

1. *Indicadores ou estruturadores de listas:* indicam a presença de itens enumerados.
Ex: First the economy is beginning to recover, and *secondly* unemployment figures have not increased this month.
2. *Aditivos:* indicam uma soma, ou seja, o enunciado que vem a seguir faz uma somatória de tudo o que foi dito antes.

Ex: He lost his watch, his car broke down, and he got a letter of complaint from a customer: *all in all*, he had a bad day.

3. *Reformulativos:* indicam uma nova expressão do segmento anterior.

Ex: They took with them some chocolate, cans of beer and fruit juice, a flask of coffe, a pack of sandwiches: *in other words,* enough refreshments.

4. *Resultativos:* indicam conseqüência.

Ex: She arrived late, gave answers in an offhand manner, and *of course* displeased the interviewing panel.

5. *Inferenciais:* indicam uma conclusão baseada em lógica e suposição.

Ex: You haven't answered my question; *in other words,* you disapprove of my proposal.

6. *Contrastivos:* indicam uma oposição com o que foi dito antes.

Ex: He expected to be happy but *instead* he felt miserable.

7. *Transicionais:* indicam uma mudança de tópico ou de evento temporariamente relacionado.

Ex1: I want to tell about my trip, but, *by the way,* how is your mother?

Ex2: He saved a great deal of money but *in the meantime* his house deteriorated very badly.

Vale ressaltar nesse momento, que uma análise crítica dos modelos de marcadores discursivos apresentados não são o foco central deste trabalho, mas sim, o levantamento de modelos de caracterização desse tipo de elemento lingüístico, e sua aplicação em nosso córpus de estudo e na ferramenta computacional utilizada para a anotação automática desses elementos em um texto. Para tanto, primamos por não escolher classificações não empíricas, nas quais acontece uma classificação pela classificação, não interessando, pois ao nosso trabalho, que parte da extração de elementos (MDs) de contexto de usos reais, portanto, precisando de caracterizações que também compartilhem desse mesmo pressuposto. Conforme Halliday (1965) propôs, é muito difícil estabelecer cortes em categorias lingüísticas. Em nosso estudo, percebemos que alguns itens lexicais, classificados em uma dada categoria, poderiam também ser incluídos em outras, devido as diferentes funcionalidades que podem adquirir em um texto. No entanto, a classificação que se fez, como toda classificação, pode ser problematizada, mas se tratou de uma tentativa de investigar os marcadores selecionados de maneira mais funcional e sistemática.

Importante também reforçar que defendemos neste trabalho, que as escolhas de marcadores discursivos são largamente determinadas pela estrutura interna do gênero artigo científico, que, por sua vez, é moldada a partir das expectativas e experiências da comunidade científica, a qual o gênero pertence. Entendemos, assim, que os meios retóricos, dos quais os

marcadores fazem parte, podem ser muito semelhantes em diferentes culturas de escrita (*writing cultures*), mas suas freqüências e usos preferenciais diferem (Mauranen, 1993:5[85] *apud* Mirahayuni, 2002). São essas diferenças que nos interessam, pois, uma vez que culturas diferentes produzem textos diferentes, é fundamental que o escritor reconheça quais são as características da língua inglesa necessárias para produzir seu texto de maneira adequada.

No Apêndice 3 deste trabalho, é apresentado um quadro com os marcadores discursivos retirados do córpus Met, a qual também servirá como insumo para compor a ferramenta computacional que identifica automaticamente os MDs em um dado texto.

Depois de se ter delineado constituintes lingüísticos como as expressões formulaicas e os marcadores discursivos, vale citar também a existência de um outro tão interessante quanto, ao ensino-aprendizagem de línguas: os termos de especialidade. No nosso caso, tratam-se dos termos existentes em córpus científicos.

3.5.6. Concordâncias

A concordância é uma listagem, na qual um dado item (palavra isolada, composta, estrutura, etc...) aparece com palavras (co-textos) ao seu redor (Berber-Sardinha, 2000b). Um exemplo de concordância é apresentado na Figura 3.6. O item em destaque na concordância é conhecido por nódulo, palavra-nódulo, nó, palavra de busca ou palavra-chave. Os tipos de concordâncias mais comuns são a KWIC (Key Word In Context) e KWOC (Key Word Out of Context), sendo a primeira a mais convencional, por mostrar a palavra de busca no centro da listagem acompanhada pelas palavras que ocorreram no texto junto a ela. As concordâncias atualmente são feitas por computador, por meio de programas especializados (concordanciadores), embora, na ausência de equipamento, é possível fazer concordâncias à mão, na lousa (Willis, 1998[86] *apud* Berber-Sardinha, 2000b).

[85] MAURANEN, A. Cultural Differences in Academic Rhetoric: A Textlinguistic Study. Peter Lang, Frankfurt, 1993.
[86] WILLIS, J. Concordances in the classroom without a computer. In TOMLINSON, B. (Ed.) *Materials development in language teaching*, Cambridge, 1998.

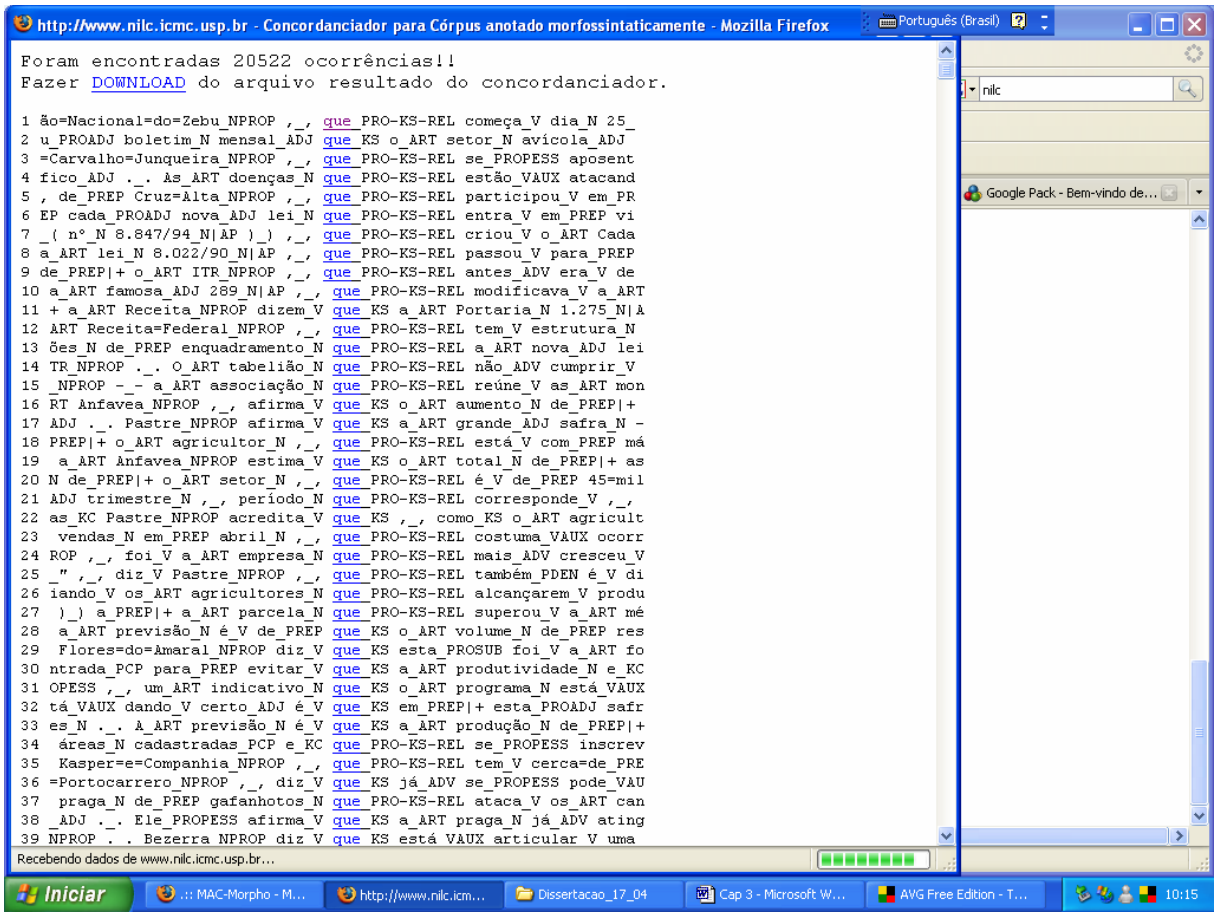


Figura 3.6: Concordância realizada com o concordanciador existente no projeto LacioWeb, que trabalha com córpus do português.(http://www.nilc.icmc.usp.br/lacioweb/macmorpho.php). Como pode ser observado, a palavra selecionada aparece em destaque em meio ao contexto do qual se encontra. Com um clique nesse nó em destaque, é mostrado ao usuário o contexto (texto) ao qual a sentença pertence. Vale também dizer que, especificamente nesse concordanciador, o córpus em uso está anotado morfossintaticamente. Há outros concordanciadores que apenas apresentam as sentenças de um córpus cru (sem anotação), sem que as mesmas estejam com suas informações morfossintáticas, por exemplo, em destaque.

A história de utilização de concordâncias na literatura e na análise lingüística teve início bem antes da era do advento dos computadores. Tribble and Jones (1990[87] *apud* Berber-Sardinha, 2000b) fazem uma síntese sobre a história da origem desse recurso no século 13, quando Hugo de San Charo recrutou quinhentos monges para produzir uma concordância completa da Bíblia em Latim. No entanto, o uso de concordâncias como ferramenta para o ensino-aprendizagem de língua é um fenômeno muito mais recente, que data dos anos 80, com a entrada em cena dos micro-computadores pessoais. Possui como principal característica o fato de estar voltado à instrução limitada de itens do vocabulário de uma

[87] TRIBBLE, C.; JONES, G. Concordances in the classroom: a resource book for teachers. Londres, Longman, 1990.

língua-alvo, e como um de seus maiores representantes o professor Tim Johns, atualmente pesquisador da Universidade de Birmingham, Reino Unido.

Segundo Johns, o primeiro fator se deve a autenticidade conferida ao processo de aprendizado, uma vez que lida com material autêntico de língua em uso. Em segundo lugar, porque os aprendizes têm controle total de seu processo de aprendizagem e, por fim, porque por meio das concordâncias a aprendizagem acaba merecendo a metáfora de pesquisa, pressuposto defendido pelas teorias do aprendizado dirigido por dados (*Data-Driven Learning*: DDL) - da qual Johns é filiado -, que constrói a competência lingüística dos alunos fornecendo aos mesmos fatos do desempenho lingüístico, "nós apenas fornecemos a evidência necessária para responder as perguntas do aprendiz e contamos com a inteligência o aprendiz para encontrar respostas" [88].

Li & Pemberton (1994[89] *apud* Thurstun & Candlin, 1998) também são favoráveis à visão de Johns e dizem que:

> Alunos não precisam necessariamente dominar amplamente os termos acadêmicos de uma área para escreverem artigos que possam ser aceitos. Eles realmente precisam, no entanto, ser usuários competentes de um conjunto restrito de vocabulário 'semitécnico'. [90]

Segundo estudo realizado por Bush *et al* (1996[91] *apud* Thurstun & Candlin, 1998), no qual pesquisadores de quatro universidades australianas foram questionados quanto às suas expectativas em relação à escrita científica de seus alunos, percebeu-se que o uso apropriado do vocabulário acadêmico é extremamente importante. Mas, que há, também, muito mais interesse em fazer com que os alunos comuniquem claramente suas idéias, do que fazerem com que se esforcem para utilizar a linguagem especializada da área em que atuam. São, portanto, esses tipos de estudos e comentários que dão suporte ao ponto de vista desta pesquisa. A abordagem, portanto, mais útil ao auxílio de estudantes, ainda não familiarizados com a escrita acadêmica, seria aquela que os fizesse entrar em contato com os itens lexicais mais importantes, em seu pleno contexto real de uso, conforme as realizações requeridas pelas

[88] "we simply provide the evidence needed to answer the learner's questions, and rely on the learner's intelligence to find answers". (Johns, 1991a:2)
[89] LI, S.E. & PEMBERTON, R. (1994). An investigation of students' knowledge of academic and subtechnical vocabulary, In FLOWERDEW, L & TONG, A.K.K. (Eds.), Entering text, p. 183-196, 1994.
[90] Students, do not necessarily need to master a wide range of academic terms in order to write acceptable academic essays. They do, however, need to be competent users of a restricted set of 'semi technical' vocabulary items.
[91] BUSH, D., CADMAN, C., de LACEY, P., SIMMONS, D., & THURSTUN, J. Expectations of academic writing at Australian universities: work in progress. Paper presented at the *First National Conference on Tertiary Literacy: Research and Practice*. Melbourne, 1996.

funções retóricas dos textos científicos. E o uso de concordanciadores pode propiciar essa rica experiência de linguagem.

Assim, as concordâncias ou o uso de um concordanciador para a verificação das mesmas será utilizado neste projeto (mais detalhes ver Etapas 6 e 7 do Capítulo 4) para oferecer aos usuários deste concordanciador, a oportunidade de condensar e intensificar o processo de aprendizado de vocabulário por meio da exposição a exemplos múltiplos de determinados itens lexicais de forma contextualizada (*keywords* da área de especialidade do escritor). De acordo com Nattinger (1988: 63[92] *apud* Thurstun & Candlin, 1998), "deduzir vocabulário a partir de contexto é a maneira mais freqüente de se descobrir o significado de palavras novas"[93]. Assim, o objetivo maior que permeia a produção das duas últimas etapas do processo apresentado no capítulo 4 é auxiliar o desenvolvimento da competência lingüística de escritores em língua estrangeira. De modo que possam, sozinhos, descobrir os significados existentes, padrões importantes da linguagem em uso investigada pelo concordanciador, e, também, estruturas gramaticais que devem ser empregadas. Essa aquisição de consciência quanto aos termos de especialidade, por exemplo, pode ser acompanhada do despertar para uma prática da investigação, não só científica, já realizada por eles, mas também para a pesquisa lingüística, com a investigação de padrões e formas de organização da língua.

Berber-Sardinha (2000), um dos pesquisadores pioneiros da Lingüística de Córpus no Brasil, defende que o vocabulário não é um fenômeno que deve ser visto de forma isolada da sintaxe. Ele descreve padrões léxico-gramaticais que são igualmente importantes para o ensino de vocabulário. São eles: 1. Colocação (associação entre itens lexicais), 2. Coligação (associação entre itens lexicais e gramaticais. Ex. 'start' é mais comum com sintagmas nominais e orações /ing/, enquanto 'begin' é mais comum com um complemento 'to') e 3. Prosódia semântica (associação entre itens lexicais e conotação - negativa, positiva ou neutra - de campos semânticos. Ele cita como exemplo a palavra 'cause' que se associa com palavras desfavoráveis (*problems, damage, death*) e 'provide' que se associa com palavras positivas ou neutras (*assistance, care, job*)). Um termo geral que abarca os padrões léxico-gramaticais acima citados é *chunk* (agrupamentos, porções). Este termo é normalmente empregado em trabalhos voltados ao ensino de línguas (Lewis, 1993[94], 1997[95] *apud* Berber-Sardinha, 2000).

[92] NATTINGER, J. Some current trends in vocabulary teaching. In CARTER, R. & Mccarthy, M. (orgs) *Vocabulary and language teaching*. New York: Longman, 1988.
[93] "guessing vocabulary in context is the most frequent way we discover the meaning of new words".
[94] LEWIS, M. The lexical approach: the state of ELT and a way forward. Hove, LTP, 1993.
[95] LEWIS, M. Implementing the lexical approach – Putting theory into practice. Hove: LTP, 1997.

De modo geral, a padronização é a regularidade expressa na recorrência sistemática de unidades co-ocorrentes de várias ordens (lexical, gramatical, sintática, etc.). Como definem Hunston & Francis (2000:37[96] *apud* Berber-Sardinha 2000):

> Os padrões de uma palavra podem ser definidos como todas as palavras e estruturas que são regularmente associadas com a palavra e que contribuem para o seu significado. Um padrão pode ser definido se uma combinação de palavras ocorre relativamente de maneira freqüente, se ela é dependente de uma escolha particular de palavra e se há um significado claro a ela associado. [97]

Os vários tipos de padrão estão interligados, e essa interligação é particularmente importante para o ensino de línguas estrangeiras, visto que para um aluno é importante saber como os vários ângulos de descrição da léxico-gramática estão interligados (Hoey, 2000[98] *apud* Berber-Sardinha 2000b).

Conforme já mencionado acima, a concordância é um recurso/instrumento típico da investigação em Lingüística de Córpus, mas que também pode ser empregado no ensino-aprendizado de línguas, via diferentes abordagens de ensino, como a *Lexical Approach*, a *Data Driven Learning*, entre outras (Tribble & Jones, 1990[99] *apud* Berber-Sardinha, 2000b).

Assim como em qualquer outra abordagem ou metodologia de ensino-aprendizagem, também há críticas ao uso de concordâncias. Em geral, essas críticas alertam que o computador, o córpus e as concordâncias não devem ser considerados os únicos instrumentos para o ensino de línguas, mas sim utilizados com consciência de suas vantagens e limitações. Entretanto, a crítica mais conhecida, diz respeito à possível incompatibilidade entre o uso de concordâncias e o ensino comunicativo de línguas, já que as concordâncias promoveriam a descontextualização da língua, pelo fato de mostrarem pequenos trechos provenientes de vários textos (Aston, 1995[100] *apud* Berber-Sardinha, 2000b). Este problema pode ser evitado por meio do acesso a um concordanciador que ofereça a visualização dos textos de um córpus na íntegra, como é o caso do concordanciador da Figura 3.6. O concordanciador gerado ao final das etapas apresentadas no Capítulo 4, para a obtenção de uma ferramenta de auxílio à

[96] HUNSTON, S., FRANCIS, G. Pattern grammar: a corpus-driven approach to the lexical grammar of English. Amsterdã/Filadélfia, John Benjamins, 2000.
[97] "The patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it".
[98] HOEY, M. A world beyond collocation: New perspectives on vocabulary teaching. In LEWIS, M. (org). *Teaching collocation: further developments in the lexical approach*. Hove, LTP, p. 224-43, 2000.
[99] TRIBBLE, C. & JONES, G. *Concordances in the Classroom*. London: Longman, 1990.
[100] ASTON, G. Corpora in language pedagogy: Matching theory and practice. In COOK, G. & SEIDLHOFER, B. (Eds), *Principle and practice in applied linguistics*, Oxford: Oxford University Press, p. 257-270, 1995.

escrita científica, também possibilitará o acesso ao texto completo. Vale dizer, que os artigos científicos acoplados nessa mesma ferramenta estarão com suas informações retóricas, marcadores discursivos e expressões formulaicas também em destaque, para também estimularem o escritor-aprendiz a identificar que tipos de padrões, vocabulários, estruturas e organizações ocorrem em um texto científico de sua área.

Além de permitir a descoberta e o ensino de padrões autênticos encontrados em córpus, a exploração lingüística via concordâncias também pode ser justificada do ponto de vista da psicolingüística, conforme argumenta Hoey (2000: 238):

> Assim como nós aprendemos nossa primeira língua, nós construímos em nossa cabeça um perfil das palavras que estamos aprendendo. O tão conhecido Language Acquisition Device que existe na cabeça de um bebê é provavelmente uma adaptação de um concordanciador que nos habilita a encontrar regularidades e traços recorrentes em nossa experiência lingüística, do que um dispositivo gerador de gramática.[101]

Por fim, podemos dizer que a exploração lingüística via concordâncias pode servir ao propósito de derrubar alguns mitos existentes no ensino de línguas, conforme apontada Berber-Sardinha (2000). Segundo ele, a Lingüística de Córpus, bem como seu ferramental de investigação, dentre eles as concordâncias, propiciam a exposição de algumas 'verdades' frente à 'mitologias' existentes e difundidas por materiais didáticos e de referência. Em linhas gerais, essa verdade está baseada no fato de a linguagem não ser estruturada pelo princípio do preenchimento de lacunas (Sinclair, 1991), mas sim, padronizada, isto é, caracterizada por traços lingüísticos que não co-ocorrem aleatoriamente, mas de modo estatisticamente significante (Biber, Conrad, & Reppen, 1998). E são esses traços lingüísticos, em geral, e o léxico, em particular, que criam as *relações de expectativa* de Eggins (1994[102] *apud* Berber-Sardinha, 2000), cuja manutenção pelos usuários da língua, transmite ao ouvinte ou leitor a sensação de naturalidade e fluência (Pawley & Syder, 1983[103] *apud* Berber-Sardinha, 2000).

Uma conseqüência direta do confronto desses mitos e verdades sobre a língua para o ensino-aprendizagem de línguas é a negação da separação entre o léxico e a sintaxe, ou seja, a defesa da existência de um nível do sistema lingüístico, que engloba o vocabulário e a gramática, conhecido por léxico-gramática. E é essa a posição assumida pela Lingüística de

[101] "As we learn our first language, we build up in our heads a profile of the words we are learning. The so-called Language Acquisition Device in a baby's head is more likely to be a set of concordancing 'software' that enables us to find regularities and recurrent features in our linguistic experience, rather than any abstract grammar-making device".
[102] EGGINS, S. An Introduction to Systemic Functional Linguistics. London: Pinter, 1994.
[103] PAWLEY, A., SYDER, H. Two puzzles for linguistic theory: native-like selection and native-like fluency. In RICHARDS, J., SCHIMDT, R. (orgs). *Language and communication*. Londres, Longman, p. 191-226, 1983.

Córpus, e sua implementação no ensino pode ser resumida, por exemplo, nas seguintes palavras de McEnery & Wilson:

> Exemplos de corpus são importantes no aprendizado de língua já que eles expõem aprendizes que estão no início de um processo de aprendizagem a tipos de sentenças e vocabulário que eles irão encontrar lendo textos autênticos da língua ou utilizando a língua em situações reais de comunicação. (McEnery & Wilson, 1996:104)[104]

Em conclusão, podemos dizer que recursos utilizados em investigações lingüísticas como, por exemplo, as concordâncias e os córpus eletrônicos, têm provado seu potencial de favorecer a descoberta de informações lingüísticas até então não pensadas ou não tratadas corretamente. E, quando utilizados com prudência e sabedoria, podem se tornar importantes instrumentos no ensino-aprendizagem de línguas, despertando o interesse de aprendizes pela investigação lingüística.

3.5.7 Rubrica

Segundo glossário consultado (pals.sri.com/pals/guide/glossary.html.), a rubrica pode ser caracterizada como um guia, composto por dimensões para avaliar o desempenho de estudantes. Para tal, possui uma escala para medir o desempenho em cada uma dessas dimensões.

Quando aplicadas à avaliação de textos, esses sistemas de avaliação de qualidade podem ser de dois tipos: os de Conteúdo, cujo foco é a análise de significado e os de Estilo, que tentam mensurar a qualidade de estruturas textuais, a adequação de estilo e a fluência de um texto. Esse segundo tipo de avaliação é utilizado em nosso trabalho, mais especificamente na Etapa 2, pois é objetivo dessa etapa avaliar, manualmente, a qualidade seções de artigos científicos escritos em inglês, segundo alguns critérios que serão expostos abaixo.

Vale dizer, que essa tarefa de avaliação de qualidade textual pode ser realizada de maneira automática. Embora pareça demasiadamente complicada, a avaliação automática de qualidade textual tem obtido ótimos resultados, quando aplicada na análise de redações (Kukich, 2000). Esses resultados podem ser observados em testes de larga escala, como o *Graduate Management Admission Test* (GMAT), o *Test of English as a Foreign Language*

[104] Corpus examples are important in language learning as they expose students at an early stage in the learning process to the kinds of sentences and vocabulary which they will encounter in reading genuine texts in the language or in using the language in real communicative situations McEnery & Wilson (1996:104).

(TOEFL), o *Graduate Record Examination* (GRE) e o *General Certificate of Secundary Education* (GCSE), por exemplo.

Especificamente para o gênero científico, vem sendo desenvolvida no NILC, desde 2004, (Aluisio *et al*, 2005; Schuster *et al*, 2005), uma rubrica voltada para a avaliação da qualidade de *abstracts* (resumos de artigos científicos em inglês) produzidos por estudantes não-nativos. São três os objetivos que motivam a construção dessa rubrica: (1) ser incorporada ao SciPo-Farmácia, como recurso automático para a avaliação de textos de estudantes, (2) auxiliar especialistas da área (orientadores) a melhorar tanto o conteúdo, via seleção de conteúdos esquemáticos necessários, quanto a linguagem de textos científicos e (3) avaliar artigos científicos candidatos a serem inseridos no córpus utilizado em ferramentas de suporte à escrita. Em outros contextos de uso, essa rubrica: 1) pode ser um mecanismo unificador de avaliação de textos tanto para estudantes quanto para professores; 2) pode ser um padrão de avaliação para anotadores envolvidos na tarefa de avaliação de textos; 3) pode possibilitar a promoção de *feedback* consistente ao aluno e 4) pode ser um tipo de medida do desempenho de alunos (Schuster *et al*, 2005).

Além disso, o fato de se criar uma rubrica inspirada em outra(s) já existente, pode facilitar muito o processo de elaboração de uma rubrica personalizada para os objetivos de avaliação pretendidos, principalmente, se os objetivos da rubrica de inspiração casam com os da rubrica a ser elaborada. Assim, decidiu-se citar como ponto de partida para a elaboração de uma rubrica personalizada, esse trabalho que vem sendo produzido por uma parceria entre pesquisadores do NILC e da Northern Essex Community College, EUA, desde 2004.

No momento em que essa rubrica estiver totalmente implementada computacionalmente – atualmente só duas de suas sete dimensões o estão - poderá ser permitida, a uma ferramenta de auxílio à escrita, a identificação de erros contidos em *abstracts*, bem como a sugestão de formas mais adequadas de se escrever uma dada informação. Interessante dizer, que podemos dividir esse conjunto de critérios da rubrica citada em duas classes: a classe dos critérios que são dependentes de um domínio e os que não o são. A primeira classe pode ser adaptável para seções de um artigo científico do domínio-alvo que se queira avaliar, enquanto que os critérios da segunda classe podem ser diretamente aplicados nessas seções de texto da área-alvo. Esses últimos critérios foram baseados em estudos feitos por um americano dessa parceria NILC-Northern Essex Community College, que investigou os erros gramaticais mais comuns cometidos por brasileiros ao escreverem em inglês, independentemente da área do conhecimento em que eles escrevem. Nessa investigação, foram analisados 114 *abstracts* provenientes de alunos de áreas como Farmácia, Química, Biologia/Genética, Física e

Ciências da Computação. Assim, adaptar um conjunto de critérios ou simplesmente aproveitar alguns e criar outros, facilita o trabalho de se elaborar uma nova rubrica personalizada para a avaliação desejada.

Atualmente, essa rubrica é composta por sete dimensões, que abordam diferentes aspectos da qualidade de escrita de um resumo escrito em inglês, principalmente, os que tendem a ser críticos para escritores não-nativos do inglês, conforme Figura 3.7.

D1 - Caracterização, Organização e Desenvolvimento. Esta dimensão trata da estrutura do resumo, enfocando tanto a presença de componentes essenciais quanto a sua ordem no texto

Alto:

- Componentes principais presentes e são apresentados em ordem: Propósito, Metodologia (se houver), Resultados principais e Conclusão.
- Se houver uma Lacuna, deve ser seguida pelo Propósito.
- Se existir Contexto e Lacuna, a Lacuna deve aparecer depois do Contexto. Mas é também possível haver ciclos de Contexto e Lacuna.

Baixo:

- Caso contrário.

D2 - Balanceamento entre os componentes. Os resumos em geral não devem ultrapassar um limite de 200 a 300 palavras, o que impõe restrições ao uso de certos componentes estruturais, como contextualização.

Alto:

- Propósito existe e foi escrito em apenas uma sentença.
- Conclusão existe e foi escrita em apenas uma sentença.
- Se existir Contexto, não deve ultrapassar 30% das palavras do *abstract*.

Baixo:

- Caso contrário.

D3 - Coerência entre os componentes. Os componentes de um resumo devem ser relacionados entre si, de forma a contribuir com a coerência do texto.

Alto:

- Se o Propósito estiver relacionado com a Lacuna em uma relação de *fullfilment*.

Note: Como a Lacuna não é necessária, se ela não está presente, o Propósito é assumido como padrão.

- Se os Resultados principais estiverem relacionados com o Propósito em uma relação de *accomplishment*.
- Se a Conclusão estiver relacionada com os Resultados principais em uma relação de *generalization*.

Relações: *Fullfilment* – desejo de realizar alguma tarefa

Accomplishment – realização, alcance

Generalization - obtenção de idéias gerais a partir de instâncias.

Padrão:

- Outras sentenças

Baixo:

- Determinado para os componentes iniciais (destacados acima) se não houver as relações citadas entre eles.

D4 – Marcadores de Coesão. As sentenças de cada componente devem ser coesas. A coesão pode ser alcançada por meio do uso de marcadores discursivos, referências pronominais e reintrodução de nomes.

Alto:

Se cada sentença é relacionada com pelo menos uma outra sentença da mesma categoria esquemática.

Baixo:

Caso contrário.

Padrão:

Se a categoria esquemática é representada por apenas uma sentença.

Note: Ciclos de Contexto e Lacuna são considerados como um único componente nessa dimensão.

D5 - Erros técnicos.

Um dos seis tipos elencados abaixo:
WU (Uso incorreto de uma palavra para expressar um significado pretendido)
ART- (Ausência de um artigo necessário em Inglês)
P (Pontuação)
SP (Ortografia)
WUCol (Uso incorreto de itens lexicais e colocações recorrentes)
ART+ (Presença de um artigo não necessário em Inglês)

Alto: Sem erros

Baixo: Se houver pelo menos um erro gramatical na sentença de um dos seis tipos elencados.

D6 - Estilo. Espera-se que um texto científico não tenha um estilo coloquial e empregue expressões características do gênero.

Alto:

Se a escrita não contém estilo pessoal ou coloquial, com presença de termos como *I, my, lot, for sure, I think, kind of, you know I mean, I think I assume, sort of.*

Baixo: Caso contrário.

D7 - Informação factual. Embora alguns autores prefiram resumos indicativos, espera-se que os resumos sejam informativos, ou seja, tragam informações relevantes sobre o trabalho em questão.

Alto:

Se as sentenças de Resultados principais e Conclusão contêm material informativo.

Baixo:

Caso contrário.

Padrão:

Sentenças de outras estruturas esquemáticas.

Figura 3.7: Rubrica para avaliação de resumos escritos em inglês. **Alto** e **Baixo** são os dois valores que cada dimensão recebe conforme presença ou ausência de dadas características no texto. Há um terceiro valor, **Padrão,** utilizado quando a dimensão não se aplica. Entretanto, esses valores "Alto e Baixo" estão sendo ainda estudados e podem ser alterados na dimensão 5, pois, segundo um dos pesquisadores que estão desenvolvendo a rubrica apresentada acima, é necessário analisar a possibilidade de um resumo ter até 3 erros dos 6 tipos possíveis e, ainda assim, ser aceito.

3.6 Considerações Finais

Apoiando-nos em reflexões feitas durante e após a investigação das ferramentas computacionais de suporte à escrita, e nas teorias lingüísticas apresentadas neste capítulo, cujos conteúdos se relacionam às características de uma produção científica adequada às expectativas da comunidade acadêmica, elaboramos como proposta de trabalho desta pesquisa, um processo, constituído por etapas, para a construção de recursos lingüísticos necessários em ferramentas de suporte à escrita científica, em uma dada área de especialidade.

O próximo capítulo é dedicado à apresentação detalhada desse processo proposto.

4. Processo para Construção e Alocação de Recursos Lingüísticos em Ferramentas de Suporte à Escrita Científica (CECARL)

4.1 Considerações Iniciais

Apoiando-nos em reflexões feitas durante a investigação de ferramentas computacionais e teorias, cujas abordagens visam auxiliar a produção de uma escrita científica adequada às expectativas da comunidade acadêmica, elaboramos um processo para a geração de recursos lingüísticos aplicáveis em ferramentas de suporte à escrita científica, com funcionalidades semelhantes às apresentadas pelo Scipo-Farmácia. Esse processo recebeu a sigla CECARL - CECARL – Conjunto de Etapas para Criação e Alocação de Recursos Lingüísticos.

Esse processo de construção inicia-se com a descrição de como compilar um córpus (conjunto de textos necessários na construção da base de casos da ferramenta) e segue até a indicação dos diretórios nos quais serão alocados adequadamente os recursos lingüísticos produzidos ao longo do processo, para que se tenha ao final, uma ferramenta de suporte à escrita em inglês, personalizada para a área do conhecimento do córpus nela inserido.

O conteúdo das seções e subseções seguintes se dividem em:

1) Apresentação de um diagrama com uma descrição sucinta de todas as etapas envolvidas no processo de geração de uma ferramenta de suporte à escrita científica em inglês, personalizada para uma dada área de especialidade. Assim o usuário de nosso CECARL poderá ter uma idéia geral de todo o processo.

2) Apresentação do conteúdo de cada uma dessas etapas de maneira detalhada: a) diagrama da etapa a ser descrita, b) instrução genérica de como a atividade nela contida deve ser realizada, e c) instanciação em algumas etapas dos procedimentos descritos, ou seja, é apresentado um exemplo de como o processo foi realizado. Em tais exemplificações é utilizado nosso estudo de caso, a construção do Córpus Met, utilizado para a construção e implementação da seção "Metodologia" do SciPo-Farmácia.

4.2 Diagrama do Processo para Construção e Alocação de Recursos Lingüísticos em ferramentas de Suporte à escrita Científica – CECARL

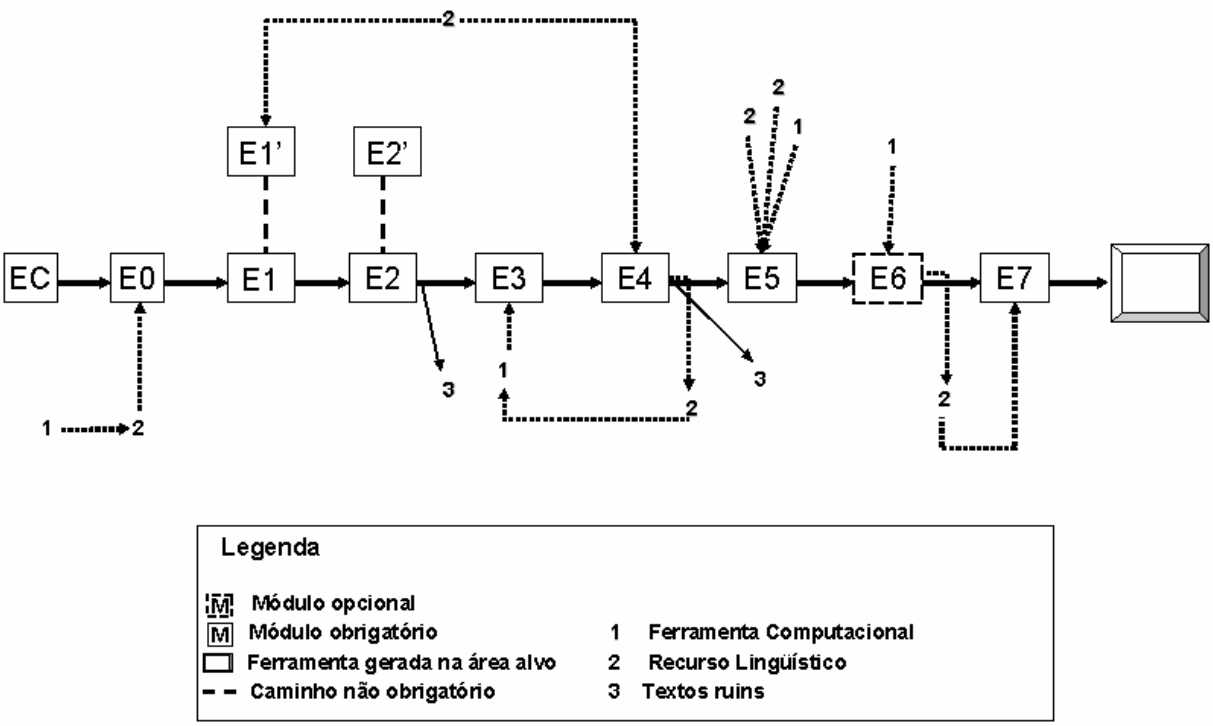


Figura 4.1: Diagrama da proposta de um processo para a geração de recursos lingüísticos aplicáveis em ferramentas de suporte à escrita, composto por 11 etapas. Conforme pode ser observado, esse processo descreve as etapas (passos) para se construir um córpus, extrair recursos lingüísticos dele, acoplá-los adequadamente em um servidor, para, então, obter uma ferramenta de auxílio à escrita científica. Esse processo inicia-se com a Etapa EC, segue por E0, E1 ou E1', E2 ou E2', E3, E4, E5 e E7. O usuário de nosso processonão precisará realizar as etapas de E6, porque E6 ainda não está concuído.

A Etapa EC - *Etapa de compilação de córpus* - fornece diretrizes desde o momento de escolha das fontes de coleta dos textos para compor um córpus até o momento de armazenamento dos textos.

A Etapa E0 - Etapa de Balanceamento das seções de artigos científicos – recebe como entrada um dado córpus e, a partir de informações relevantes retiradas do mesmo, avalia-se seu balanceamento.

A Etapa E1 - *Etapa de Anotação Automática da Estrutura Esquemática* – cada seção de artigo científico é submetida a uma ferramenta computacional (categorizador automático), que irá detectar automaticamente os componentes da estrutura esquemática contidos em uma seção, seja ela a introdução, a metodologia, a conclusão e assim por diante.

A Etapa E1'- *Etapa de Anotação Manual da Estrutura Esquemática* – é utilizada caso não seja possível ou desejável utilizar o auxílio automático do categorizador citado em E1. Em E1' estão as etapas necessárias à tarefa de anotação manual da estrutura esquemática de cada seção textual de um córpus de artigos científicos.

A Etapa E2 - *Etapa de Avaliação Automática de Qualidade de Escrita* – nessa etapa, há a submissão dos textos a uma ferramenta computacional, que avalia automaticamente a qualidade textual de um córpus, segundo critérios específicos de qualidade, separando-se, assim, as seções de artigos científicos em "boas e "ruins". É importante notar que o conteúdo científico de cada seção de artigo científico não é avaliado, uma vez que é assumido ter conteúdo científico adequado, dada a sua publicação em fóruns de excelência.

A Etapa E2' - *Etapa de Avaliação Manual de Qualidade de Escrita* – nessa etapa, temos a avaliação manual da adequação das seções de um artigo científico com o auxílio de uma rubrica particular a avaliação de cada seção.

A Etapa E3 - *Etapa de Anotação Automática de Marcadores Discursivos e Expressões Formulaicas* – nessa etapa, as seções de artigos científicos são submetidas a uma ferramenta computacional, que detecta automaticamente padrões lexicais reutilizáveis, como os marcadores discursivos e as expressões formulaicas.

A Etapa E4 - *Etapa de Revisão Manual e Parcial da Qualidade dos textos, da Estrutura Esquemática e dos Marcadores Discursivos* – essa etapa manual visa corrigir eventuais falhas cometidas pelos processos automáticos, bem como fornecer insumo para as ferramentas computacionais melhorarem sua precisão.

A Etapa E5 - *Etapa de Anotação Manual das Estratégias Retóricas* – é uma etapa manual de anotação das possibilidades de realização lingüística de cada componente da estrutura esquemática, contidos em uma seção de artigo científico específica.

A Etapa E6 - *Etapa de Extração Automática de Termos* – futuramente, nessa etapa, será efetuada a submissão de artigos científicos a uma ferramenta computacional, que fará a extração dos termos específicos da área a qual o córpus pertence. A lista desses termos será

submetida a um concordanciador, que os apresentará em seu contexto de uso. Assim, o usuário desse concordanciador poderá observar de que forma os termos importantes de sua área devem se apresentar organizados.

A Etapa E7 - *Etapa de Inclusão dos Recursos Lingüísticos do processo em uma ferramenta de suporte à escrita genérica* – dá-se a formatação de todos os recursos lingüísticos, produzidos ao longo das etapas, e a inclusão desse conhecimento em uma ferramenta genérica, isto é, sem uma base de recursos lingüísticos, o Scientific Writing.

Todas essas etapas e procedimentos são descritos com mais detalhes nos próximos tópicos deste trabalho, a partir da Seção 4.3.

Antes de partirmos para essa descrição com detalhes, é interessante notar que as etapas E1 e E1' realizam quase que a mesma tarefa. A etapa E1' faz, além da anotação das estratégias retóricas, a anotação da estrutura esquemática de um córpus de uma seção de artigo, com a diferença de que o primeiro a realiza de forma automática, e o segundo, manual. A abordagem automática é extremamente interessante para agilizar a tarefa, embora tenha uma precisão ainda longe de ser a desejada.

Da mesma forma, as etapas E2 e E2' também realizam a mesma tarefa de avaliação da qualidade dos textos – a primeira de forma automática; e a segundo, manual.

Assim, a Figura 4.1 apresenta vários caminhos para o processo de construção de recursos lingüísticos aplicáveis em ferramentas de suporte à escrita, geradas com nosso processo:

1) se a etapa E1 for escolhida, as seguintes etapas a serem utilizados são: E2 ou E2´, E3 a E7
2) se a etapa E1' for escolhida, as seguintes a serem utilizados são E2 ou E2', E3, somente parte de E4 (a estrutura esquemática não precisa ser revista), E6 e E7.

4.3 Etapa EC – Etapa de Compilação de corpus

4.3.1 Instruções para a realização da Etapa EC

Nessa etapa, é apresentada uma descrição de como se construir um córpus, passo de considerável importância dentro de todo o processo, uma vez que o córpus constitui o núcleo das ferramentas geradas com nosso processo, ou seja, a parte principal da qual serão retirados os recursos lingüísticos.

A primeira e crucial etapa que precede a compilação de um córpus é o questionamento acerca do conjunto de critérios que determinarão a coleta dos textos, definidos com base nos propósitos segundo os quais o córpus será compilado (Sinclair, 1991; Atkins, Clear, Ostler, 1992; Biber, 1993; Quirck, 1992; Kennedy, 1998; Biber, Conrad e Reppen, 1998). Entre esses questionamentos, podemos citar qual o tipo de pesquisa que utilizará o córpus a ser compilado, quais as possíveis fontes para coleta de textos pertinentes ao córpus, qual o tamanho ideal desse córpus, que tipo de textos farão parte desses córpus, e assim por diante. Uma discussão dos critérios mais relevantes que devem ser considerados pode ser encontrada em Renoulf (1984), por exemplo.

Essas perguntas, quando previamente formuladas, possibilitam: (1) planejar o custo e os esforços necessários para a viabilização da compilação de um córpus; (2) estruturar o córpus de maneira adequada às necessidades do estudo pretendido e, acima de tudo, (3) uma economia de tempo, já que o córpus produzido com planejamento prévio servirá corretamente aos propósitos requeridos pelo estudo. A realização desse planejamento prévio é aconselhada tanto em trabalhos que visam à construção de um grande córpus de referência como nos trabalhos com um pequeno córpus de especialidade, pois, segundo Sinclair "(...) as decisões tomadas sobre o que inserir em um corpus e como a seleção deve ser organizada; [essas decisões controlam quase tudo que acontece subseqüentemente" (Sinclair, 1991:13)[1].

No nosso caso, o objetivo de se construir um córpus é o de utilizá-lo na extração de recursos lingüísticos aplicáveis em uma ferramenta de auxílio à escrita científica em língua inglesa. Assim, critérios como escolher textos publicados em revistas bem conceituadas, escolher textos escritos por nativos do inglês, variar os autores dos textos coletados para que não se corra o risco de obter exemplos de um ou pouco estilos de escrita, etc., podem ser critérios interessantes para aumentar (mas não garantir) a boa qualidade dos recursos lingüísticos a serem construídos.

A próxima seção indica qual o primeiro passo que se deve dar para a coleta de textos, e como essa coleta deve ser feita.

[1](…) the decisions that are taken about what is to be in the corpus, and how the selection is to be organized; [these decisions] control almost everything that happens subsequently. (Sinclair, 1991:13)

4.3.1.1 Estudo da área de especialidade e posterior elaboração de uma árvore de domínios dessa área

Conforme pode ser observado na literatura especializada, a ausência do conhecimento sobre a área de especialidade do córpus, suas subáreas e sua classificação hierárquica podem influenciar o balanceamento e a representatividade do material compilado (Pardo, 2004).

Portanto, o primeiro passo que é sugerido antes de se realizar uma pesquisa a respeito das possíveis fontes para a coleta de textos é o conhecimento da área de especialidade em que esses textos se encontram inseridos, assim como o de sua abrangência. Um possível modo de se obter tal conhecimento é por meio da pesquisa e obtenção de uma árvore de domínios.

Mas qual é a função de uma árvore de domínios? A árvore de domínios auxilia a estabelecer uma estrutura de organização interna do córpus, pode servir como guia na busca de textos para a construção do córpus e pode também auxiliar a nomeação dos textos a fim de facilitar sua posterior identificação e consulta. É importante também decidir o nível de estratificação[2] adequado de uma árvore de domínios, pois uma hierarquia interna muito estratificada pode vir a dificultar a busca e a coleta de textos no momento de construção do córpus, bem como dificultar a sua consulta e posterior utilização.

É importante dizer que a organização de uma árvore de domínios de uma área já é motivo suficiente para gerar discussão e divergência de opinião entre os especialistas quanto à organização e limites das áreas e subáreas de uma dada árvore. No entanto, o que se pretende com essa pequena árvore de domínios é uma simples organização do material escolhido a fim de se obter um córpus balanceado quanto ao número de textos divididos entre as subáreas de determinado domínio.

Para tanto, podemos pesquisar em sites acadêmicos e de apoio à pesquisa científica, como o do Cnpq (www.cnpq.br) e o da CAPES (www.capes.gov.br), que fornecem esse tipo de árvore.

A estrutura da árvore de domínios será utilizada na etapa E0, quando relatarmos o procedimento de verificação do balanceamento do córpus em construção. Antes de se realizar essa verificação, é necessário possuir um conjunto de textos. Assim, as seções seguintes discutem os diferentes tipos de fontes existentes para a coleta de textos, bem como as formas como essa coleta pode ser feita.

[2] Estratificação: nível de subdivisão, ramificação da árvore de domínios.

4.3.1.2 Fonte e Coleta de textos para a composição de um córpus

A construção de um córpus em formato eletrônico e computável, que permita sua disponibilização em *sites* ou outros meios de comunicação eletrônica, como também a manipulação de seus dados por ferramentais computacionais específicos (concordanciadores, estudos de *n-gramas*, etc.), é uma tendência que não pode ser ignorada, pois possibilita as seguintes vantagens:

a) Integração científica: a disponibilidade de ferramentas capazes de manipular dados em formato eletrônico é uma evidência de que a Lingüística, cada vez mais, busca auxílio tecnológico para o trabalho repetitivo, poupando tempo e recursos do pesquisador.

b) Organização do córpus: os métodos de coleta e armazenamento propostos pela Lingüística de Córpus permitem a geração de córpus altamente organizados, que podem ser disponibilizados na Web para consultas;

c) Processamento dos dados: existe uma grande quantidade de programas que podem ser utilizados por lingüistas, poupando tempo.

Como já visto, assumimos que um córpus deva ser construído em formato eletrônico, independente do propósito principal que motivou sua construção. Assim, o problema do pesquisador passa a ser como fazer isso. Existem vários meios e o mais comum seria a coleta de textos diretamente disponíveis na *Web*. O procedimento é bastante simples e só necessita de comandos básicos que podem ser realizados no próprio ambiente de interface do *browser* (navegador, que são mais conhecidos por seus nomes comerciais, como *Internet Explorer*, por exemplo). Ao pesquisar pela Internet, o usuário geralmente utiliza um navegador que possui, no alto da janela, um menu ou barra de ferramentas (uma linha com botões *arquivo*, *editar*, *exibir*, etc).

Escolhido um texto para a compor o córpus, basta que o usuário clique em *arquivo* e escolha a opção *salvar como*. Irá aparecer uma janela na qual ele deverá indicar em que pasta o arquivo será salvo, devendo indicar, também, qual o formato em que esse arquivo deverá ser salvo. O diretório (pasta) no qual serão armazenados os textos já deverá ter sido previamente criado na área de trabalho do pesquisador de acordo com os critérios escolhidos para a organização. O formato do arquivo deverá ser, no mais das vezes, do tipo .txt (arquivo de texto sem formatação), que permite sua utilização por parte de grande maioria das ferramentas computacionais disponíveis.

Essa situação de coleta, no entanto, não é a única existente. Não raro, o pesquisador se defronta com a necessidade de elencar como córpus conjuntos de textos não disponíveis na

Internet. Nesse caso, uma saída viável é a digitalização dos textos através do uso de scaners e programas *de reconhecimento ótico de caracteres (OCR)*, que transformam a imagem coletada pelo scaner em arquivos de texto manipuláveis por computador. Deve-se estar atento para o fato de que a mera digitalização do texto por um scaner, sem o uso dos referidos programas, gera tão somente uma foto do documento, que não poderá ser utilizada como entrada em editores de textos, por exemplo. Frente a isso, foram criados *softwares* que podem oferecer 99,9% de precisão no reconhecimento de caracteres, como é o caso do *Recognita Corp*[3] (*Recoginta 5 Plus*) e do *Image Recognition Integrated Systems*[4] *(ReadIris)*. Mas esse fato não descarta a necessidade de revisão do produto gerado, uma vez que não devem ser adicionados ao córpus textos que contenham erros de codificação de caracteres. Esses processos de digitalização de documentos são eficientes, quando a qualidade de impressão é boa. No entanto, o estado físico dos textos nem sempre pode estar em boas condições. Nesses casos, a única solução possível é a digitação do texto no computador, apesar de ser um método custoso e demorado.

A escolha entre um ou outro método vai depender muito da natureza do projeto, da qualidade dos materiais disponíveis e das fontes disponíveis para a compilação. Mas mesmo os processos que envolvem o scaner e o *software* OCR requerem uma revisão para assegurar uma boa caracterização do texto, pois apenas a boa revisão dos textos convertidos manualmente ou semi-automaticamente para o formato digital é que pode promover precisão e conformidade das fontes geradas com os textos originais (Hockey, 1998:107).

Independente do método adotado para a obtenção dos textos, o desejável é que o córpus coletado seja diversificado, ou seja, que as fontes de coleta e os autores dos textos sejam variados. Com um número reduzido de fontes pode-se correr o risco de obter um número maior de textos de uma área em detrimento de outra, ou vários textos de mesma autoria, com uma terminologia altamente recorrente, especificamente utilizada pelo autor em particular e não por todos os autores da mesma área.

Ao coletar os textos devemos estar atentos também para que não haja duplicações. A nomeação do arquivo é uma maneira de se evitar a duplicação, pois no nome estariam explícitas determinadas informações-chave sobre o texto escolhido, as quais poderiam evitar com que o mesmo pudesse ser salvo mais de uma vez.

[3] Informações adicionais em http:// www.caere.com/recognita
[4] Informações adicionais em http://www.irislink.com/UK/index.html

4.3.1.3 Direitos autorais

As pessoas que desejam compilar um córpus são obrigadas a se assegurarem legalmente sobre a utilização dos textos coletados para sua pesquisa. Se o texto foi publicado há muito tempo, há a possibilidade dos direitos autorais terem expirado. Esse tempo exato de validade dos direitos varia de país para país, fato que precisa ser conferido de acordo com o local em que se pretende desenvolver a pesquisa.

Portanto, o primeiro passo para o pesquisador que pretende coletar textos via Internet é se informar detalhadamente sobre os direitos de posse que estão relacionados a um texto em particular ou ao editor do mesmo. Uma vez identificado o proprietário dos direitos autorais, este deve ser consultado para saber se há a possibilidade de utilizar o(s) texto(s) para fins de estudo científico. A melhor política para se realizar esse tipo de pedido é destacando e dando detalhes do papel importante que ele(s) ocupa(m) na pesquisa. Assim, são apresentadas as reais e boas intenções em se utilizar os textos dos autores. Outro fator que deve ser considerado é o agradecimento formal por meio de citações em produtos originados do córpus, dos eventuais editores/autores que colaboraram fornecendo textos para o córpus.

Com esses cuidados, aumenta-se a chance de se obter uma resposta afirmativa sobre a concessão de textos, especialmente se estiver destacado que o córpus não visa a uma reprodução de seu conteúdo científico[5].

Com relação à carta para pedido de autorização de uso de texto(s), no site do NILC, por exemplo, mais precisamente no link do projeto Lacio-Web[6], pode ser encontrada uma carta endereçada aos autores dos textos utilizados por esse projeto. Nesse mesmo site, por exemplo, pode ser consultada uma lista com os nomes das pessoas que contribuíram com suas produções escritas na composição dos córpus produzidos no Lacio-Web.

4.3.1.4 Edição de textos

Conforme discutido anteriormente, os textos coletados devem ser, preferencialmente, salvos em formato .txt, exigido por ferramentas computacionais que extraem dados de textos, como é o caso do *WordSmith Tools*, que é utilizado em nosso córpus.

---

[5] Há um site brasileiro no qual o pesquisador interessado em coletar textos via WEB pode encontrar mais informações sobre plágio e direito autoral na Internet no Brasil: http://www.persocom.com.br/brasilia/plagio1.htm.

[6] Site do projeto Lacio-Web: http://www.nilc.icmc.usp.br/lacioweb

Existem gratuitamente na Web conversores do formato .pdf para o .txt (*ABC PDF Converter 1.0; ABC Amber Txt Converter 2.16*, por exemplo). Para os casos em que é preciso transformar um arquivo .doc para um .txt., basta abrir o texto no editor *Word*, da *Microsoft* por exemplo, clicar na opção *Arquivo*, *Salvar Como*, *Salvar como Tipo* e selecionar a opção *Texto sem Formatação*.

Uma vez convertidos para o formato .txt, é preciso realizar a edição desses textos convertidos, porque esse tipo de formato não permite que informações não textuais sejam salvas. Portanto, será necessário isolar, por exemplo, com etiquetas em linguagem *extended markup language* (XML),[7] elementos como tabelas, fórmulas, quadros, figuras, etc., visto que elas não fazem parte do corpo do texto, mas não podem ser retiradas por trazerem informações e dados da pesquisa realizada, como também pelo fato de ser importante conservar a estrutura original do texto fonte.

4.3.1.5 Criação de cabeçalhos

Os cabeçalhos são, segundo Berber-Sardinha (2004:145), trechos demarcados contendo informação não veiculada verbalmente no evento comunicativo, que fornecem detalhes acerca de, por exemplo, proveniência, tipologia e autoria dos textos. Em outras palavras, corresponde à seção superior do arquivo.txt, na qual, utilizando-se linguagem XML, por exemplo, poderão ser inseridas informações extratextuais como o endereço eletrônico do texto escolhido, o nome do(s) autor(es), etc. Para tanto, poderão ser seguidas as normas do TEI[8] (*Text Enconding Initiative*), sempre consultadas, em geral, pelos projetos de córpus.

O cabeçalho poderá ser inserido manualmente ou por meio de uma ferramenta computacional que faz a inserção das informações do cabeçalho de maneira semi-automática, em XML, como o fizeram os anotadores dos córpus do projeto Lácio-Web, com o auxílio de um editor de cabeçalhos. A utilização de tal auxílio se justifica na tentativa de se evitar que sejam cometidos erros de digitação no momento de inserção de informações no cabeçalho dos textos, bem como agilizar esse tipo de processo que deve ser realizado em linguagem XML, o qual requer muita atenção na colocação dos caracteres. Para tanto, os anotadores necessitavam apenas preencher campos cujas informações já estavam previamente inseridas em um campo

[7] Informações sobre padrões internacionais de codificação e anotação de córpus: XCES (Córpus Enconding Standard for XML), http://www.cs.vassar.edu/XCES/; CES (Córpus Encondig Standard): http://www.cs.vassar.edu/CES/; EAGLES (Expert Advisory Group for Language Engineering Standards): http://www.ilc.cnr.it/EAGLES96/home.html. XML é um tipo de linguagem utilizada para identificar informações em um dado texto.

[8] *Site* com normas internacionais para a criação de cabeçalhos: http://www.tei-c.org/Guidelines2/index.html.

com barra de rolagem. Com apenas um clique, as informações eram inseridas e, ao final, quando salvos, os textos saiam já em formato XML.

No caso de se optar pela inserção manual de cabeçalho, as informações devem ser digitadas cuidadosamente, pois um erro de digitação ou de sintaxe da linguagem XML poderá comprometer a extração de dados em uma futura análise com o córpus.

A inserção de cabeçalho em um córpus possibilita vantagens, como: preservar informações importantes sobre os textos e colocá-las imediatamente à disposição dos usuários como também auxiliar o computador na localização de textos específicos. Vários programas que auxiliam na investigação de córpus, como, por exemplo, o *WordSmith Tools*, lêem cabeçalhos, permitindo ao usuário escolher os tipos de textos com os quais deseja trabalhar. Daí a importância em não se utilizar esquemas de cabeçalhos caseiros, mas sim os já institucionalizados internacionalmente.

4.3.1.6 Nomeação dos textos

Tarefa aparentemente simples, mas que requer certo cuidado e coerência em seu desenvolvimento, é a nomeação dos textos do córpus. Assim como os diretórios que contém os textos devem refletir em seu nome o seu interior, os nomes dos textos já editados e formatados também devem ter esse tipo de funcionalidade.

4.3.1.7 Organização do córpus

Uma vez coletados, formatados e nomeados, o próximo passo é a organização dos arquivos em uma estrutura coerente e de fácil manuseio. Segundo a literatura, não há regras para esse tipo de procedimento. Alguns córpus vêm organizados em pastas hierarquizadas, outros com textos salvos em arquivos separados, outros ainda com um texto em cada pasta, por exemplo. Entretanto, há recomendações que, segundo Berber-Sardinha (2004:72), são importantes e devem ser consideradas em qualquer tipo de organização de córpus:

1ª - Os textos devem estar em uma pasta principal em que só existam textos do córpus.

2ª - Seja criada uma subpasta que indique a versão atual do córpus, por exemplo, 00.

3ª - As subpastas criadas devem refletir seu conteúdo, isto é, que tenham nomes que indiquem o tipo de texto, o assunto, etc...

4.3.1.8 Aproveitamento de diferentes partes de um mesmo artigo científico

Ainda em relação à atividade de organização do córpus, outra questão importante a ser citada diz respeito ao aproveitamento de diferentes partes de um mesmo artigo. Mais explicitamente, quando se tem em mãos um texto (artigo científico em inglês) que se deseja adicionar ao córpus, é preciso salvar as seções desse artigo, ou seja, as partes Resumo, Introdução, Metodologia e assim por diante, em arquivos diferentes, mesmo sendo partes de um mesmo texto. Isto é, no documento R1.doc, por exemplo, poderá ser salvo o resumo do primeiro artigo coletado para o córpus. No documento I1, poderá ser salva a Introdução desse mesmo primeiro artigo coletado, e assim por diante. Isso porque, ao longo do processo de extração dos recursos lingüísticos desses textos, cada seção constitutiva de um artigo científico é trabalhada em separado. Em outras palavras, para que os recursos sejam gerados de maneira adequada, é preciso que cada seção de um artigo tenha seus recursos lingüísticos gerados em separado das outras seções. Primeiro, por exemplo, são extraídos os recursos existentes na seção "Resumo". Posteriormente, podem ser extraídos os recursos, por exemplo, contidos na seção "Conclusão". Mas sem deixar de lembrar que devem ser extraídos todos os recursos de uma seção de artigo antes de iniciar a extração de recursos de uma outra seção qualquer.

4.3.2 Instanciação da Etapa EC

As idéias apresentadas até o momento são importantes; no entanto, sua validade se mostra ainda mais eficaz quando apresentadas em forma de exemplo ou estudo de caso. Para tanto, serão apresentadas as etapas que envolveram a coleta e armazenamento de um córpus utilizado na seção "Metodologia" do SciPo-Farmácia, um dos ambientes inspiradores deste estudo e que serviu de estudo de caso para a elaboração da proposta deste trabalho. A construção desse córpus, córpus Met doravante, realizada pela autora desta pesquisa, foi motivada primeiramente pela necessidade de se descrever em detalhes a construção e anotação de um córpus requerido por ambientes de auxílio à escrita nos moldes do Scipo-Farmácia, uma vez que faz parte deste projeto de estudo possibilitar que pessoas não conhecedoras da área de PLN e de Lingüística de Córpus sejam capazes de construir os recursos lingüísticos necessários na geração desse tipo de ferramenta computacional e gerar uma ferramenta de suporte à escrita personalizada para sua área de pesquisa. Para tanto, buscamos embasamentos em situações e dificuldades reais encontradas durante o processo de

construção do córpus citado, de modo a promover uma descrição mais adequada para o usuário do processo proposto por este projeto de mestrado. Essa descrição dos procedimentos envolvidos incluem detalhes sobre os critérios utilizados, dificuldades e procedimentos envolvidos na tarefa de compilação e extração de recursos lingüísticos que geraram o córpus Met. Como resultado colateral dessa atividade, obtivemos a implementação da única seção de auxílio à escrita que ainda não estava implementada no SciPo-Farmácia, a seção "Metodologia" de artigos científicos da área em questão.

Como o córpus de nosso estudo de caso precisava ser da área de Ciências Farmacêuticas, foi necessário compor uma árvore de domínios dessa área citada. Para isso, foram consultados dois especialistas experientes, que são docentes do curso de pós-graduação em Ciências Farmacêuticas da USP-São Paulo. Apesar de haver divergências entre esses especialistas quanto à estratificação da grande área, eles chegaram a um consenso, que resultou na seguinte estrutura:

Árvore de Domínios gerada para a área de Ciências Farmacêuticas

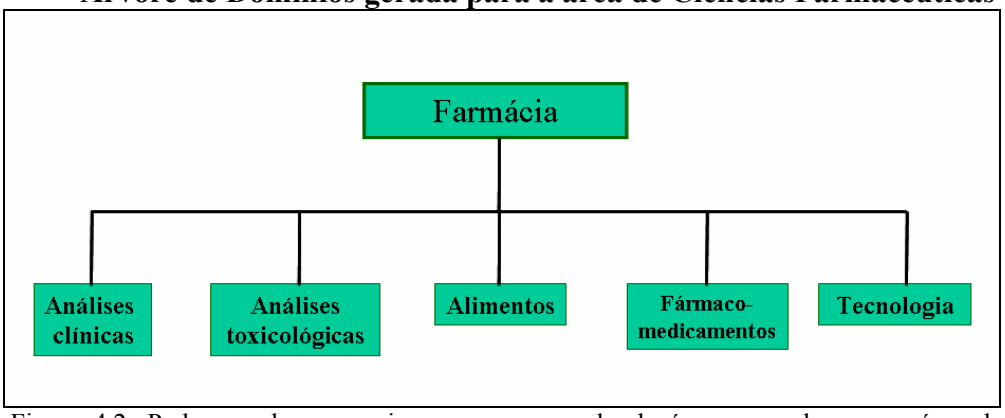


Figura 4.2: Podemos observar acima, que o exemplo de árvore gerada para a área das Ciências Farmacêuticas apresenta uma estruturação simples, com ramificação em apenas um nível: a grande área isolada e localizada no topo do organograma com suas cinco subáreas a ela ligadas. Há que se considerar que se outros especialistas a tivessem elaborado, possivelmente essa estrutura teria diferenças.

No caso da construção do córpus Met, foi de fundamental importância consultar especialistas da área de Farmácia para se chegar a uma árvore de domínios condizente com a área de especialidade que precisava ser representada, uma vez que a especialista responsável pela construção desse córpus não pertence à área em questão. No entanto, como o usuário do CECARL vai construir recursos lingüísticos na área em que atua, tal consulta é opcional, uma vez que esse usuário possui condições de organizar uma árvore de domínio, mesmo que simples, da área em que atua.

Sobre a coleta de textos para compor o córpus Met, 30 seções Metodologia da área de Ciências Farmacêuticas, podemos dizer que sua obtenção foi de certa forma facilitada, pois essas seções foram retiradas de fontes on-line de divulgação científica (Tabela 4.1).

Texto	Fonte	Texto	Fonte
Met_01	PubMed	Met_16	PubMed
Met_02	PubMed Central	Met_17	PubMed
Met_03	PubMed	Met_18	JBC
Met_04	Journal of Biological Chemistry	Met_19	PubMed
Met_05	PubMed	Met_20	PubMed
Met_06	PubMed	Met_21	Elsevier
Met_07	Chemical Engineering Journal	Met_22	Elsevier
Met_08	Pharmaceutical Research	Met_23	Elsevier
Met_09	Elsevier	Met_24	Elsevier
Met_10	PubMed	Met_25	Elsevier
Met_11	PubMed	Met_26	ACS
Met_12	PubMed	Met_27	JBC
Met_13	Nature	Met_28	Elsevier
Met_14	Nature	Met_29	ACS
Met_15	PubMed	Met_30	Elsevier

Tabela 4.1: As fontes on-line apresentadas na tabela foram escolhidas por serem locais de divulgação de reconhecida importância por parte da comunidade acadêmica das Ciências Farmacêuticas.

Essas 30 seções de metodologia apresentadas na Tabela 4.1 foram construídas sob orientação de dois especialistas da área, os mesmos que decidiram a organização da árvore de domínios anteriormente citada, ou seja, depois de escolhidos, os textos foram lidos e aqueles que não eram pertinentes ao córpus foram excluídos. Antes de se chegar a esse número de textos, outros mais foram coletados e descartados por serem de má qualidade (conteúdo científico), segundo esses mesmos especialistas. É importante dizer que se primou pela escolha de textos produzidos por nativos da língua inglesa, uma vez que se tem aumentada (mas não garantida!) a probabilidade de uma escrita adequada em língua inglesa. Também houve preferência pela coleta de textos com grande impacto na área, ou seja, aqueles que são bem citados e tidos como referência por outros pesquisadores.

A compilação desse córpus Met consistiu em retirar os textos de sua fonte original (Internet), armazená-los em um diretório no qual pudessem ser manipulados adequadamente, convertendo-os para o padrão texto sem formatação (salvando-os com a extensão .txt), sempre

respeitando o texto original. Mais especificamente, tal compilação foi efetuada conforme os seguintes procedimentos:

1. acesso à página de Qualis da Capes, http://qualis.capes.gov.br/;
2. seleção da opção *Área de Avaliação (Farmácia)* e, em seguida, da *Classificação (A)*;
3. como resultado da ação anterior, os nomes dos periódicos com avaliação "A" aparecem na tela;
4. copia do nome de um periódico para um site de busca, como por exemplo o www.google.com.br
5. acesso ao site do periódico com links de artigos;
6. seleção do trecho de interesse (seção "Metodologia" de artigos científicos, no caso de compilação do córpus Met). No entanto, nesse momento, o usuário do CECARL poderá selecionar um artigo completo e, posteriormente, no momento de gravação desse, salvar as seções do mesmo em arquivos separados. Uma vez que as tarefas descritas para a extração dos recursos lingüísticos requerem que as seções de artigos científicos sejam investigadas separadamente. Em outras palavras, de um artigo científico de uma área qualquer, o usuário do CECARL poderá obter, por exemplo, 6 textos salvos: um com a introdução do artigo, outro com o resumo, um terceiro com a metodologia, um quarto com resultados e outros dois com as discussões e as conclusões, todos retirados de um mesmo artigo;
7. cópia do texto de interesse (o processo de recuperação dos textos da Internet é o padrão "copiar e colar");
8. em uma página do editor de texto *Microsoft Word*, por exemplo, pode-se colar o conteúdo copiado da Internet. Nessa operação, certas informações, além de formatação específica, podem ser perdidas. Nesse caso, o texto fonte (site) foi mantido aberto, o que facilitou na visualização e posterior identificação dos caracteres problemáticos;
9. quebra da linha entre os parágrafos do texto, deixando espaço de uma linha em branco. Observação: As quebras de parágrafo respeitam a paragrafação do texto-fonte. Para a execução desta etapa, mantivemos, novamente, o texto fonte (site) aberto;
10. conversão para o formato texto sem formatação (extensão .txt) utilizando o editor *Microsoft Word*. O objetivo da escolha do formato .txt é para permitir o tratamento computacional (avaliação da concordância da anotação do córpus com vários anotadores via estatística *Kappa* (mais detalhes na Etapa E1'), como também manuseá-lo com o auxílio da ferramenta *WordSmith Tools,* que requer um arquivo nesse formato.

Na transferência dos textos do córpus Met do formato .doc para o .txt, houve problemas com os seguintes caracteres:

- Potências: o texto salvo em formato .txt não possibilita a elevação de potências. Dessa maneira, optamos pela inserção de um acento circunflexo antes da escrita do número que seria elevado a potência, por exemplo, 10^2 foi substituído por 10^2.

- Letras gregas e sinais matemáticos: não foram possíveis de serem salvos no formato requerido e foram substituídos por sua forma em extenso: ≥ <maior ou igual>, alfa, teta, lâmbda, mi, e assim por diante.

11. Depois de formatados, os textos foram nomeados. Ex: Met_01, o que significa dizer que se trata de um texto da seção "Metodologia" (Met_) e que se refere ao primeiro texto que compõe o córpus construído.

Quanto à autoria dos textos contidos no córpus Met, os mesmos são citados no cabeçalho de cada texto autorizado, uma vez que os textos foram retirados de sites *on-line* de divulgação científica (conforme mostra a Tabela 4.1), para os quais as submissões são precedidas por um termo de autorização. Nesses sites, depois de submetidos, os textos são disponibilizados livremente para o uso, sem a necessidade de se pedir nova autorização para tal. No entanto, a ressalva por eles feita é a de que os autores sejam citados sempre que seus trabalhos forem utilizados em outras pesquisas.

Para o isolamento dos dados extratextuais do córpus Met, utilizamos as seguintes etiquetas[9].

<figura> <tabela1> <formula1>

Dessa maneira, a ferramenta computacional escolhida para a extração de dados do córpus poderá incluir ou excluir esses elementos extratextuais na seleção de dados que serão analisados. Após o isolamento dos dados extratextuais, é aconselhado uma inserção de informações sobre o texto no próprio texto, utilizando-se para isso um cabeçalho. No córpus Met, optamos pelo seguinte cabeçalho:

[9] As etiquetas são pequenos trechos inseridos no corpo do texto, demarcados por símbolos específicos, por exemplo, <Autor=Dimeinstein>, como identificador do autor do texto (Berber-Sardinha, 2004:145).

1) Link para a versão original e completa do artigo coletado, uma vez que utilizamos apenas as seções metodologia de cada um deles. Sempre que necessário ou desejado, a versão original poderá ser consultada.
2) Título do artigo, o qual dá identidade e também dá dicas sobre o conteúdo da seção "Metodologia".
3) Autores do artigo, respeitando a autoria dos textos coletados e possibilitando que os textos possam, se desejado, ser reunidos e/ou separados por autores.

É importante dizer que, para os propósitos de construção do córpus Met, esse cabeçalho, apesar de simples, satisfaz as necessidades do projeto. Se o usuário de nosso processo desejar adicionar mais informações sobre os textos que coletou, poderá fazê-lo sem problemas, desde que para isso leve em consideração os padrões convencionalizados para realizar esse tipo de tarefa, os quais possibilitam que o cabeçalho padronizado seja (re)utilizado por diferentes ferramentas computacionais. Tão importante quanto a criação de um cabeçalho e a organização dos textos em diretórios que reflitam seu conteúdo, é a nomeação dos textos.

Em nosso córpus Met, a nomeação seguiu a seguinte padronização: Met_01, Met_02, Met_03 ... Met_30. O que significa, respectivamente: primeiro texto da seção "Metodologia", segundo texto, terceiro... e o qüinquagésimo texto da seção "Metodologia".

Há um momento dentro do processo de extração dos recursos lingüísticos do córpus, mais precisamente no momento de verificação do balanceamento do córpus (Seção 4.4), em que esses textos poderão sofrer alteração de seus nomes, se desejável, pois poderão ser organizados segundo a subárea a que pertencem.

Em relação aos procedimentos de organização e armazenamento dos textos coletados para o córpus Met, foram construídos os seguintes diretórios: criação de uma pasta (diretório) com o nome do córpus "córpus Met"; criação de subpastas, dentro de córpus Met, com informações sobre o córpus do tipo: pasta com córpus anotado, pasta com córpus sem anotação, etc. Dentro de córpus anotado, por exemplo, se encontram as pastas com nomes de seus respectivos anotadores (Figura 4.3).

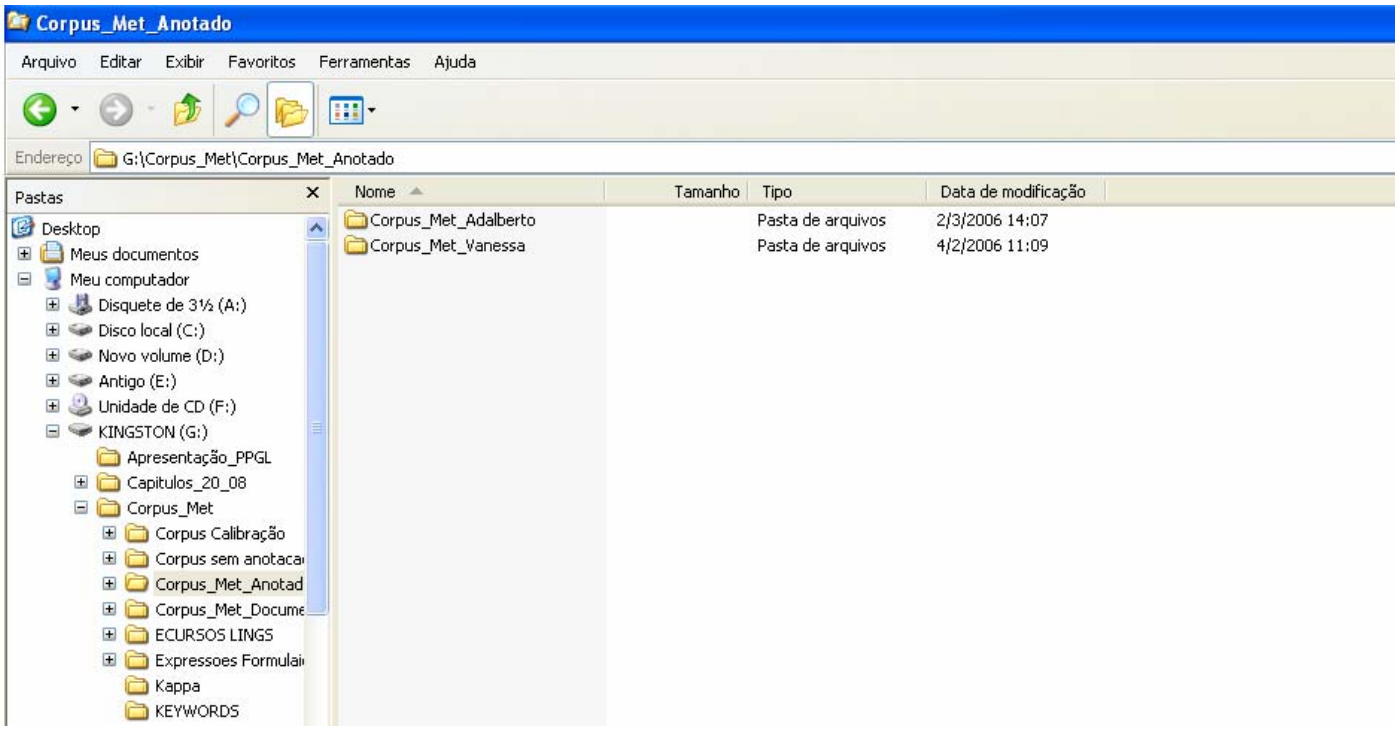


Figura 4.3: Estruturação de diretórios utilizada na organização do córpus Met.

Como pode ser observado na proposta acima, há um grande diretório intitulado Córpus_Met. Em seu interior há diferentes subpastas com conteúdos a ela relacionados: o Córpus de Calibração utilizado para familiarizar os anotadores (qunado houver mais de uma pessoa anotando os textos) quanto às categorias que deveriam utilizar para marcar o córpus; há uma pasta que contém o córpus cru, isto é, sem anotação; há outra pasta com o nome de Córpus_Met_Anotado, que possui em seu interior subpastas com o nome dos respectivos anotadores. Na pasta Córpus_Met_Documentação, há as versões do manual de anotação, e as tabelas que comparam a anotação realizada pelos diferentes anotadores.

4.4. Etapa E0 – Etapa de Balanceamento das seções de artigos científicos coletados

Trata-se de uma etapa, na qual, a partir de um dado córpus formado por seções "Metodologia", por exemplo, são extraídas informações referentes ao tipo de conteúdo de cada seção de artigo científico e, posteriormente, é feita a distribuição dessas seções sob a árvore de domínios construída na etapa anterior, a EC. A partir dessa distribuição, é feita uma avaliação da distribuição equivalente/balanceada das seções em cada subárea.

Vale ainda dizer, que pode ser possível que uma pessoa queira construir recursos lingüísticos de uma única subárea de especialidade. Nessa situação, não é necessário realizar o balanceamento, uma vez que haverá textos de um único ramo (subárea) desse conhecimento.

A Figura 4.4 apresenta uma visão geral dessa etapa, a qual recebe como entrada um córpus não-balanceado e produz como saída, informações relevantes para a avaliação de seu balanceamento.

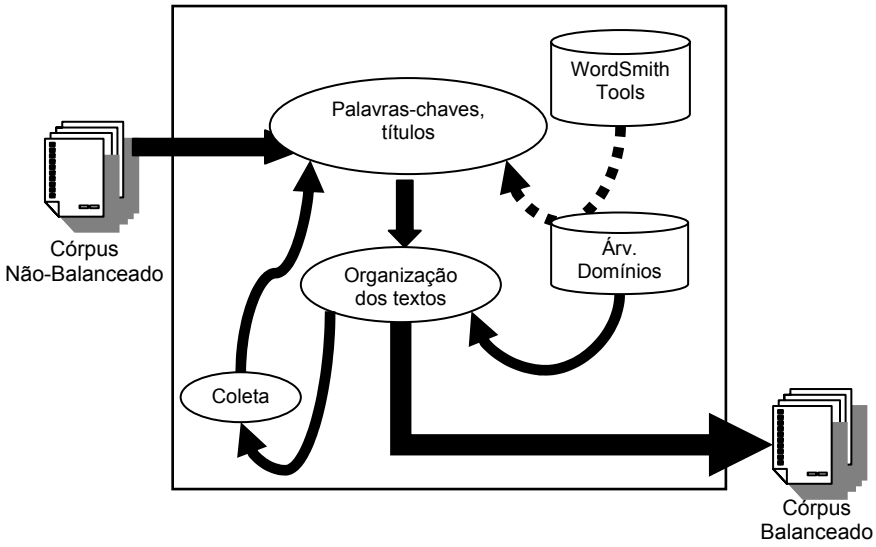


Figura 4.4: Diagrama da etapa de balanceamento dos textos contidos no córpus.

Conforme é apresentado, o balanceamento pode ser feito com o auxílio de uma ferramenta computacional, no caso é sugerido o Word Smith Tools, para a realização da extração de palavras-chaves. Essa ferramenta poderá ser utilizada se as palavras-chaves existentes nos artigos científicos e os títulos não forem suficientes para realizar a organização das seções de artigos científicos sob a arvora de domínios gerada em EC.

4.4.1 Instrução da Etapa E0

Depois de apresentadas as possibilidades de coleta de textos (via Web ou não), formatá-los, nomeá-los e organizá-los, ou seja, depois desse pré-processamento dos textos coletados, a próxima etapa a ser descrita é a avaliação do balanceamento do córpus.

Como pode ser observado, o diagrama da Figura 4.4 ilustra o processo de avaliação do balanceamento de um córpus, que pode ser feito por meio da identificação do conteúdo tratado em cada texto e posterior distribuição dos mesmos, segundo o tipo de informação que trazem, sob a árvore de domínios construída na etapa anterior. A identificação do tipo de informação pode ser feita com base nas palavras-chaves trazidas nesses textos, nos títulos dos mesmos, ou ainda, com base nas palavras-chaves que poderão ser extraídas dos textos com o auxílio de uma ferramenta computacional (Mais informações ver em 4.4.2.1). A vantagem de

se extrair palavras-chaves com o auxílio de uma ferramenta computacional está no fato dessas palavras-chaves serem eleitas como chaves segundo um dado método, no caso, o estatístico e não apenas pela eleição subjetiva feita pelo(s) autor(es) dos textos, como acontece regularmente. A opção por utilizar as palavras-chaves já escolhidas pelo autor ou por extraí-las via ferramenta computacional vai depender da dificuldade em se classificar os textos quanto às subáreas que pertencem. Ou seja, caso o usuário do CECARL esteja com dificuldades para definir o conteúdo de um dado texto com base apenas nas palavras-chaves trazidas pelos textos e pelos títulos dos mesmos, ele poderá optar por fazer um levantamento estatístico dessas palavras-chaves. A ferramenta computacional sugerida para essa tarefa é o *WordSmith Tools*.

O *WordSmith Tools* é um *software*, desenvolvido por Mike Scott e publicado pela Oxford University Press desde 2001, somente obtido pela Internet, nos seguintes endereços: www.liv.ac.uk/~ms2928/;www.lexically.net/; www.oup.com/elt/global/isbn/6890/ (Berber-Sardinha, 2004; 1999). Nesses endereços, o usuário baixa a versão demo e se desejar a versão completa, precisa pagar uma licença para receber um código que o habilitará para converter a versão demo para uma completa. É de fácil manuseio e, por isso, seu uso se estende em diferentes áreas da comunidade lingüística. A *Oxford University Press*, por exemplo, a utiliza em trabalhos de lexicografia, que envolvem a preparação de dicionários; professores de língua, estudantes e pesquisadores na análise de padrões de uma dada língua podem por sua vez, utilizá-la na investigação de concordâncias, por exemplo.

Instruções de como utilizar esse tipo de ferramenta na extração de palavras-chaves podem ser obtidas no Apêndice 9.

4.5 Etapa de Anotação Automática dos Componentes da Estrutura Esquemática

A Figura 4.5 apresenta uma visão geral da etapa de Anotação Automática da Estrutura Esquemática, composta por uma ferramenta computacional que identifica os componentes da estrutura citada de maneira automática. Essa ferramenta foi desenvolvida por um mestrando em Ciências da Computação do ICMC-USP, sob a mesma orientação do mestrado em tela. Depois de pronta, essa ferramenta foi adicionada ao processo aqui proposto, de modo que o nosso usuário possa utilizá-la para agilizar o processo de identificação das estruturas esquemáticas de todas as seções de artigos científicos. Na disponibilização *on-line* do CECARL, pode ser encontrado o *link* de acesso para essa ferramenta.

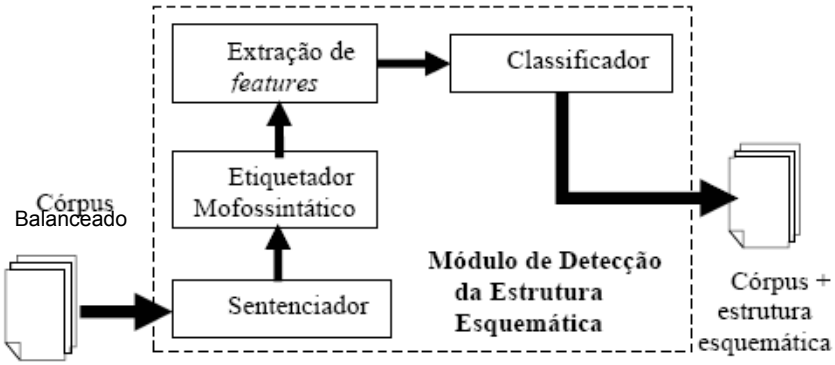


Figura 4.5: Diagrama da etapa de anotação automática dos componentes da estrutura esquemática de seções de artigos escritos em inglês (Marquiafável *et al*, 2006).

4.5.1 Instrução da Etapa E1

Como observado na Figura 4.5, o procedimento para sua utilização é bem simples e consiste na submissão de um grupo de seções de artigos científicos a esse classificador, por exemplo, um grupo de resumos, introduções e assim por diante. A seguir, esse classificador segmenta automaticamente os textos em sentenças e depois em *tokens*[10], que recebem etiquetas morfossintáticas (trabalho feito pelo tokenizador/etiquetador morfossintático de um pacote com ferramentas lingüísticas TTT[11]). Após esse pré-processamento de texto, o categorizador extrai, de textos submetidos, os valores de sete traços – *features* - (Tabela 4.2) utilizadas pelo AZEA[12] na anotação automática dos componentes da estrutura esquemática presente nos textos submetidos, obtendo-se ao final um córpus categorizado quanto aos componentes esquemáticos que possui. O AZEA é um classificador automático da estrutura esquemática de resumos construído por um mestrando em computação. Depois de pronto, poderá ser utilizado pelo usuário do CECARL para que esse obtenha os componentes esquemáticos de seus resumos, por exemplo, identificados de maneira automática.

[10] Token: em computação é um segmento de texto ou símbolos que podem ser manipulados por um parser (analisador sintático), em outras palavras, é um conjunto de caracteres (de um alfabeto, por exemplo) com um significado coletivo.
[11] http://www.ltg.ed.ac.uk/software/ttt/index.html
[12] AZEA: Classificador automático da estrutura esquemática de resumos construído por um mestrando em computação. Depois de pronto, poderá ser utilizado pelo usuário de nosso processo semi-automático para que esse usuário tenha os componentes esquemáticos de seus resumos, por exemplo, identificados de maneira automática.

Feature	Descrição	Valores Possíveis
Tamanho (length)	Tamanho da sentença	*Small, medium, big*
Localização (post_sent)	Posição da sentença no texto	*Fir, sec, third, méd penult, last*
Tempo (tense)	Tempo do primeiro verbo finito da sentença	*BaseForm, Gerund, Past, PastPart, Pres3, PresNo3, NoVerb*
Modal (modal)	Se o primeiro verbo finito da sentença é ou não modal	*Modal, NoModal ou NoVerb*
Histórico (history)	Categoria da sentença anterior	*None, Background, Gap, Purpose, Method, Result, Conclusion*
Expressões Formulaicas (formulaic)	Tipo de expressão-padrão contida na sentença	19 tipos de expressões formulaicas ou *none*
Agente (agent)	Tipo de agente contido na sentença	14 tipos de agente ou *none*

Tabela 4.2: *Features* utilizadas no AZEA para a anotação automática de componentes esquemáticos de resumos (Marquiafável *et al*, 2006).

O esquema de funcionamento desse detector automático de componentes esquemáticos se dá, de maneira resumida, com a entrada de dois córpus: um com componentes esquemáticos anotados (córpus de treinamento) e outro, novo, a ser anotado. Ambos os córpus devem ser do mesmo gênero e conter textos da mesma seção de artigo científico. A partir desse córpus de treinamento, é realizada a indução do classificador apresentado, que é aplicado ao córpus novo para a anotação automática de componentes esquemáticos.

Para que a categorização automática de tais componentes pudesse ser realizada com todas as seções de artigos científicos, o procedimento descrito acima foi repetido para cada uma das seções (Resumo, Introdução, Conclusão, Discussão, Metodologia e Resultado) de artigos científicos contidos no SciPo-Farmácia, utilizados, portanto, como córpus de treinamento do categorizador citado.

O início de desenvolvimento dessa ferramenta computacional se deu para a análise da estrutura esquemática de resumos. Essa é, portanto, a seção mais trabalhada e que, por isso, apresenta os melhores resultados de categorização esquemática. Para realizar a categorização esquemática de resumos, a ferramenta busca identificar primeiramente a existência de uma sentença que possa ser identificada como Propósito do Resumo analisado. A seguir, essa ferramenta computacional classifica as sentenças anteriores e posteriores a da sentença propósito identificada. A estatística *Kappa* utilizada para medir a concordância da anotação automática dessa tarefa indicou k=0,667, resultado bom, considerando-se o grau de dificuldade da tarefa automática realizada.

Esse categorizador acima apresentado, chamado AZEA-Web, pode ser acessado pelo endereço http://www.nilc.icmc.usp.br/azea-web/.

A próxima etapa apresenta como o procedimento realizado por essa ferramenta computacional citada pode ser feito manualmente.

4.6 Etapa de Anotação Manual dos Componentes da Estrutura Esquemática

A Figura 4.6 apresenta uma visão geral da etapa de Marcação Manual dos Componentes da Estrutura Esquemática, o qual recebe como entrada um córpus balanceado e como saída um córpus que teve os componentes de sua estrutura esquemática identificados manualmente, com o auxílio de um editor de textos ou de uma ferramenta computacional. A utilização dessa etapa é descartada, se a etapa de anotação automática de estruturas esquemáticas (M1 apresentado na seção anterior) for utilizada.

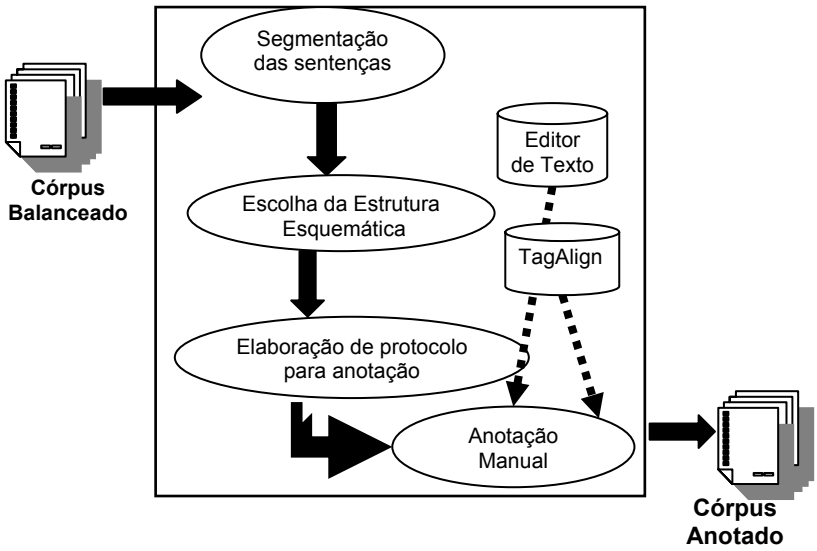


Figura 4.6: Diagrama da etapa de Anotação Manual de Estruturas Esquemáticas.

4.6.1 Instrução da Etapa E1'

Para se realizar a anotação manual da estrutura esquemática de uma seção qualquer de um artigo científico, é interessante separar as sentenças da seção a ser analisada com espaço de uma linha entre elas para facilitar o processo de anotação da estrutura esquemática contida em cada uma delas. Esse procedimento pode ser realizado manualmente em um editor de textos qualquer ou de maneira automática, com o emprego de uma ferramenta computacional como, por exemplo, o *Senter* (Pardo, 2006). O *Senter* é uma ferramenta computacional que

serve para segmentar automaticamente um texto (em inglês ou português) em sentenças. Tal ferramenta está disponível gratuitamente para *download* em http://www.icmc.usp.br/~taspardo/senter.html e depois de instalada, basta executar a linha de comando "senter.exe myfile.txt". A seguir, o texto segmentado será salvo em um arquivo com o mesmo nome do arquivo submetido à segmentação + ".seg", por exemplo, resumos.seg no qual haverá uma sentença por linha. O arquivo de entrada dessa ferramenta é do tipo texto sem formatação, ou seja, arquivos com o formato .txt.

Depois dessa organização, o próximo passo é escolher os componentes esquemáticos a serem identificados em cada sentença do texto. Neste trabalho foram utilizados, principalmente, os trabalhos de Swales (1990) e Weissberg & Buker (1990), muito respeitados na área de estudos de gênero e que propuseram modelos deste tipo de estrutura para todas as seções de artigos científicos. Interessante ainda dizer que os modelos de componentes esquemáticos propostos por estes autores foram baseados em análises de textos de diferentes áreas do conhecimento, o que possibilita a replicação destes modelos com quaisquer seções de textos científicos das três grandes áreas: Exatas, Humanas e Biológicas. No entanto, vale ressaltar que esses modelos não podem ser vistos como verdades absolutas, mas sim como possibilidades de tipos de componentes esquemáticos que podem ser encontrados em uma dada seção de artigo científico. Assim, se durante a anotação de um texto for observada a ausência de um dado componente não contido no modelo escolhido, mas que tenha sido identificado no córpus, é indicado optar pela inclusão do mesmo na anotação, uma vez que o texto mostrou necessidade de tal inserção, que o modelo não foi possível de prever.

Depois de escolhido o modelo de estrutura esquemática, é preciso que se elabore um protocolo (manual) de anotação dos componentes esquemáticos. Esse protocolo consiste em um tipo de documentação escrita na qual, além do modelo de estrutura esquemática adotado, também deve constar exemplos de sentenças nas quais os componentes esquemáticos ocorrem, como também, procedimentos indicados em momentos de dúvida no processo de anotação. A confecção de um manual é aconselhável uma vez que facilita a replicação da tarefa de anotação por diferentes pessoas que o poderão consultar sempre que sentirem necessidade. Para a anotação da Seção "Metodologia", por exemplo, foi preparado um manual de anotação. Além disso, foram desenvolvidos manuais para a anotação de componentes da estrutura esquemática para as outras seções que um artigo científico pode apresentar. Estes manuais correspondem aos Apêndices 1, 2, 5, 6, 7 e 8.

Depois de elaborado um manual, inicia-se a anotação dos textos. Essa anotação pode ser feita por um editor de textos, como o *Microsoft Word*® ou por uma ferramenta computacional

que tenha as funcionalidades, por exemplo, da *TagAlign* (Caseli *et al*, 2002) apresentada na Figura 4.7. Como pode ser observado na Figura 4.7, o funcionamento desse tipo de ferramenta é simples e pode facilitar o trabalho do anotador. Trata-se de uma interface de fácil interação, bastando apenas ao usuário submeter à ferramenta um arquivo com as etiquetas que vai utilizar no processo de anotação, como também o texto a ser anotado. Selecionando a sentença a ser anotada, basta um clique na etiqueta a ser adicionada à sentença para que seja finalizado o processo de anotação. Assim, o usuário repete esse procedimento até o fim do texto, quando salvará o arquivo que será automaticamente salvo em formato XML, formato de arquivo requerido por ferramentas de auxílio à escrita semelhantes ao SciPo-Farmácia. Portanto, no caso de se optar por realizar a anotação dos textos em um editor do tipo *Word*, será necessário que se converta esse arquivo do tipo .doc para o formato XML, pois é esse formato que recupera e exibe ao usuário da ferramenta de suporte exemplos reais de seções de artigos científicos.

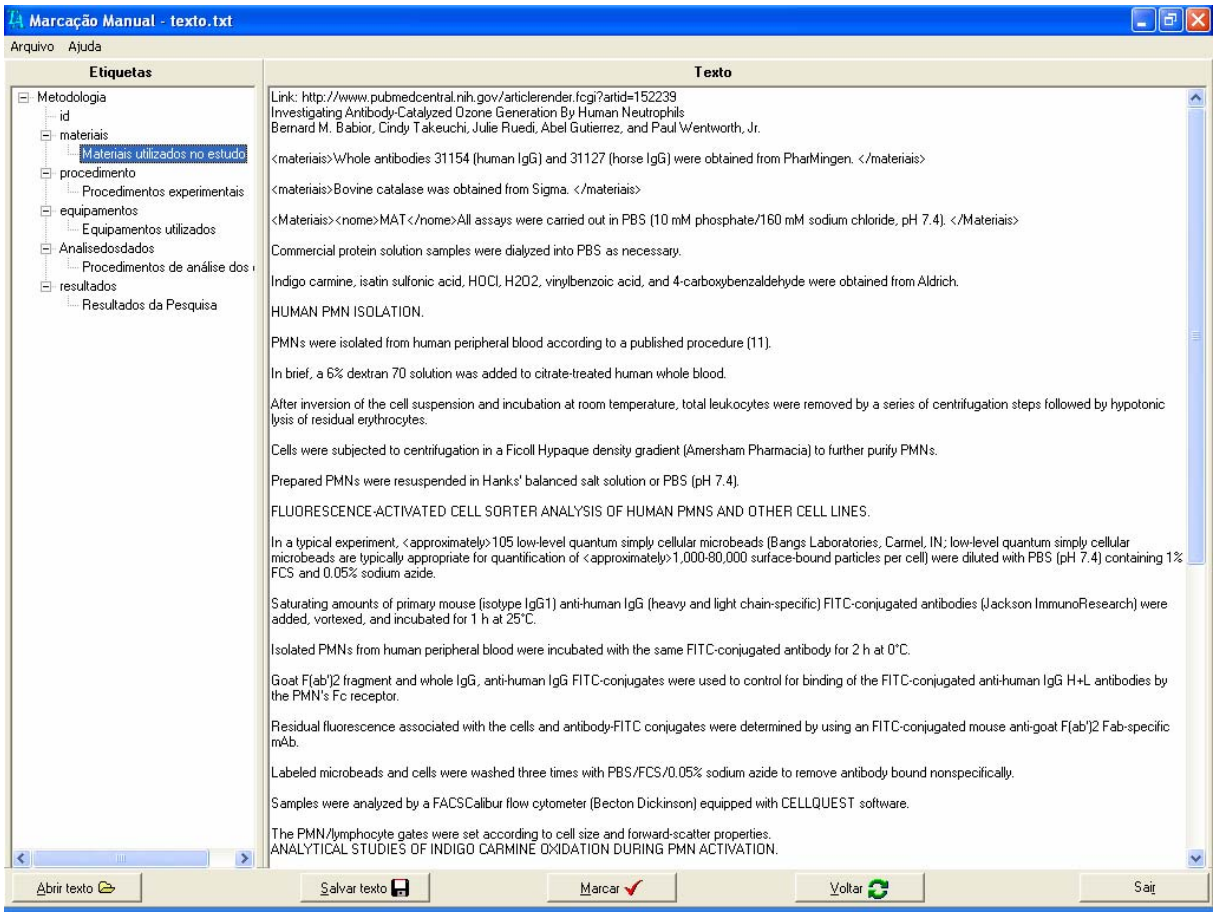


Figura 4.7: Tela da Ferramenta de auxílio à anotação *TagAlign*. O fato de utilizar apenas um botão do mouse para fazer a anotação em vez de ter que digitar as siglas das categorias faz com que a tarefa de anotação seja efetuada mais rapidamente e sem riscos de digitação incorreta das siglas, além da interface ser de fácil entendimento e manuseio. Interessante também dizer que é simples alterar as etiquetas para trabalhar com textos de outras seções. Para isso, é preciso apenas alterar o arquivo de etiquetas, que deve ser inserido num dado formato, cujo exemplo está inserido no pacote de instalação dessa ferramenta. Além disso, dá para indicar mais de uma etiqueta para sentenças com mais de uma função retórica. No entanto, poderiam ser sugeridas algumas alterações, como: (1) Ao carregar o texto a ser anotado, que deve estar em formato .txt, seria interessante que a separação existente entre as linhas (espaço de uma linha) se conservasse, pois facilitaria a leitura do texto. (2) Seria interessante que a fonte dos textos apresentados na tela dessa ferramenta fosse um pouco maior, facilitando a leitura. (3) Se as etiquetas fossem coloridas facilitaria a identificação da distribuição das funções nas sentenças e no texto em geral.

4.6.2 Instanciação da Etapa E1'

A anotação manual dos componentes esquemáticos da seção "Metodologia" foi realizada no âmbito deste projeto por duas razões: na ferramenta Scipo-Farmácia, a seção "Metodologia" era a única que ainda não havia sido implementada e porque o classificador para a anotação automática da mesma ainda não havia sido induzido.

A partir de 30 textos da seção "Metodologia", retirados de fontes *on-line* de divulgação científica da área de Ciências Farmacêuticas, realizou-se a anotação manual do córpus Met. Nessa primeira etapa de anotação, cada sentença dos textos foi anotada segundo os

componentes esquemáticos e as estratégias retóricas contidos. Vale lembrar que os três níveis escolhidos para serem identificados nos textos ao longo de todo o CECARL são (1) componentes esquemáticos, (2) marcadores discursivos e (3) estratégias retóricas. Esses três níveis de marcação foram escolhidos por refletirem os níveis de auxílio que a ferramenta de suporte gerada com nosso processo irá oferecer: organização do texto (componentes esquemáticos), modo de escrita de cada estrutura esquemática (estratégias retóricas) e os padrões lexicais utilizados em cada sentença de texto, de acordo com a estratégia retórica escolhida (marcadores discursivos).

A seguir, descrevemos o modelo de componentes da estrutura esquemática adotado para a anotação manual do córpus Met, bem como todas as etapas contidas nesse processo de anotação.

4.6.2.1 Modelo de Componentes de Estrutura Esquemática para a Seção "Metodologia"

Em linhas gerais, a seção "Metodologia" pode ser definida como a materialização lingüístico-discursiva de uma pesquisa científica com o objetivo de apresentar a descrição dos métodos, materiais e procedimentos utilizados nessa pesquisa (Coracini, 1991: 26). Entretanto, além de descrever a investigação metodológica e procedimental de um fenômeno, a seção "Metodologia" se constitui como um texto argumentativo em que o pesquisador/escritor tem como objetivo principal persuadir o seu leitor a crer na veracidade da teoria, do método e dos dados que foram por ele investigados e analisados (princípio da validade) e, se for de interesse, replicar o experimento que está sendo apresentado e descrito (princípio da replicabilidade) (Swales,1990:121). Huckin & Olsen (1991:362) complementam esse princípio contido na seção "Metodologia" dizendo que a mesma deve conter detalhes suficientes para permitir a qualquer pesquisador experiente em sua área reproduzir seus resultados com exatidão. Nesse sentido, para que o leitor possa construir sentido sobre o que está sendo apresentado e até mesmo possa repetir, é preciso que o texto seja coerente. Além disso, para que a pesquisa obtenha êxito e mérito na comunidade científica da qual o pesquisador faz parte, não basta conduzir a escrita da investigação segundo o paradigma vigente dessa comunidade; é preciso também que o texto construído por este pesquisador apresente uma organização retórica e léxico-gramatical que possa ser compartilhada pelos membros desse contexto de produção.

Assim, a explicitação da organização retórico-lingüística contida nesse tipo de seção contribui como tentativa de auxílio a amenização das principais dificuldades sentidas no

momento de redação. Em vista disso, esta pesquisa busca subsídios na Análise de Gêneros, especialmente no trabalho de John Swales para identificar a configuração retórica contida na seção "Metodologia", que é marcada segundo Swales (1990), pela utilização de passos retóricos que visam apenas descrever objetiva e cronologicamente os procedimentos de coleta, análise e interpretação dos dados, com ênfase na descrição dos materiais e métodos utilizados na realização da investigação. A opção por esse teórico se deve ao fato de ter baseado seus modelos em análises de textos reais (córpus) em vez de prescrever padrões sem verificação de sua ocorrência, mostrando a rica variedade de padrões contida nos movimentos e também a freqüência com que ocorre, postura também compartilhada por essa pesquisa.

De acordo com o modelo de Swales (1990), a seção "Metodologia" possui quatro movimentos. O movimento consiste em uma unidade textual funcional, utilizada com algum propósito retórico identificável. Os movimentos podem variar em tamanho, mas normalmente possuem, no mínimo, uma proposição (Mauranen, 1993:225). No caso desse estudo, os movimentos ilustrados na Tabela 4.3 foram identificados ao longo de cada sentença contida nas trinta seções Metodologia compiladas para nosso trabalho.

Sigla	***Categoria***	***Descrição da Categoria***
MAT	Materiais	Materiais utilizados no estudo
PRO	Procedimentos	Procedimentos necessários à execução correta da metodologia
EQU	Equipamentos	Equipamentos utilizados no experimento
PAD	Análise de Dados	Procedimentos de análise dos dados
RES	Resultados	Resultados da Pesquisa

Tabela 4.3: Note que a sigla da categoria é composta sempre por letras contidas em suas respectivas categorias, de forma a facilitar a memorização e fácil identificação do significado da categoria que deverá ser empregada nas sentenças do córpus.

A anotação foi realizada por quatro juízes (anotadores) a fim de se assegurar maior confiabilidade na identificação das partes dos textos e adotou-se como modelo de estruturação esquemática o esquema de *Moves* (Movimentos) proposto por Swales (1990), apresentado na Tabela 4.3. Embora o modelo apresentado por Swales seja amplamente aceito na literatura, observou-se, com o auxílio do córpus em estudo, que a inserção de mais um passo seria necessária: adicionou-se o *Move* 5 – "Resultados", e sua estratégia retórica, "Resultados da Pesquisa". Isso porque foram encontrados relatos breves dos resultados dos estudos descritos, na seção "Metodologia" dos artigos científicos anotados.

Antes de realizar a anotação, os anotadores receberam um manual que descreve o modo como esse processo deveria ser feito e que se encontra no Apêndice 1 deste trabalho. Esse manual apresenta orientações para se realizar a anotação manual da seção "Metodologia" de artigos da área de Ciências Farmacêuticas. Esse manual inicia com a apresentação das partes constitutivas de um artigo, com foco para a seção "Metodologia" e segue com a apresentação das categorias a serem identificadas em 30 seções metodologia de artigos da Farmácia. Vale dizer que nesse manual está também incluso a tarefa de anotação contida na etapa de anotação manual das estratégias retóricas, uma vez que a anotação das estratégias retóricas e das estruturas esquemáticas foi feita simultaneamente.

Cada uma das estratégias retóricas e estruturas esquemáticas foi apresentada com exemplos retirados do próprio córpus no momento de "treinamento", isto é, familiarização dos anotadores com as categorias a serem identificadas. Depois de esclarecidas algumas dúvidas sobre o significado de cada sigla utilizada na anotação bem como a maneira de se realizar a anotação dos textos, deu-se início ao procedimento de familiarização com o modelo adotado e com os textos a serem trabalhados. Nessa etapa, foram utilizados cinco textos (81 sentenças), o esquema de estruturação retórica apresentado na Tabela 4.3 e do manual de anotação, que serviu de protocolo para o processo. Discussões foram realizadas entre os anotadores até que se chegasse a um consenso sobre a classificação das sentenças. Gerou-se, então, uma tabela tida como versão final da classificação dessas sentenças-teste, a qual serviu também como base de exemplos para a anotação do restante do córpus.

Durante o processo de anotação, reconheceram-se padrões de estruturação esquemática citados acima como também desvios desse mesmo esquema adotado, os quais podem ser prejudiciais, uma vez que podem comprometer o entendimento do trabalho, dificultar a análise retórica e, conseqüentemente, gerar discordâncias entre os anotadores. Os principais desvios observados no Córpus Met foram:

- organização estrutural ineficiente (aspecto de texto recortado, sem fluência)
- formulação de sentenças inadequadas (organizadas aparentemente de forma aleatória)
- trechos confusos que suscitaram discordância total entre os anotadores, pois não foi possível reconhecer a função retórica utilizada pelo autor nesses tipos de sentenças

Diante desse fato, dividiram-se os textos do córpus em bons e ruins, sendo os primeiros os escolhidos para serem inseridos na ferramenta SciPo-Farmácia.

Vale dizer ainda, que assim como Bruce (1983:8), Weissberg (1984) e Swales (1990) identificaram certas organizações em textos de Metodologias, as mesmas puderam ser identificadas em nosso córpus de estudo. Ao analisar seções Metodologia de um periódico da

medicina, Bruce (1983) observou que o texto desse tipo de seção pode parecer à primeira vista incoerente, isto é, uma ausência praticamente completa de elementos referenciais, mas que na verdade a coerência é preenchida com o conhecimento sobre os procedimentos investigativos e suas seqüências apropriadas, trazidos ao texto pelo leitor. Nesse mesmo contexto, Weissberg (1984) analisou 20 Metodologias de diferentes disciplinas, encontrando apenas dez itens lexicais coesivos. Com base em investigações semelhantes a essa que Weissberg (1984) realizou, Swales (1990:168) chegou à conclusão de que as seqüências textuais desse tipo de seção de texto são caracterizadas por uma linearidade quebrada (*broken linear*), isto é, em muitas Metodologias, as sentenças parecem uma corrente de ilhas, apenas aqueles que possuem um conhecimento especializado e experiência conseguem facilmente pular de uma para outra. No entanto,Swales alerta que esse tipo de caracterização não é, em geral, identificada em textos da área das Ciências Humanas, na qual há uma descrição passo-a-passo massivamente subsidiada por dadas referências anafóricas e repetições lexicais. Swales justifica essa diferença pelos fenômenos sociológicos e intelectuais que constituem a comunidade discursiva da área de humanas.

A seguir será apresentada a etapa de anotação automática da Qualidade de Escrita, cuja ferramenta foi desenvolvida pelo mesmo mestrando de Ciências da Computação responsável pela etapa E1'.

4.7 Etapa de Avaliação Automática de Qualidade de Escrita

A Figura 4.8 apresenta uma visão geral da etapa de Avaliação Automática da Qualidade de Escrita, cuja ferramenta computacional foi desenvolvida por Genoves, 2007 (no prelo) como parte de seu projeto de mestrado. Esta etapa possui duas funções: *Justificar os resultados da avaliação automática* de um texto de um usuário da Ferramenta de Suporte à Escrita e *Garantir a qualidade do córpus* a ser utilizado em uma Ferramenta de Suporte à Escrita, criada pelo processo proposto no mestrado em tela.

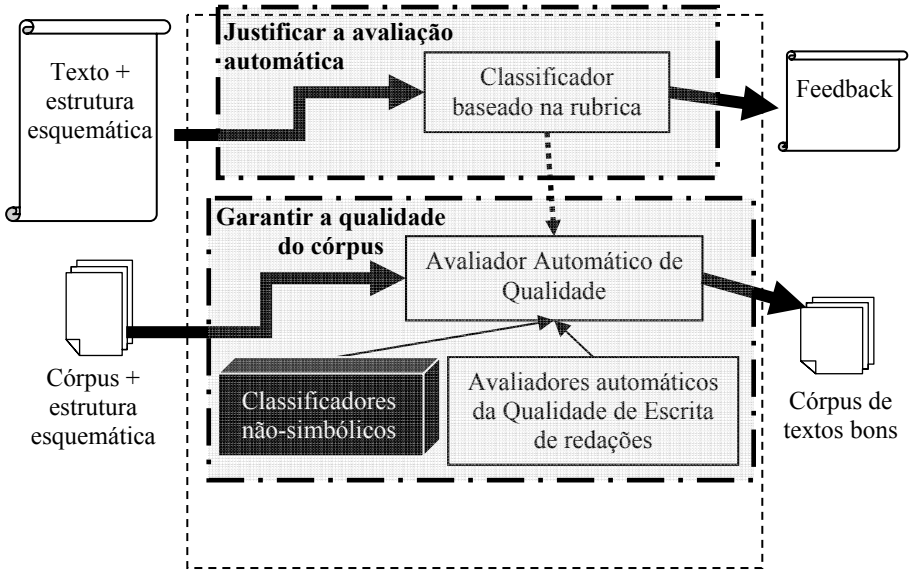


Figura 4.8: Diagrama da Etapa de Avaliação Automática da Qualidade de Escrita.

A função que avalia a qualidade de um córpus é de interesse especial para o nosso mestrado, pois pode fornecer subsídios para auxiliar o usuário do CECARL na avaliação do córpus a ser usado na ferramenta de suporte à escrita. Como mostra a Figura 4.8, para garantir a qualidade do córpus, o avaliador automático de qualidade pode ser auxiliado por outros classificadores não-simbólicos, como os baseados em redes neurais e redes complexas, por adaptações dos avaliadores automáticos da qualidade de redações (gênero mais trabalhado na literatura), e inclusive pelo classificador baseado em uma rubrica dedicada ao gênero científico, em desenvolvimento por um grupo de pesquisadores do NILC e que é descrita abaixo. É importante ressaltar que o objetivo desta etapa é avaliar a qualidade de escrita do texto, e não o seu conteúdo.

A próxima etapa visa apresentar uma maneira de como essa atividade desempenhada automaticamente pode ser realizada de maneira manual.

4.8 Etapa de Avaliação Manual da Qualidade de Escrita

A Figura 4.9 apresenta uma visão geral da etapa de avaliação manual da qualidade de seções de artigos científicos em inglês, o qual recebe como entrada uma seção de artigo científico, por vez, a ser avaliada. Em nosso estudo, a seção eleita é a "Resumo" e a saída, portanto, desse processo é um resumo, cuja qualidade foi avaliada. Essa qualidade é avaliada com base em um conjunto de critérios, conforme apresenta a Figura 4.9.

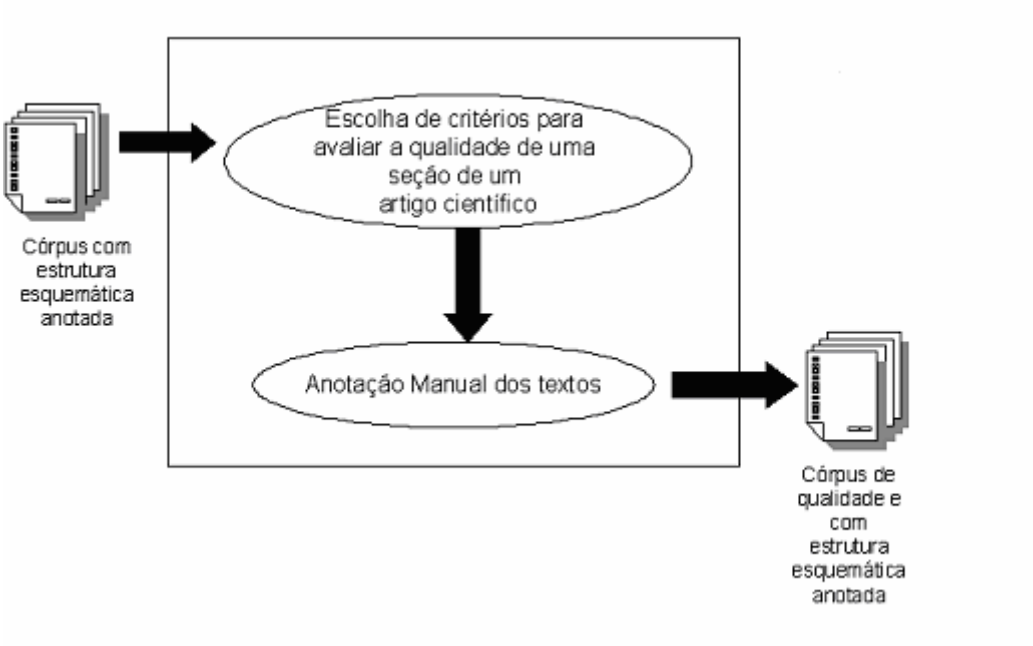


Figura 4.9: Diagrama do processo de avaliação manual da qualidade de uma dada seção de artigo científico contido na Etapa E2' do CECARL.

4.8.1 Instrução da Etapa E2'

Como pode ser observado na Figura 4.9, o procedimento contido na etapa E2' é bem simples. A entrada dessa etapa é uma seção de artigo científico, no nosso caso a seção "Resumo", que será avaliada quanto a sua adequação textual. Essa seção possui sua estrutura esquemática anotada, produto da etapa anterior. Como saída, tem-se um texto adequado ou inadequado, dependendo dos critérios e da relação deles com a seção do texto em análise.Os bons textos serão utilizados na etapa seguinte, enquanto que os ruins serão descartados. Como núcleo desse procedimento tem-se um conjunto de critérios – rubrica – responsável em grande parte pelo bom resultado dessa etapa.

A seguir serão apresentados os sete critérios (ou dimensões) da rubrica citada. Vale dizer, que os mesmos foram elaborados por Aluisio *et al* (2005), Schuster *et al* (2005) e Genoves *et al* (2007), para realizar a avaliação da adequação textual de *abstracts* da área de Ciências Farmacêuticas escritos por brasileiros. Assim, para se avaliar uma outra seção qualquer dessa mesma área ou até mesmo um resumo em inglês de uma outra área, é necessário realizar adaptações nos critérios dependentes de domínio. Os não dependentes como, por exemplo, o quinto critério (ou dimensão), não precisa ser adaptado, porque foi desenvolvido com base em textos de diferentes áreas. Das sete dimensões, as três primeiras avaliam o texto como um todo e as outras cada uma das sentenças, atribuindo um dos três

valores: Alto, Baixo e Padrão (ou N/A, isto é "não-aplicável"). Diferentemente do modelo de rubrica apresentado no Capítulo 3, no qual as duas primeiras dimensões apresentadas avaliam o texto como um todo.

Dimensão 1 – Organização e Desenvolvimento de um texto: esse critério é indicado para investigar a estrutura esquemática contida em uma dada seção de um artigo científico, no caso, a seção "Resumo". Ela objetiva tanto a anotação de **componentes essenciais** a essa seção em foco, quanto a verificação da ordem que esses componentes devem aparecer no texto. Para tal, são utilizados dois valores: Alto e Baixo. O valor Alto é atribuído quando os componentes principais da estrutura esquemática estão presentes e são apresentados em ordem lógica. Por exemplo, na seção "Resumo" os componentes da estrutura esquemática principal apresentaria a seguinte ordem: Propósito, Metodologia (se houver), Resultados principais e Conclusão. Como nem todos os resumos apresentam a mesma ordem proposta por esse modelo ideal de componentes esquemáticos de um resumo adequado às especificações dos pesquisadores sobre escrita científica, a ordem dos componentes presentes deve obedecer a uma lógica que satisfaça as expectativas do leitor, ou seja, deve conter uma ordem que apresente de maneira lógica as informações descritas. Assim, se houver uma Lacuna, esta deve ser seguida pelo Propósito. Se existir Contexto e Lacuna, a Lacuna deve aparecer depois do Contexto. Mas é possível também haver ciclos de Contexto e Lacuna. O valor Baixo é atribuído quando as condições descritas acima não forem satisfeitas.

Como cada área tem um conjunto de características específicas que fazem variar esse conjunto de componentes principais, sugere-se então, que seja feita uma avaliação empírica para obtenção desse dado. Coletar-se-ia, por exemplo, um conjunto de resumos considerados clássicos na área, isto é, aqueles cuja qualidade é indiscutível. Posteriormente, poderia ser feita uma anotação das estruturas esquemáticas nele contidas, para então se fazer um levantamento daquelas mais recorrentes. Esse tipo de procedimento foi realizado por Biasi-Rodrigues e Jucá (2004) e parte dos resultados obtidos são apresentados na Tabela 4.4, como sugestão de auxílio para se eleger os principais componentes da estrutura esquemática a serem considerados dentro de uma dada área. Vale dizer que se tratou de uma investigação baseada em 32 resumos em inglês de oito áreas diferentes, um número, entretanto, ainda baixo para avaliação estatística.

Área Científica	*Apresentação Pesquisa*	*Contextualização Pesquisa*	*Apresentação Metodologia*	*Sumarização Resultados*	*Conclusão Pesquisa*
Eng. Elétrica	0%	50%	100%	100%	0%
Eng. Mecânica	100%	100%	50%	0%	0%
Sociologia	50%	100%	0%	50%	0%
Educação	50%	50%	50%	100%	0%
Lingüística	100%	100%	0%	50%	0%
Farmácia	50%	0%	100%	100%	0%
Enfermagem	100%	100%	0%	50%	0%
Economia	100%	100%	0%	0%	0%
Total	69%	75%	38%	56%	0%

Tabela 4.4: Tabela ilustrativa de parte dos resultados obtidos nas investigações de Biasi-Rodrigues e Jucá (2004) sobre os componentes da estrutura esquemática mais recorrentes em determinadas áreas. Conforme observado, o componente "Contextualização da Pesquisa" é o mais recorrente em todas as áreas, com 75% de freqüência. Em contrapartida, o componente "Conclusão da Pesquisa" não ocorreu em nenhum dos *abstracts* investigados. Vale ainda dizer que a estrutura esquemática apresentada na tabela difere um pouco da adotada por nós nessa pesquisa.

Dimensão 2 – Balanceamento entre os componentes: essa dimensão visa verificar o balanceamento do tamanho de cada uma das seções de um artigo científico, em separado. Por exemplo, os resumos, em geral, não devem ultrapassar um limite de 200 a 300 palavras,o que implica na imposição de algumas restrições ao uso de dados componentes utilizados em resumos, como, por exemplo, não supervalorizar a escrita de um contexto com várias sentenças. Para tal verificação, são também utilizados os valores Alto e Baixo. O valor Alto é atribuído para resumos escritos em inglês na área de Ciências Farmacêuticas quando: 1) o Propósito existe e foi escrito em apenas uma sentença; 2) a Conclusão existe e foi escrita em apenas uma sentença; 3) se existir Contexto, não deve ultrapassar 30% das palavras de um *abstract*. O valor Baixo é atribuído quando as condições descritas acima não forem satisfeitas.

Para se fazer a verificação do balanceamento das estruturas esquemáticas em outras seções de artigos científicos, que sejam ou não da área de Farmácia é interessante realizar o mesmo procedimento sugerido na dimensão anterior: realizar um levantamento empírico com textos da área e seção de artigo científico para a qual se deseja verificar o tipo de balanceamento da estruturação esquemática mais recorrente.

Dimensão 3 – Coerência entre os componentes: essa dimensão visa avaliar a coerência entre os componentes da estrutura esquemática de uma seção, ou seja, verificar se os

componentes estão relacionados entre si de forma a contribuir com a coerência do texto. A coerência pode, grosso modo, ser definida como o resultado de uma não-contradição entre os diversos segmentos de um texto, que devem estar encadeados logicamente. Cada segmento textual é pressuposto do seguimento que vem a seguir, que por sua vez será pressuposto para o(s) que lhe sucederem, formando assim uma corrente, uma cadeia na qual todos os segmentos estejam concatenados de maneira harmônica. Quando um segmento está em contradição com um anterior, perde-se coerência textual. Para a verificação de tal coerência, também são utilizados os valores Alto e Baixo. Abaixo serão apresentados os critérios que devem estar presentes no momento de verificação da coerência de resumos em inglês da área de Farmácia. Para as outras seções de artigos científicos, outros critérios de coerência devem ser criados, com base em uma investigação empírica, de modo que as características peculiares da seção a ser avaliada sejam consideradas. Além disso, o bom senso também pode ajudar a avaliar a coerência ou lógica existente entre o fluxo da informação contida na seção de texto a ser analisada. Assim, o valor Alto é atribuído à coerência de um resumo em inglês da Farmácia:

1) Se o Propósito estiver relacionado com a Lacuna, em uma relação de *fullfilment,* isto é, se o propósito realmente preenche a(s) lacuna(s) levantada(s) no componente Lacuna. Interessante notar que como a Lacuna não é um item obrigatório, quando não está presente, ao Propósito é atribuído o valor-padrão (N/A, isto é, "não-aplicável").

2) Se os Resultados principais estiverem relacionados com o Propósito, em uma relação de *accomplishment*, isto é, se os resultados que se esperavam encontrar com a pesquisa foram realmente alcançados.

3) Se a Conclusão estiver relacionada com os resultados principais, em uma relação de *generalization*, isto é, quando o autor consegue generalizar seus resultados no componente Conclusão.

Dimensão 4 – Marcadores coesivos: essa dimensão visa avaliar se as sentenças contidas em cada estrutura esquemática estão coesas, isto é, contém uma relação lógica/coerente estabelecida entre elas. Essa coesão pode se dar por meio de marcadores discursivos, referências pronominais e re-introdução de nomes. O valor Alto é então atribuído se uma dada sentença constitutiva de uma estrutura esquemática estabelece uma relação com pelo menos uma outra sentença da mesma estrutura esquemática. Caso contrário o valor atribuído é Baixo. Se um dado componente possui apenas uma sentença, o valor a ser atribuído é Padrão

(ou N/A, isto é, "não-aplicável"). A seguir são apresentados exemplos de coesão de cada elemento lingüístico citado acima:

- Coesão por marcador discursivo: "Catalase decreased the rate of cysteine oxidation, *but* the sensitivity to iron was similar in the presence and absence of catalase". Uma lista com exemplos de marcadores discursivos pode ser consultada no Apêndice 3. (Exemplo retirado do SciPo-Farmácia).

"Dogma dictates that the lethal blow is delivered to microbes by reactive oxygen species (ROS) and halogens, products of the NADPH oxidase, whose impairment causes immunodeficiency. *However*, recent evidence indicates that the microbes might be killed by proteases, activated by the oxidase through the generation of a hypertonic, K+ rich and alkaline environment in the phagocytic vacuole". (Exemplo retirado do SciPo-Farmácia).

- Coesão por referência pronominal: "In contrast, *SAA* was not a ligand or agonist for FPR, the high affinity fMLP receptor. Thus, *it* is the first chemotactic ligand identified for FPRL1". (Exemplo adaptado do SciPo-Farmácia).

- Coesão por re-introdução de nomes: "In contrast, *SAA* was not a ligand or agonist for FPR, the high affinity fMLP receptor. Thus, *SAA* is the first chemotactic ligand identified for FPRL1". (Exemplo adaptado do SciPo-Farmácia).

Dimensão 5 – Erros técnicos/gramaticais: esse critério tem como motivação identificar possíveis erros técnicos cometidos por brasileiros em geral na escrita de *abstratcs*. Dessa maneira, é uma rubrica interessante para ser utilizada por professores que queiram desenvolver uma lista de critérios para o julgamento de um texto produzido por um aluno. Mas isso não descarta a possibilidade de se usar todos ou apenas alguns deles para se avaliar a adequação de um resumo a ser inserido na base de casos de um ambiente de auxílio à escrita gerado com nosso CECARL.

Como essa dimensão foi elaborada para explorar a natureza dos erros técnicos comumente cometidos por brasileiros e para encontrar um meio de auxiliar esses alunos a corrigi-los, Genoves *et al* (2007) analisaram[13] 114 *abstracts* provenientes de alunos de áreas como Farmácia, Química, Biologia/Genética, Física e Ciências da Computação.

A Tabela 4.5 apresenta os erros apontados pelo estudo realizado, que são divididos em uso lexical, precisão sintática e correção mecânica.

[13] A análise citada no trabalho foi realizada por um dos autores que é um pesquisador nativo do inglês.

Tipos de erros

Correção mecânica	
P	Pontuação
CAP	Capitalização
SP	Ortografia
Uso Lexical	
WU	Uso incorreto de uma palavra para expressar um significado pretendido
WUCol	Uso incorreto de itens lexicais e colocações recorrentes
WF	Uso incorreto de formas como (*this/these, that/those* e pronomes possessivos ou não)
Precisão sintática	
ART	substituição de um artigo por outro (definido *versus* indefinido)
ART –	ausência de um artigo necessário em Inglês
ART+	Presença de um artigo não necessário em Inglês
WO_NP	Ordem incorreta das palavras em sintagmas nominais complexos
WO_ADJ	Ordem incorreta no emprego de adjetivos
WO_S	Ordem incorreta do sujeito contido na cláusula principal
WO	Ordem incorreta da palavra
S+VO	Sujeito extra
S-VO	Sujeito ausente
SV-O	Verbo ausente
POS	Classe gramatical
VU	Erro no uso do tempo verbal
VF	Erro na forma do verbo
SVA	Erro na concordância entre verbo e sujeito
S/PL	Erro no uso de substantivo no singular ou plural
S/PL_ADJ	Usar forma plural para um adjetivo
PORT	Utilização de uma palavra da língua portuguesa na escrita em inglês

Tabela 4.5: Tabela traduzida de Genoves *et al* (2007) sobre a categorização de erros técnicos.

A Tabela 4.6 apresenta os seis erros mais comuns levantadas por esse mesmo estudo.

Erros mais comuns cometidos por estudantes brasileiros	**%**
WU (Uso incorreto de uma palavra para expressar um significado pretendido)	25.8
ART- (Ausência de um artigo necessário em Inglês)	13.4
P (Pontuação)	8.6
SP (Ortografia)	7.6
WUCol (Uso incorreto de itens lexicais e colocações recorrentes)	5.7
ART+ (Presença de um artigo não necessário em Inglês)	4.9

Tabela 4.6: Erros mais comuns cometidos por estudantes brasileiros segundo estudo realizado por Genoves *et al* (2007).

O valor Baixo é atribuído se a sentença possui pelo menos um erro da Tabela 4.6 e recebe em contrapartida o valor Alto se não houver nenhum dos erros da Tabela 4.6.

Dimensão 6 – Estilo: esse critério, como seu próprio nome indica, visa averiguar o estilo de um texto científico, seja ele resumo, ou qualquer outra das seções contidas em um artigo científico. Dessa maneira, é esperado pela comunidade acadêmica de qualquer área que não haja um estilo coloquial presente na escrita, mas sim expressões lingüísticas características do

gênero científico. O valor Alto é atribuído se há ausência de indicadores estilo coloquial/pessoal, por exemplo, *I, my, me, frankly, by the way*, de enfáticos (*a lot, for sure, really*), de partículas discursivas de início de sentenças como *well, now, anyway*, de sentenças do tipo *I mean, I think, I assume, sort of, kind of, you know*. Caso contrário, o valor atribuído é Baixo.

Dimensão 7 – Informação Factual: embora haja autores cuja preferência seja a produção de artigos científicos indicativos, a comunidade acadêmica espera encontrar resumos informativos. Resumos indicativos não dispensam a leitura do artigo científico do qual faz parte, pois descreve apenas a natureza, a forma e o propósito do trabalho. Já o resumo informativo contém as principais informações do trabalho apresentado ao longo do artigo, dispensando assim, se desejável, a leitura do texto completo para se saber qual assunto é nele abordado. Dessa maneira, resumos indicativos acabam sendo interpretados como simples ponteiros do conteúdo a ser tratado em todo o artigo e não como um ponto que apresente de forma sucinta as principais informações que serão tratadas com mais detalhes ao longo do artigo científico do qual o resumo faz parte.

A seguir, são apresentados exemplos de cada um dos tipos de resumos citados. Esses exemplos foram retirados do centro de pesquisa jurídica Sílvia Mota, acessível pelo endereço http://www.silviamota.com.br/direito/artigos/resumo.htm.

- Exemplo de Resumo Indicativo: ROCCO, Maria Thereza Fraga. *Crise na linguagem*: a redação no vestibular. São Paulo: Mestre Jou, 1981. 184 p.

Estudo realizado sobre redações de vestibulandos da FUVEST. Examina os textos com base nas novas tendências dos estudos da linguagem, que buscam erigir uma gramática do texto, uma teoria do texto. São objetos de seu estudo a coesão, o clichê, a frase feita, o 'não-texto' e o discurso indefinido. Parte de conjecturas e indagações, apresenta os critérios para a análise, o candidato, o texto e farta exemplificação.

- Exemplo de Resumo Informativo: ROCCO, Maria Thereza Fraga. *Crise na linguagem*: a redação no vestibular. São Paulo: Mestre Jou, 1981. 184 p.

Examina 1500 redações de candidatos a vestibulares (1978), obtidas da FUVEST. O livro resultou de uma tese de doutoramento apresentada à USP em maio de 1981. Objetiva caracterizar a linguagem escrita dos vestibulandos e a existência de uma crise na linguagem escrita, particularmente desses indivíduos. Escolheu redações de vestibulandos pela oportunidade de obtenção de um *corpus* homogêneo. Sua hipótese inicial é a da existência de uma possível crise na linguagem e, através do estudo, estabelecer relações entre os textos e o nível de estruturação mental der seus produtores. Entre os problemas, ressaltam-se a carência

de nexos, de continuidade e quantidade de informações, ausência de originalidade. Também foram objeto de análise condições externas como família, escola, cultura, fatores sociais e econômicos. Um dos critérios utilizados para a análise é a utilização do conceito de coesão. A autora preocupa-se ainda com a progressão discursiva, com o discurso tautológico, as contradições lógicas evidentes, o nonsense, os clichês, as frases feitas. Chegou à conclusão de que 34,85 dos vestibulandos demonstram incapacidade de domínio dos termos relacionais: 16,95 apresentam problemas de contradições lógicas evidentes. A redundância ocorreu em 15,25 dos textos. O uso excessivo de clichês e frases feitas aparece em 69,05 dos textos. Somente em 40 textos verificou-se a presença de linguagem criativa. Às vezes o discurso estrutura-se com frases bombásticas, pretensamente de efeito. Recomenda a autora que uma das formas de combater a crise estaria em se ensinar a refazer o discurso falho e a buscar a originalidade, valorizando o devaneio.

Uma sentença receberá o valor Alto se for observado material informativo nas estruturas esquemáticas "Resultados Principais" e em sentenças de Conclusão, a ponto de o leitor não precisar recorrer ao texto todo para obter as informações que deveriam estar contidas nesses dois componentes, informações estas específicas do estudo desenvolvido.

Depois de apresentado esse conjunto de critérios que poderão servir como ponto de partida para o usuário do CECARL gerar um conjunto personalizado para a área em foco da seção de artigo científico que queira avaliar será apresentado o conteúdo da próxima etapa do processo citado, a Etapa E3. Essa trata de outro tipo de recurso lingüístico possível de ser identificado nas seções de artigos científicos, os marcadores discursivos.

4.9 Etapa de Anotação Automática de Marcadores Discursivos e Expressões Formulaicas

A Figura 4.10 apresenta uma visão geral da etapa de Anotação Automática de Marcadores Discursivos e Expressões Formulaicas, o qual recebe como entrada um córpus de textos considerados "bons", da área em questão, e produz como saída um córpus, cujos marcadores discursivos e expressões formulaicas aparecem destacados. Essa anotação é realizada automaticamente por meio de uma ferramenta computacional, que tem como base uma lista desses dois recursos lingüísticos, gerada a partir do córpus do Scipo-Farmácia e de fontes bibliográficas consultadas.

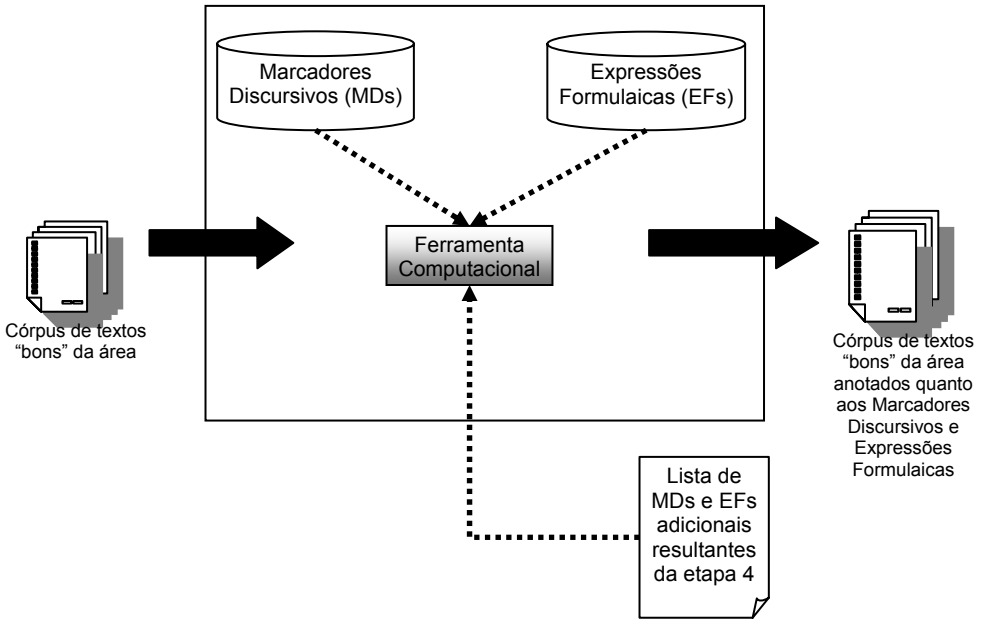


Figura 4.10: Diagrama da Etapa de Anotação Automática de Marcadores Discursivos e Expressões Formulaicas do córpus (M3).

4.9.1 Instrução da Etapa E3

Como pode ser observado nesse diagrama, o procedimento para se realizar a tarefa de marcação automática de um córpus é bem simples e se inicia com a submissão de um córpus constituído por textos bons (resultantes da avaliação da Etapa E2) a uma ferramenta computacional (anotador automático), que detecta automaticamente os possíveis marcadores discursivos e expressões formulaicas existentes nesses textos. Para tal, essa ferramenta tem como base uma lista de marcadores discursivos e de expressões formulaicas a qual pode ser acrescida com mais exemplos resultantes do processo manual contido na Etapa E4 do CECARL. Essa lista poderá ser incrementada com elementos retirados de material autêntico, assim como foi elaborada a lista acoplada ao anotador automático. Isso faz com que o usuário da ferramenta de suporte à escrita gerada com nosso processo tenha acesso apenas aos marcadores discursivos e expressões formulaicas que realmente ocorrem em determinadas seções de artigos científicos e, portanto, essenciais de serem conhecidos e utilizados na escrita de seções de novos artigos. A inserção desses elementos adicionais auxilia no aumento da precisão de anotação dos mesmos com os próximos córpus submetidos. Ao final dessa Etapa E3, o usuário tem como saída/produto do anotador automático o córpus submetido com os marcadores discursivos nele existentes destacados.

A Figura 4.15 traz a lista de Marcadores Discursivos organizados por funções que podem desempenhar em textos científicos, como contraste/oposição, adição, conseqüência/resultado, e assim por diante, utilizada no SciPo-Farmácia. A ferramenta gerada pelo CECARL também oferecerá esse tipo de recurso ao seu usuário.

Importante dizer que o objetivo dessa Etapa E3 são dois: (1) a identificação desses dois tipos de recursos lingüísticos auxilia posteriormente na anotação das estratégias retóricas, que deverão ser anotadas manualmente na Etapa E5 e (2) as listas de marcadores discursivos e expressões formulaicas utilizadas na base dessa ferramenta de anotação automática serão retornadas ao usuário em forma de listas organizadas por funções que desempenham (ver Figura 4.11), bem como destacados em seus respectivos contextos de uso, isto é, nos textos em que foram encontrados (ver Figura 4.12). Assim, o usuário terá exemplos de como, com quais palavras e em que momento da sentença esses recursos aparecem, e o que é melhor: observando exemplos autênticos da linguagem científica em seu contexto de uso.

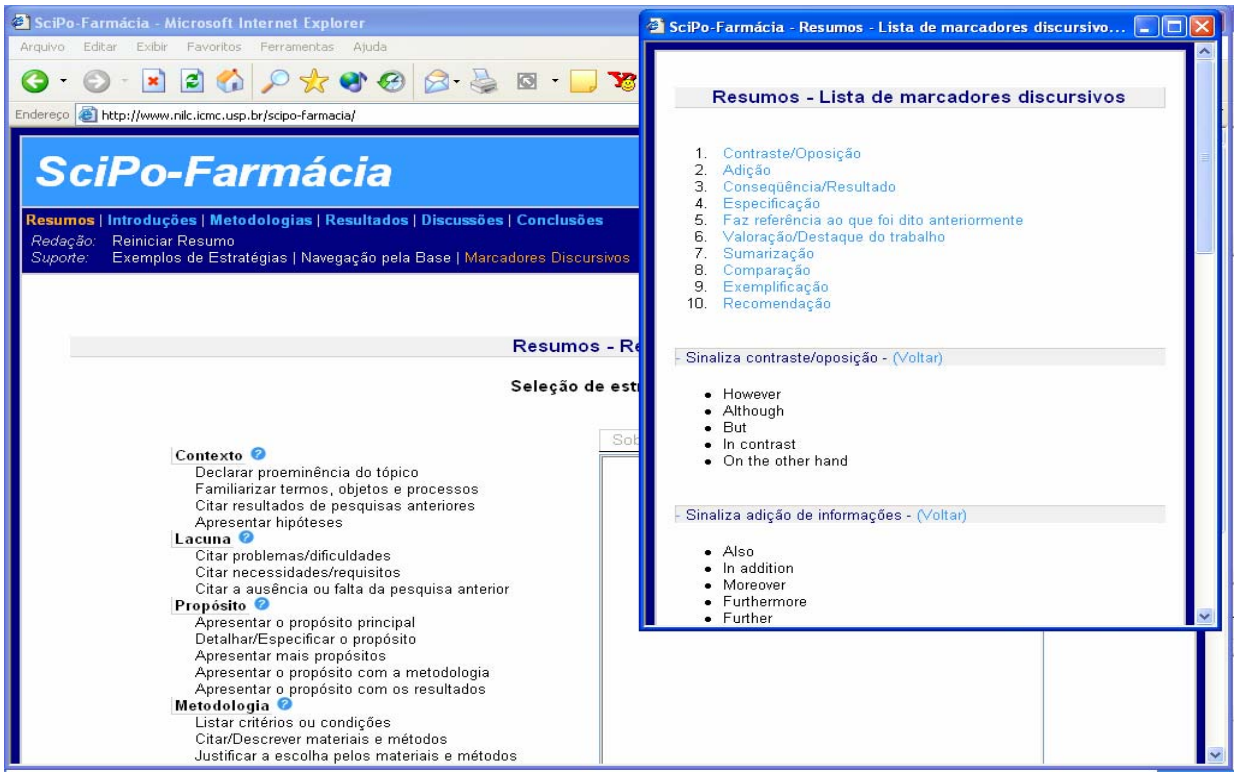


Figura 4.11: Marcadores Discursivos organizados por funções que podem desempenhar em textos científicos do SciPo-Farmácia.

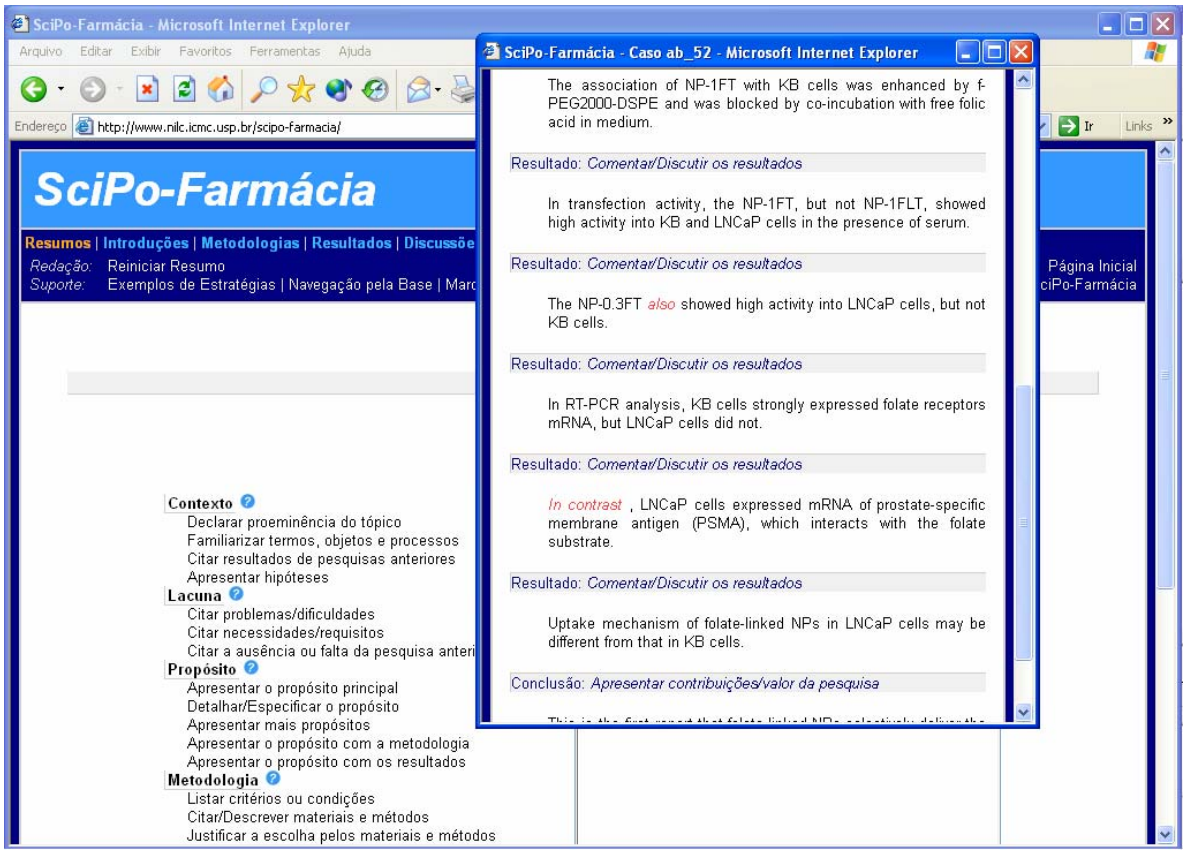


Figura 4.12: Como pode ser observado, os marcadores discursivos destacados (em vermelho) aparecem em seu contexto de uso, isto é , nas sentenças em que foram encontrados. Em azul, estão as funções que desempenham nessas sentenças.

4.9.2 Instanciação da Etapa E3

Em relação ao nosso córpus Met, essa tarefa foi realizada manualmente, uma vez que esse tipo de ferramenta ainda não foi implementada.

O Apêndice 3 reúne sob as teorias estudadas a respeito de Marcadores Discursivos (Quirk *et al*, 1985 e Fraser, 1993; 1995; 2005) os marcadores retirados do córpus Met, bem como os outros marcadores também retirados de textos de outras seções da área de Farmácia. Estes serão inseridos tanto no SciPo-Farmácia, na parte de navegação de Marcadores Discursivos da seção "Metodologia", como também na ferramenta de anotação automática da estrutura esquemática de tal seção.

Interessante dizer que do Córpus Met foram retirados 103 advérbios. Um número alto quando comparado ao número de advérbios encontrados nas outras seções de artigo que compõem o SciPo-Farmácia. Além deles, foram encontrados mais sete marcadores discursivos, também adicionados à lista acima. O restante dos outros marcadores adicionados na lista acima já constituíam a base do SciPo-Farmácia e acabaram também sendo

incorporados. Importante também dizer que, conforme é afirmado por Houaiss em seu Dicionário Houaiss da Língua Portuguesa, os advérbios são uma classe de palavras de difícil definição pela variedade de comportamentos sintáticos, peculiaridades semânticas, divergências de funções e classificações duvidosas que abrange. Portanto, é natural haver divergências quanto à classificação feita acima.

A próxima etapa a ser apresentada diz respeito a um passo manual de revisão dos recursos lingüísticos gerados pelas etapas anteriores, a E1 ou E1', o E2 e o E3.

4.10 Etapa de Revisão Manual da Estrutura Esquemática, Marcadores Discursivos, Expressões Formulaicas e da Qualidade dos textos

A Figura 4.13 apresenta uma visão geral da etapa de Revisão Manual, a qual recebe como entrada um córpus com textos anotados quanto aos marcadores discursivos, expressões formulaicas e estruturas esquemáticas, e se realiza uma avaliação/revisão manual desses recursos lingüísticos. Ao final desse processo, pode ser obtida mais uma triagem de textos, separando-os em bons e ruins, bem como uma lista de marcadores discursivos e expressões formulaicas que poderão ser adicionadas à ferramenta computacional que as anota automaticamente.

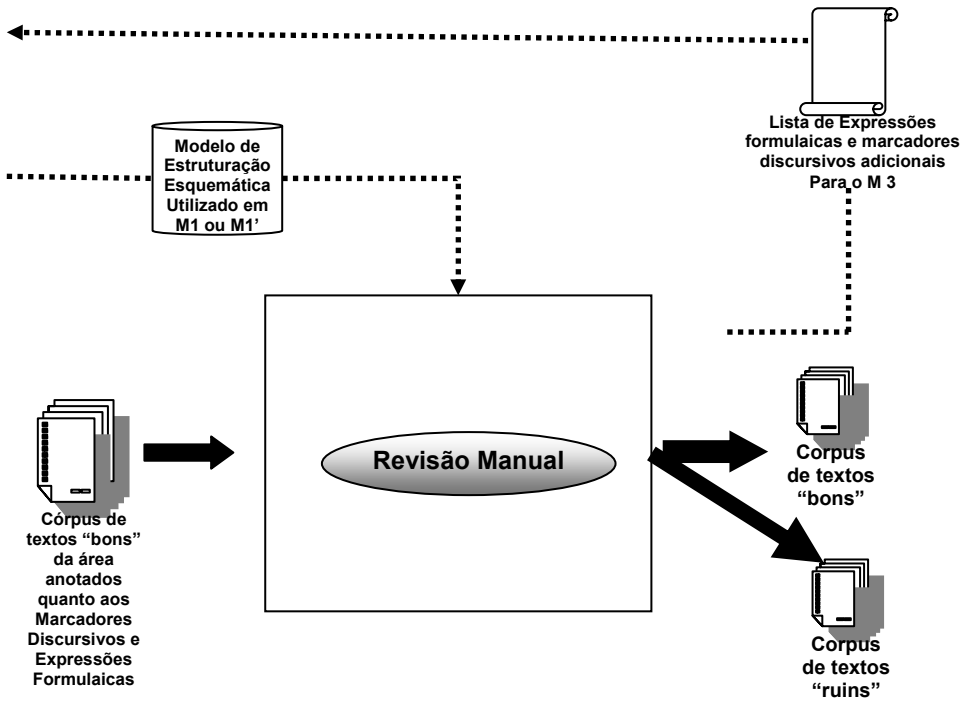


Figura 4.13: Diagrama da etapa de revisão manual da qualidade dos textos e das partes reutilizáveis.

4.10.1 Instrução da Etapa E4

Nessa etapa, temos a tarefa manual de correção de eventuais falhas cometidas pelos processos automáticos contidos nas etapas anteriores, bem como fornecer insumo para essas mesmas ferramentas melhorarem sua precisão, conforme pode ser observado na Figura 4.13.

A primeira correção sugerida é quanto às estruturas esquemáticas, isto é, avalia-se se as siglas utilizadas pelo categorizador automático estão indicando corretamente as macro-funções (estruturas esquemáticas) de cada uma das sentenças contidas nos textos. A seguir, é feita uma revisão da quanto à identificação e posterior anotação de possíveis marcadores discursivos e expressões formulaicas que apareceram no texto, mas que não constavam da lista desses elementos já existente na ferramenta computacional que os anota automaticamente. Por último, é realizada uma avaliação quanto à qualidade dos textos. Nessa etapa, é avaliado se os textos resultantes do processo automático de avaliação da qualidade textual são realmente bons para comporem o córpus, utilizando para isso, a própria rubrica de forma manual.

Poder-se-ia perguntar o porquê de se revisar manualmente essas partes reutilizáveis dos textos, uma vez que, na maioria delas, se tem um ferramental computacional para tal tarefa. Uma possível resposta é o fato de já ter sido comprovado em experimentos anteriores com marcação de córpus, o fato de ser mais rápido e promover melhores resultados corrigir um texto anotado a ter que realizar a anotação em texto cru, isto é, sem anotação. Outro fator que vem reforçar essa necessidade da revisão manual dos processos automáticos realizados é o fato dessas anotações serem atualmente realizadas com precisão não tão alta, sendo natural a ocorrência de eventuais falhas. No entanto, além dessa revisão assegurar uma melhor qualidade para os textos do córpus destinados a ferramenta de auxilio a escrita, a revisão também servirá de insumo para o ferramental computacional melhorar a precisão de suas tarefas, uma vez que listas de elementos não contidos nas bases dessas ferramentas poderão ser inseridas após essa análise.

Vale dizer, que o mesmo modelo de estrutura esquemática utilizado na Etapa E1 ou E1' deve ser o mesmo a ser utilizado nessa revisão, como pode ser observado no diagrama 4.16.

4.10.2 Instanciação da Etapa E4

Em relação ao nosso córpus Met, como a ferramenta computacional que indica a qualidade dos textos ainda não se encontrava totalmente desenvolvida, a avaliação da

qualidade dos textos foi feita com o auxílio de um especialista da área de Ciências Farmacêuticas e um especialista em escrita científica. As expressões formulaicas encontradas em nosso córpus Met foram as seguintes: *In a typical experiment; For better illustration; In some experiments; At the same time; Unless otherwise stated; As a consequence.*

A seguir serão apresentadas os passos contidos na próxima etapa, para a anotação manual das estratégias retóricas de um córpus.

4.11 Etapa de Anotação Manual das Estratégias Retóricas

Na Figura 4.14 temos uma descrição dos procedimentos manuais envolvidos na tarefa de anotação das estratégias retóricas de um córpus. Para tanto, é necessário um modelo teórico que descreva esses tipos de estratégias em uma dada seção de artigo científico, um manual de anotação para que a forma, os procedimentos e exemplos de textos anotados sejam consultados sempre que haja necessidade.

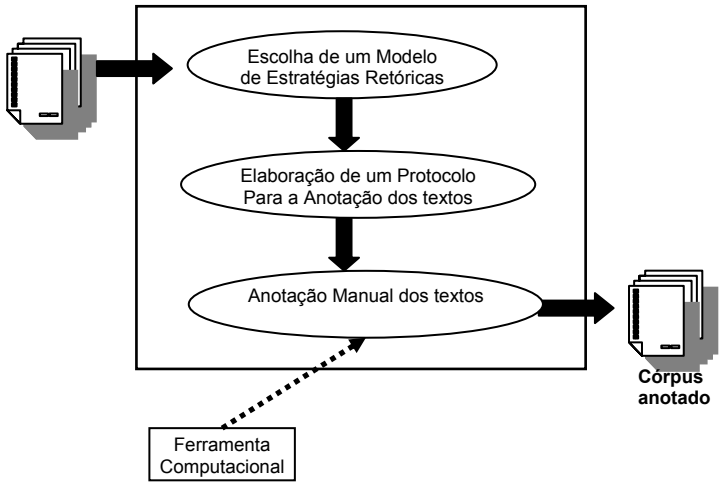


Figura 4.14: Diagrama da Etapa de Anotação Manual das Estratégias Retóricas de um córpus.

4.11.1 Instrução da Etapa E5

Conforme pode ser observado na figura 4.14, é uma etapa manual de anotação das estratégias retóricas dos textos de um córpus. Para tanto, é necessário escolher um modelo de estratégias retóricas para anotar os textos, elaborar um protocolo de anotação desses textos, isto é, uma descrição das categorias a serem classificadas, com seus respectivos exemplos, e, por fim, realizar a anotação manual propriamente dita dessas estruturas, com o auxílio de um

editor de textos, como o *Word* da *Microsoft*, ou de uma ferramenta computacional, com funcionalidades semelhantes às apresentadas pela ferramenta *TagAlign* (Caseli *et al*, 2002) apresentada na seção que descreve a Etapa E1'. Apesar de manual, pode ser interessante a existência de uma ferramenta computacional que auxilie nessa tarefa de categorizar as estratégias retóricas contidas nas sentenças, bastando um clique em um dado botão dessa ferramenta para se ter a etiqueta referente à estratégia retórica anexada à sentença.

4.11.2 Instanciação da Etapa E5

Em nosso córpus Met, essa anotação foi realizada por quatro juízes (anotadores) a fim de assegurar maior confiabilidade na anotação das partes dos textos e adotamos como modelo de anotação, o esquema de Passos (*Steps*) proposto por Swales (1990), que é apresentado na Tabela 4.7.

Siglas das categorias	Descrição das siglas
MAT-LIST	Listagem dos materiais utilizados no estudo
MAT-FONT	Detalhamento da Fonte dos materiais utilizados
MAT-INFO	Fornecimento de informações a respeito dos materiais
PRO-DOC	Detalhamento dos procedimentos utilizados para a execução correta da metodologia
PRO-DET	Detalhamento dos procedimentos utilizados
PRO-JUST	Fornecimento de justificativa sobre os procedimentos
EQU	Equipamentos utilizados
PAD	Procedimentos de análise dos dados
RES	Resultados da Pesquisa

Tabela 4.7: Note que a sigla da categoria é composta sempre por letras contidas em suas respectivas categorias, de forma a facilitar a memorização e fácil identificação do significado da categoria que deverá ser empregada nas sentenças do córpus.

Antes de realizar a anotação, os quatro anotadores receberam um manual que descreve o modo como esse processo deveria ser feito e depois de esclarecidas algumas dúvidas, iniciaram o procedimento de familiarização com o modelo adotado e com os textos a serem trabalhados, semelhantemente ao procedimento descrito na etapa E1', porém nesse momento, estão em foco a anotação das estratégias retóricas. Entre as dificuldades sentidas no processo de anotação dessas estratégias, a maior delas estava relacionada ao fato de uma sentença possuir ao mesmo tempo mais de uma estratégia retórica e esse fato ser identificado por todos os anotadores. Nem sempre os juízes identificavam todas as estratégias contidas, o que contribuiu para haver queda de concordância entre eles.

O processo de anotação dessas estratégias consistiu na anotação de agrupamentos de palavras/estratégias retóricas que pudessem ser reutilizadas em contextos distintos dos quais foram retiradas. Essa atividade, no entanto, não consiste em plágio, uma vez que sentenças completas não são reutilizadas, as informações factuais dessas sentenças, isto é, aquelas informações que dizem respeito a um dado experimento, não são anotadas/compiladas. Essas partes não reutilizáveis consistem nas lacunas a serem preenchidas com a parte factual do experimento do autor do novo artigo a ser escrito.

Assim, o autor do artigo poderá construir seu texto montando peças, isto é, por meio da identificação de diferentes combinações das estratégias que aparecem no texto original e criando, a seguir, sua própria combinação.

Assim como foi realizado um teste estatístico para avaliar o grau de concordância na anotação das estruturas esquemáticas do córpus Met, o mesmo teste foi realizado para avaliar o grau de concordância/discordância entre os anotadores em relação à anotação das estratégias retóricas do córpus anotado. Em nosso estudo com o córpus Met, o valor de *k* obtido para essa tarefa foi K=0.676 (mais informações sobre a estatística *Kappa* aqui aplicada ver Capítulo 5), o que significa dizer que houve uma boa concordância entre os anotadores. Há que se considerar para esse resultado obtido que apesar da subjetividade envolvida na tarefa e o fato de haver sentenças no córpus que possuíam mais de uma função retórica (que nem sempre era detectada por todos os anotadores), o valor *k* obtido, ainda assim, foi um bom resultado. Fato esse que mostra que o modelo de estruturação adotado em nosso trabalho foi útil para a classificação das estratégias retóricas contidas no córpus Met, as quais contribuirão enquanto exemplos que serão utilizados na seção "Metodologia" de artigos científicos da área de Ciências Farmacêuticas.

Uma dúvida que pode surgir nesse momento do trabalho diz respeito ao tamanho do córpus que se deve construir. Em geral, o tamanho de um córpus depende sempre do propósito a que ele serve. Portanto, um córpus deve ser grande o bastante para conter ocorrências dos elementos de linguagem que se queira estudar/analisar. Tribble (1997) afirma que um pequeno córpus composto por 25.000-30.000 palavras pode ser adequado à maioria dos propósitos educacionais. Vale lembrar, que uma das características mais importantes de um córpus destinado a uma ferramenta de auxílio à escrita é o de conter uma boa quantidade de estratégias utilizadas em cada componente de cada seção constitutiva de um artigo científico, uma vez que servirão de base de exemplos de consulta para o usuário desse tipo de ferramenta. Portanto, devem estar bem representados em termos de quantidade e de qualidade. Em nosso córpus Met, as estatísticas referentes às estratégias retóricas (nove estratégias)

encontradas para cada componente da estrutura esquemática (cinco componentes) podem ser observadas na tabela 4.8.

Estratégias Retóricas	**Número de exemplos**
PRO-DET	644
MAT-FONT	132
PRO-DOC	116
PAD	96
EQU	77
MAT-INFO	77
PRO-JUST	66
RES	36
MAT-LIST	12

Estruturas Esquemáticas	**Número de exemplos**
PRO	829
MAT	221
PAD	96
EQU	77
RES	36

Tabela 4.8: Estatísticas das estratégias retóricas e estruturas esquemáticas contidas em 30 textos de Metodologia das Ciências Farmacêuticas.

Como as seções do SciPo-Farmácia foram analisadas separadamente quanto a sua adequação aos modelos de estrutura e qualidade de escrita, elas possuem quantidade de material textual diferente, nem sempre vindo de uma mesma publicação. Atualmente, a base contém 43 Resumos, 39 Introduções, 26 Resultados, 11 Discussões e 22 Conclusões. A média de textos contidos no SciPo-Farmácia pode ser observada abaixo:

Total de textos do SciPo-Farmácia => 171 textos - 100%
Resumos => 43 textos - 25.14%
Introduções => 39 textos - 22.80%
Resultados => 26 textos - 15.20%
Discussões => 11 textos - 6.43%
Conclusões => 22 textos - 12.86%
Metodologia => 30 textos - 17.54%

Trinta textos podem parecer um volume grande para ser anotado manualmente, no entanto, não é pequeno para constituir uma ferramenta de auxílio à escrita. É por esse motivo que existe o trabalho de um mestrando do ICMC-USP, cujo objetivo, conforme apresentado na etapa E1, é automatizar o processo de anotação de componentes esquemáticos de um córpus, ainda que se precise revisar posteriormente. Revisar ainda é mais fácil e rápido que

anotar manualmente, como já foi comprovado em experiências de anotação morfossintática no projeto Lacio-Web[14].

No entanto, vale lembrar que existem dois objetivos em nosso trabalho com o córpus da seção "Metodologia" o córpus Met. Além de disponibilizar tal córpus para uso nessa seção do ambiente, isto é, compilar esse córpus e extrair os recursos lingüísticos existentes, também é objetivo deste trabalho, aplicar a estatística *Kappa* para verificar se as tarefas de anotação realizadas com ele foram facilmente entendida pelos anotadores. Fato comprovado pelos valores satisfatórios obtidos.

4.12 Etapa de Extração Automática de Termos

A Figura 4.15 mostra que para se obter a extração automática de termos de um córpus, basta submetê-lo a uma ferramenta computacional que extraia automaticamente esse tipo de informação lingüística e que fornecerá como saída uma lista de termos que serão adicionados, na Etapa E7, a um concordanciador.

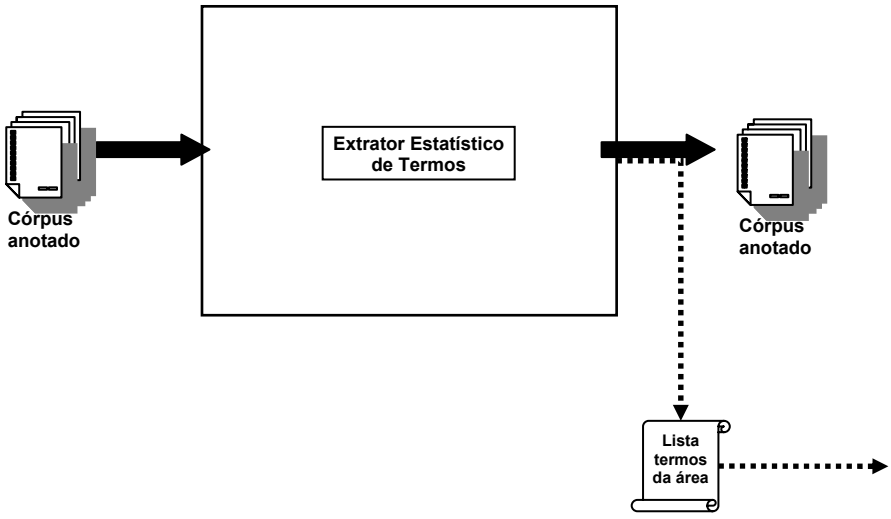


Figura 4.15: Diagrama da etapa de Extração Automática de termos da área a qual pertence o córpus.

Conforme é observado na Figura 4.15, é efetuada nessa etapa é efetuada a submissão dos textos do córpus a uma ferramenta computacional que extrai termos específicos da área a qual o córpus pertence. Em seguida, esses termos são submetidos a Etapa E7, que os retornará em seu contexto de uso, por meio de um concordanciador.

[14]http://www.nilc.icmc.usp.br/lacioweb/index.htm

Existem três tipos de abordagens para se realizar essa tarefa de extração de termos de uma dada área de especialidade: a abordagem estatística, a lingüística e a híbrida que combina as duas anteriores (Teline, 2004). Entre essas abordagens de extração, a estatística foi escolhida para o nosso trabalho por ser o tipo mais simples de ser utilizado segundo a literatura.

Entre os métodos estatísticos existentes podem ser citados as medidas estatísticas do pacote NSP (*N-gram Statistics Package*)[15], escrito em linguagem *Perl*, que foi implementado por Ted Pedersen, Satanjeev Banerjee e Amruta Purandare na Universidade de Minnesota, Duluth. Ele é constituído por um conjunto de programas que auxilia na análise de n-gramas[16] em arquivos texto. Outro método é o BootCaT (do inglês *Bootstrapping Corpora and Terms*)[17], que é composto por várias ferramentas escritas em linguagem Perl que foram projetadas para executar pequenas partes do processo de extração automática de córpus e de termos. Uma terceira ferramenta que poderia ser citada para esse tipo de tarefa de extração de termos é a *KeyWords*, parte integrante da suíte de ferramentas *WordSmith Tools* (Scott, 1998), a qual segundo Berber-Sardinha (1999a), tem sido referência para vários estudos e investigações sobre linguagem.

Entre esses três métodos estatísticos citados, o terceiro foi o escolhido para ser aplicado neste projeto, pois considerando que será um método a ser executado não só por lingüistas já acostumados ou familiarizados com o *WordSmith Tools*, mas também por pesquisadores de outras áreas do conhecimento, portanto nem tão cientes da existência ou até mesmo do tipo de uso que se pode fazer com esse ferramental computacional para a extração, no nosso caso, de termos. Entre as razões dessa escolha pode ser citado o fato de ser um programa que é executado no ambiente *Windows*, familiar para a maioria dos usuários, e pode ser obtido pela Internet mediante pagamento de licença. Ser executável no ambiente *Windows* significa não só uma interface amigável de interação, mas também a ausência de necessidade de linhas de comandos, que por mais simples que possam ser, podem intimidar o interessado em realizar a tarefa de extração de termos de um córpus.

Conforme já mencionado, a ferramenta *KeyWords* se destina à comparação de listas de palavras de um córpus de estudo com uma lista de palavras de um córpus de referência. O resultado desta comparação é uma lista de palavras-chaves, cujas freqüências no córpus de

[15] http://www.d.umn.edu/~tpederse/nsp.html.
[16] Termo composto por uma ou mais palavras, por exemplo: uni-grama:termo formado por uma palavra, bi-grama termo formado por duas palavras. Pode ser também referenciado como multipalavra.
[17] http://sslmit.unibo.it/~baroni/bootcat.html

estudo são diferentes do córpus de referência. Em outras palavras, sua função é comparar por meio de um método estatístico as palavras cujas freqüências no córpus de estudo são maiores do que no córpus de referência, que deve ser representativo. Assim, os principais componentes na extração das palavras-chaves são: 1) um córpus de estudo, representado por uma lista de freqüência de palavras, e 2) um córpus de referência, também representado como uma lista de freqüência de palavras, cuja função é a de fornecer o conjunto de palavras com o qual se fará as comparações.

Portanto, após a extração de termos realizada na Etapa E0, o próximo passo para o usuário do CECARL é reutilizar essa mesma lista de palavras-chaves anteriormente gerada como recurso para a composição de um concordanciador voltado para o córpus de especialidade. A razão da existência de tal concordanciador pode ser justificada pela facilidade que o usuário terá de acessar, sempre que preciso, os termos de especialidade de sua área dentro de seu contexto de uso, ou seja, nas possíveis realizações lingüísticas contidas no córpus compilado. Esse contexto de uso permite verificar os colocados existentes no córpus, o local dentro de uma frase onde regularmente dados termos, marcadores discursivos ou mesmo expressões formulaicas ocorrem, em geral. Pode ainda auxiliar na dúvida de uso de um dado verbo combinado com uma também dada preposição, permitindo aos estudantes fazerem suas próprias descobertas sobre linguagem (Johns, 1991a; Tribble & Jones, 1990; Swales & Lee, 2006) ao mesmo tempo em que são expostos a diferentes formas de linguagem geradas de diferentes perfis de autores, portanto advindas de diferentes conceitos de gramática, de estilo e de convenções lingüísticas.

Para isso, esse concordanciador possibilitará dois tipos de interação: uma voltada para os termos da área e outra para uma outra dada instância lingüística que o usuário desejar, ou seja, de outros elementos existentes no córpus. Assim, ao selecionar o botão "termos" desse concordanciador, o usuário terá a sua esquerda uma lista com as palavras-chaves do córpus, que após um clique em uma delas, aparecerá à sua direita, esse mesmo termo em seu contexto de uso. Se o usuário desse concordanciador optar pelo botão "todas as palavras do córpus", ele terá que digitar no campo de busca a palavra ou expressão que deseja ver na concordância. Importante ainda dizer que o texto completo do qual o termo ou expressão foi retirado aparece na tela do concordanciador quando se realiza um clique na palavra nódulo do concordanciador. A Figura 4.16 ilustra esse concordanciador descrito (ainda idealizado).

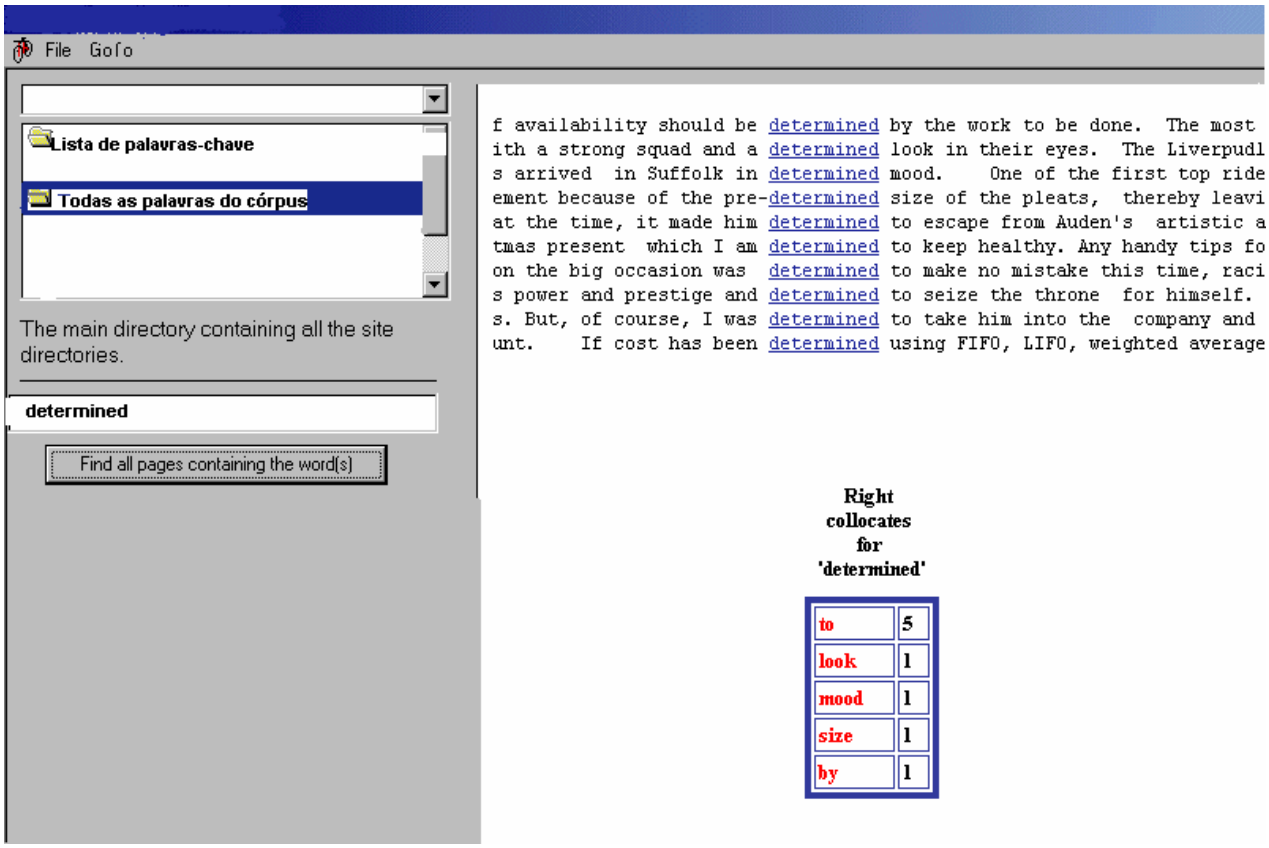


Figura 4.16: Figura montada de um Concordanciador idealizado para compor parte da ferramenta de auxílio à escrita criada pelo ambiente Web gerador de ferramentas.

Vale dizer que esse concordanciador será gerado, futuramente com a contribuição de um outro trabalho, junto da ferramenta de auxílio à escrita científica construída pelo ambiente Web gerador, o projeto maior que inclui o projeto em tela.

A próxima e última etapa a ser apresentada trata do acoplamento e submissão de todos os recursos lingüísticos que foram produzidos até o momento a uma ferramenta de auxílio à escrita científica genérica, o Scientific Writing.

4.13 Etapa de Inclusão dos Recursos Lingüísticos gerados em uma ferramenta genérica

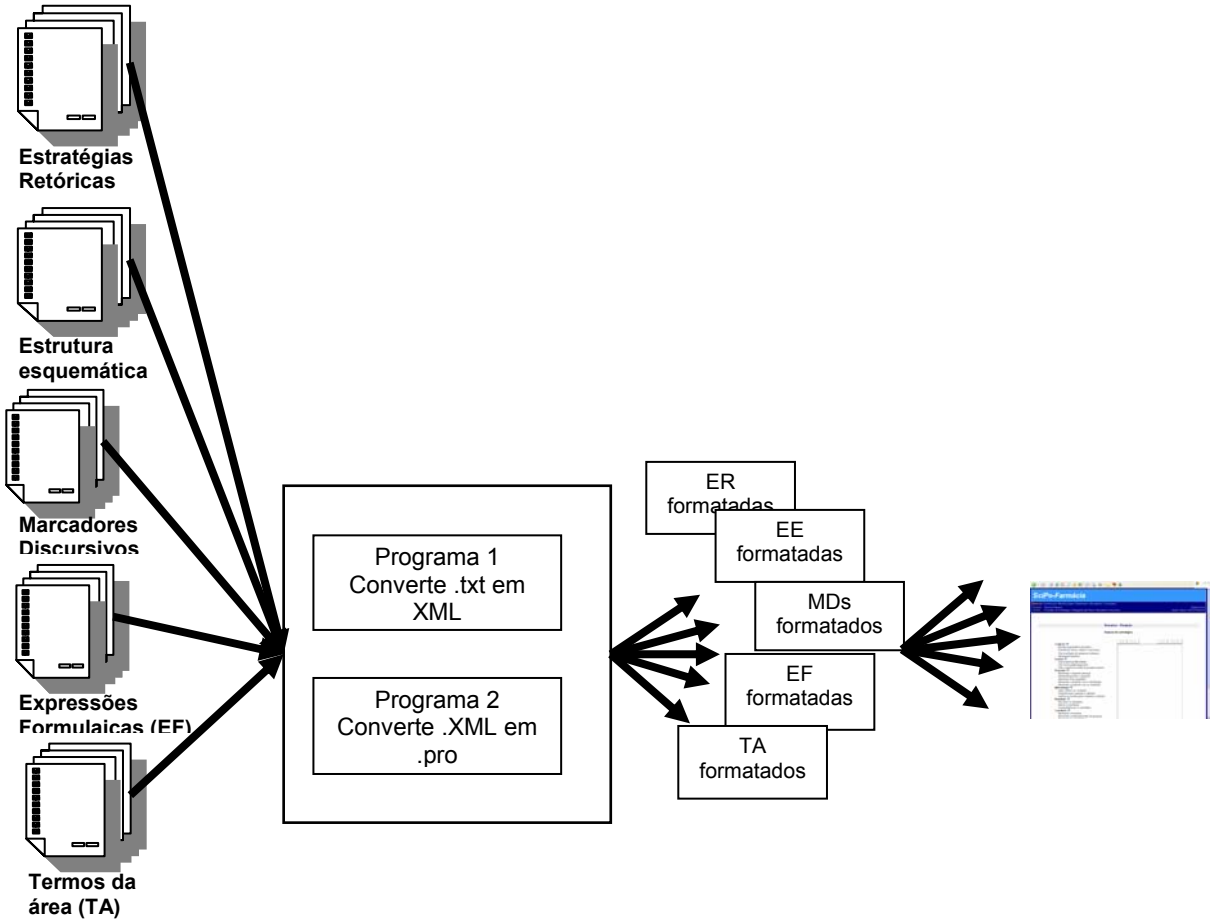


Figura 4.17: Diagrama da etapa de Inclusão dos Recursos Lingüísticos em uma ferramenta do tipo SciPo-Farmácia genérica.

Nessa última etapa de nosso processo, ocorre a **formatação** dos recursos lingüísticos produzidos para sua posterior inclusão em uma ferramenta genérica, o Scientific Writing. Esta inclusão deve, entretanto, ser realizada por um profissional da computação que possui as devidas permissões de acesso no servidor que abrigará a ferramenta de suporte à escrita. Futuramente, este trabalho será feito automaticamente pelo ambiente Web gerador de ferramentas.

Os recursos necessários ao Scientific Writing, ferramenta semelhante ao SciPo-Farmácia, devem ser alocados em um dado diretório de um servidor (no caso da ferramenta que possui conhecimento das Ciências Farmacêuticas usamos o diretório SciPo-Farmácia) e são divididos em 7 tipos:

1) **A base de casos em XML**. Cada seção possui um diretório para guardar os casos e estes devem possuir os nomes: Resumos, Introduções, Metodologias, Resultados, Discussões e Conclusões. Dentro destes os arquivos possuem extensões ".xml". Estes casos são usados

para exibição, por exemplo, na "Navegação pela Base". A Figura 4.18 mostra o texto codificado de um abstract composto de 6 orações (**caso ab_04),** que possui um único marcador discursivo anotado com etiquetas XML (<Marcador>**here**</Marcador>). Este texto é gerado automaticamente a partir de um arquivo no formato txt que possui uma oração por linha precedida do nome de seu componente e da sua estratégia retórica.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <Abstract id="ab_04">
  <Referencia>http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=152239
     Investigating antibody-catalyzed ozone generation by human neutrophils Bernard M.
     Babior, Cindy Takeuchi, Julie Ruedi, Abel Gutierrez, and Paul Wentworth,
     Jr.</Referencia>
- <Subcomponente>
  <Nome>Contexto</Nome>
- <Estrategia>
  <Nome>Citar resultados de pesquisas anteriores</Nome>
     Recent studies have suggested that antibodies can catalyze the generation of previously
     unknown oxidants including dihydrogen trioxide (H2O3) and ozone (O3) from singlet
     oxygen (1O.2) and water.
     </Estrategia>
     </Subcomponente>
- <Subcomponente>
  <Nome>Contexto</Nome>
- <Estrategia>
  <Nome>Apresentar hipóteses</Nome>
     Given that neutrophils have the potential both to produce 1O.2 and to bind antibodies,
     we considered that these cells could be a biological source of O3.
     </Estrategia>
     </Subcomponente>
- <Subcomponente>
  <Nome>Propósito</Nome>
- <Estrategia>
  <Nome>Apresentar o propósito principal</Nome>
     We report
  <Marcador>here</Marcador>
     further analytical evidence that antibody-coated neutrophils, after activation, produce
     an oxidant with the chemical signature of O3.
     </Estrategia>
     </Subcomponente>
- <Subcomponente>
  <Nome>Resultado</Nome>
- <Estrategia>
  <Nome>Comentar/Discutir os resultados</Nome>
     This process is independent of surface antibody concentration down to 50% of the
     resting concentration, suggesting that surface IgG is highly efficient at intercepting
     the neutrophil generated 1O.2.
     </Estrategia>
```

</Subcomponente>
- <Subcomponente>
<Nome>**Metodologia**</Nome>
- <Estrategia>
<Nome>**Citar/Descrever materiais e métodos**</Nome>
Vinylbenzoic acid, an orthogonal probe for ozone detection, is oxidized by activated neutrophils to 4-carboxybenzaldehyde in a manner analogous to that obtained for its oxidation by ozone in solution.
</Estrategia>
</Subcomponente>
- <Subcomponente>
<Nome>**Conclusão**</Nome>
- <Estrategia>
<Nome>**Apresentar contribuições/valor da pesquisa**</Nome>
This discovery of the production of such a powerful oxidant in a biological context raises questions about not only the capacity of O3 to kill invading microorganisms but also its role in amplification of the inflammatory response by signaling and gene activation.
</Estrategia>
</Subcomponente>
</Abstract>

Figura 4.18: Abstract abs_04 em formato XML.

2) **A base de casos em Prolog.** Existe um programa que toma os textos dos casos em txt e gera um arquivo com extensão ".pro". Os arquivos resultantes possuem os nomes: Case_base_abstracts.pro, Case_base_introductions.pro, Case_base_methodologies.pro, Case_base_results.pro, Case_base_discussions.pro, Case_base_conclusions.pro. Todos estes seis arquivos devem estar localizados na raiz do diretório que abriga o SciPo-Farmácia da área específica. As buscas para recuperação de casos similares são executadas com estes arquivos.

Vale dizer, que as ferramentas computacionais utilizadas nos processos descritos nos itens acima, de conversão de txt para XML e de XML para .pro encontram-se disponíveis no ambiente Plonetarium da Fapesp: http://gen-writingtool.incubadora.fapesp.br/portal.

3) **As interfaces das estruturas esquemáticas e suas estratégias retóricas para cada uma das seções.** São também seis arquivos alocados na raiz do diretório que abriga a ferramenta SciPo-Farmácia da área específica: abstracts_lista_estr.txt, introductions_lista_estr.txt, methodologies_lista_estr.txt, results_lista_estr.txt, discussions_lista_estr.txt, conclusions_lista_estr.txt.

4) **Helps (textos de ajuda).** Os textos de ajuda ficam no diretório Ajuda e seguem os formatos: componente_estrategia.htm.inc ou componente.htm.inc.

5) **Lista de marcadores discursivos.** Cada seção possui sua lista particular que é alocada na raiz do diretório que abriga a ferramenta SciPo-Farmácia da área específica. Seguem o formato conclusions_consult_marcadores.html.inc.

6) **Texto Sobre.** Alocado na raiz do diretório que abriga a ferramenta SciPo-Farmácia da área específica, o arquivo sobre.php possui informações sobre o projeto e pesquisadores que desenvolveram os recursos lingüísticos. O texto da ferramenta SciPo_Farmácia é o seguinte:

> "O sistema SciPo-Farmácia foi um projeto realizado no NILC, sob a orientação da Profa. Sandra Maria Aluísio (sandra@icmc.usp.br) e do Prof. Osvaldo Novais de Oliveira Jr. (chu@if.sc.usp.br), em parceria com a Faculdade de Ciências Farmacêuticas da USP-São Paulo, particularmente com os professores Adalberto Pessoa Jr. (pessoajr@usp.br) e Ana Campa (anacampa@usp.br). A análise textual dos artigos da ferramenta foi realizada pela lingüísta Aline Maria Pacífico Manfrim e posteriormente avaliada pelo Prof. Osvaldo Novais de Oliveira Jr. e Profa. Sandra Maria Aluísio.
>
> Adaptado do projeto SciPo, trabalho de doutorado de Valéria D. Feltrim, intitulado "Suporte Computacional à Escrita Científica em Português", desenvolvido no NILC (ICMC - USP/São Carlos), sob a orientação da Profa. Dra. Maria das Graças Volpe Nunes (orientadora) e da Profa. Dra. Sandra Maria Aluísio (co-orientadora), volta-se para a escrita de todos os componentes de um artigo científico (resumos, introduções, metodologias, resultados, discussões e conclusões) tendo como língua-alvo o inglês.
>
> Nesse projeto, agradecemos a participação e o empenho de Valéria Feltrim (vfeltrim@icmc.usp.br), Lucas Antiqueira (lantiq@grad.icmc.usp.br) e Leandro Henrique Mendonça de Oliveira (leandroh@nilc.icmc.usp.br)".

7) **Texto Ajuda.** Alocado na raiz do diretório que abriga a ferramenta SciPo-Farmácia da área específica tal arquivo indica em linhas gerais o que deve ser esclarecido em cada uma das seções da estrutura de um artigo cientifico relatando uma pesquisa experimental; descreve a base de textos utilizados no sistema de suporte à escrita científica com as particularidades da área de conhecimento em termos de estruturação das seções e sua apresentação nos artigos que foram escolhidos para fazer parte da base; e mostra como utilizar o sistema para escrever as seções de um artigo.

Os textos Sobre e Ajuda aparecem na tela inicial da ferramenta de escrita, como mostra a tela na Figura 4.19.

Figura 4.19: Tela inicial do SciPo-Farmácia.

4.13.1 Instanciação da Etapa E7

Os recursos lingüísticos gerados do córpus Met passaram pelos procedimentos de formatação acima descritos para que pudessem ser inseridos no SciPo-Farmácia. A Figura 4.20 ilustra as estruturas esquemáticas e as estratégias retóricas criadas para a seção "Metodologia" desse ambiente. A Figura 4.21 mostra a navegação pela base de casos também da seção "Metodologia".

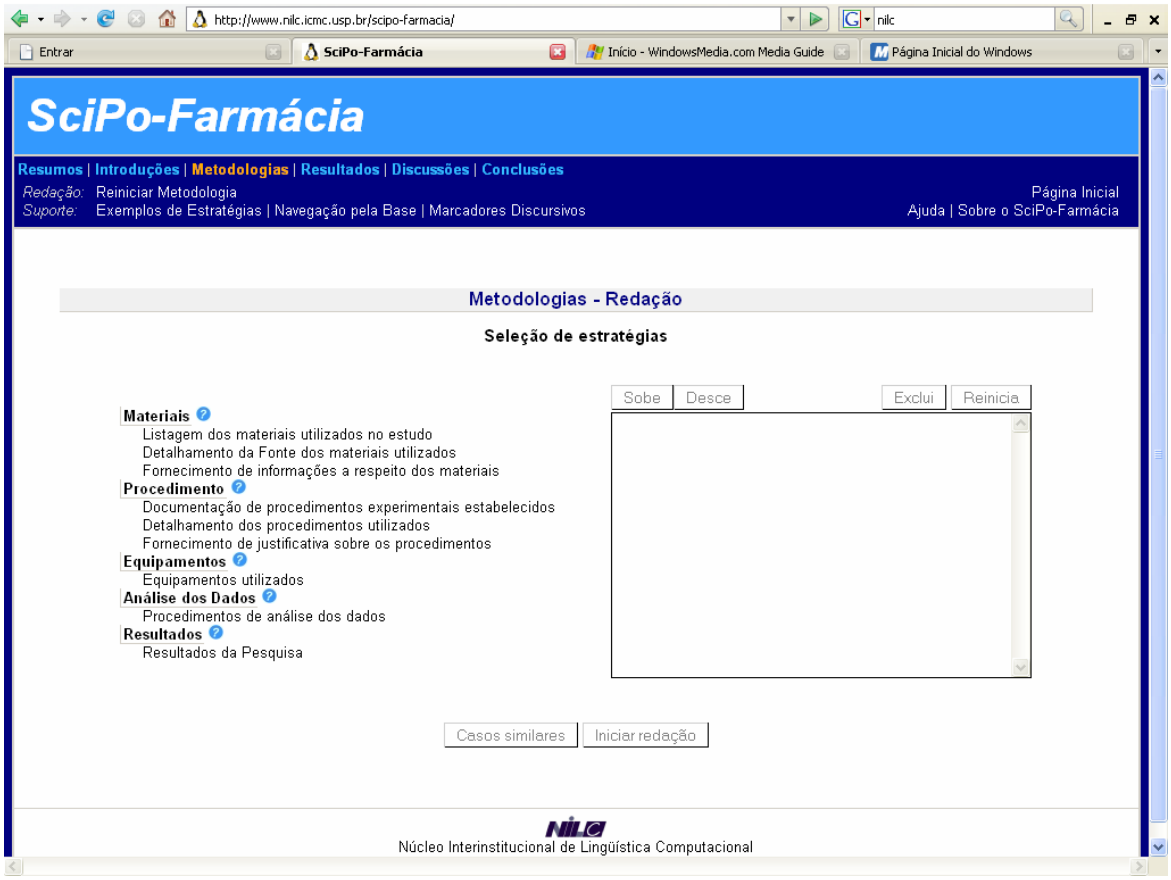


Figura 4.20: Estruturas esquemáticas e estratégias retóricas.

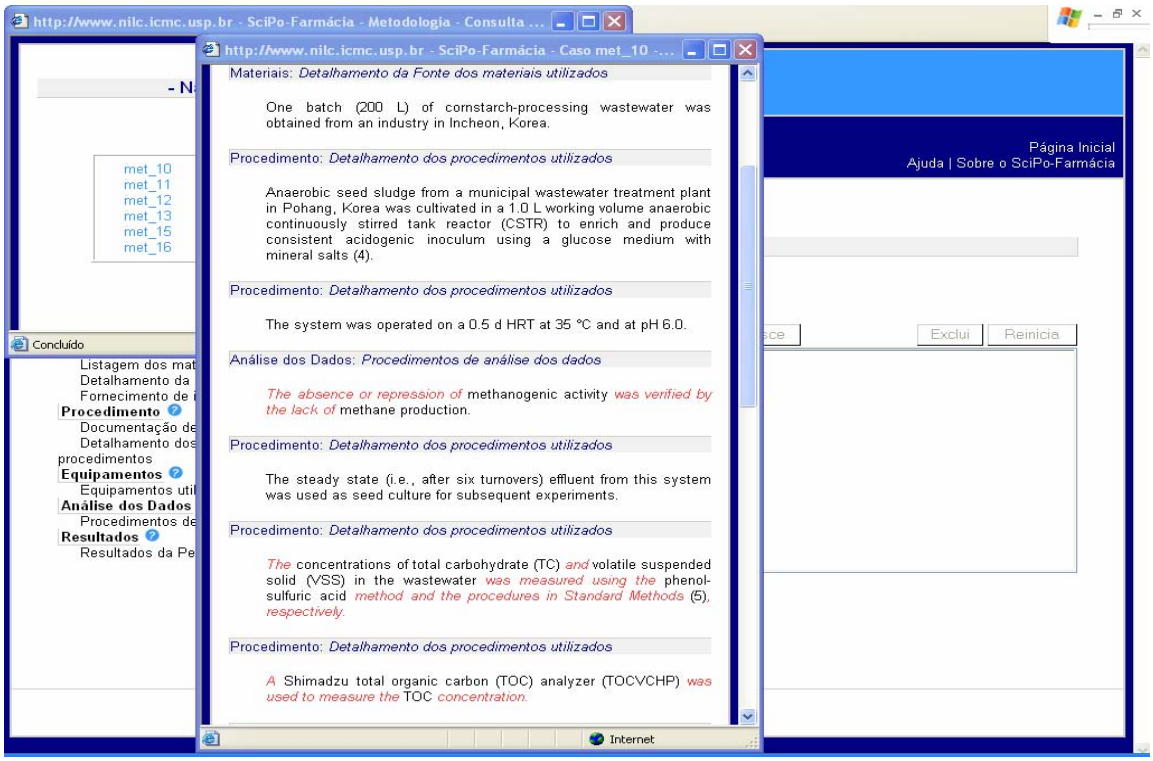


Figura 4.21: Exemplo de texto da base de casos de Metodologias do SciPo-Farmácia.

As figuras 4.22 e 4.23 Mostram a interface da ferramenta genérica, semelhante ao SciPo-Farmácia, que será disponibilizada no *site* de divulgação desta pesquisa, http://gen-writingtool.incubadora.fapesp.br/portal/, para que o usuário do CECARL possa fazer seu *download*.

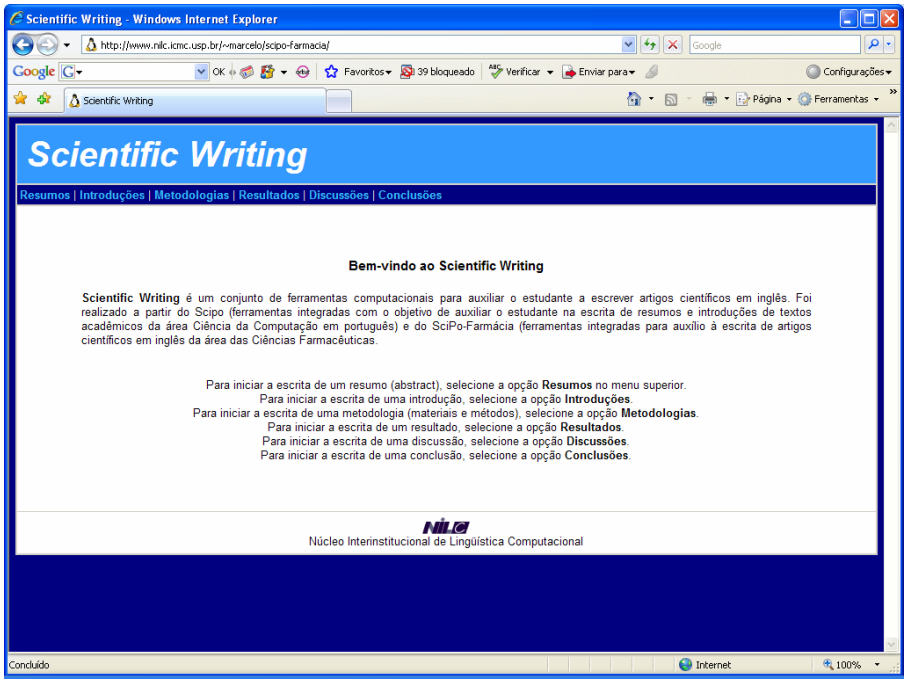


Figura 4.22.: Tela inicial do Scientific Writing, ferramenta de suporte à escrita genérica disponibilizada junto do processo proposto, e que foi inspirada no SciPo-Farmácia.

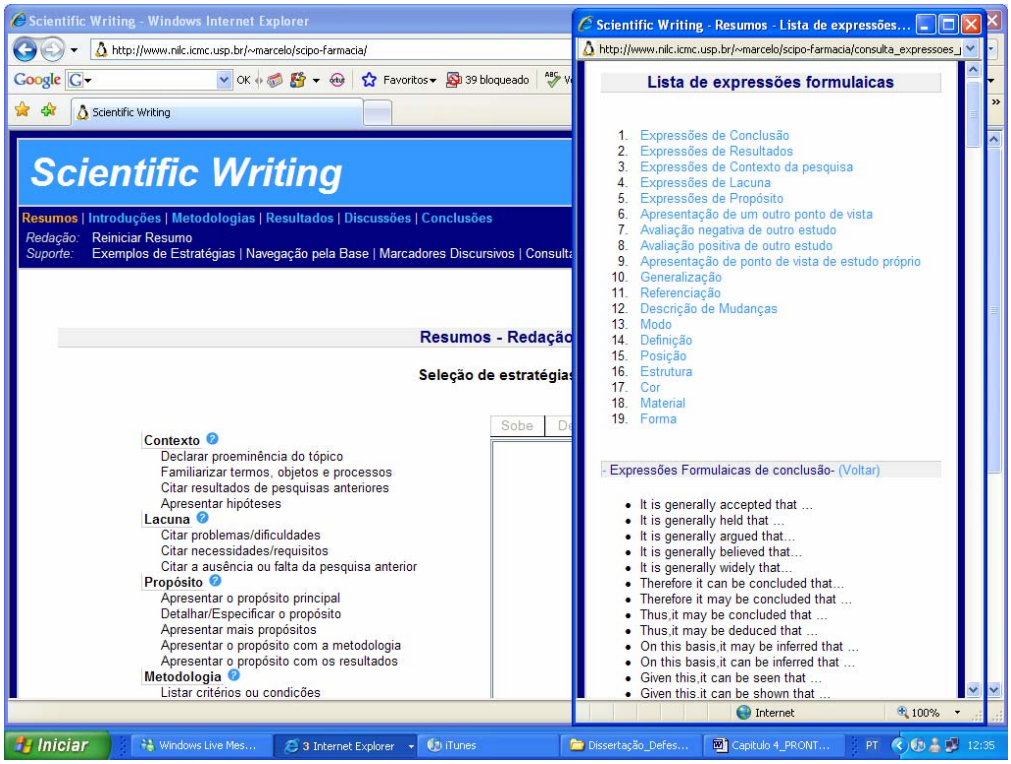


Figura 4.23: Tela do Scientific Writing com lista de expressões Formulaicas

4.14 Considerações Finais

Conforme dito anteriormente, a motivação deste capítulo foi mostrar as etapas de um processo para a construção de recursos lingüísticos aplicáveis em ferramentas de suporte à escrita em uma dada área de especialidade, possibilitando que pesquisadores de diferentes áreas possam utilizar os recursos de um sistema de auxílio à escrita científica personalizado para o domínio do conhecimento no qual se encontram inseridos.

Todos esses procedimentos foram realizados para que se obtivesse recursos lingüísticos confiáveis. Assim, é interessante que duas ou mais pessoas identifiquem nos textos os recursos lingüísticos existentes, ou quando sozinha, é aconselhável, no caso de dúvidas, consultar um especialista em escrita científica para que o mesmo possa avaliar a qualidade do material produzido.

5. Avaliação do Processo

Para avaliar se o processo proposto atinge seu objetivo, dividimos a avaliação em dois momentos. O primeiro visa averiguar a possibilidade de reprodução de etapas desse processo e esta possibilidade só se verifica se as instruções forem claras e completas. O segundo, por sua vez, visa avaliar a consistência da anotação realizada em três etapas escolhidas, bem como elaborar uma estimativa de tempo gasto, em média, na execução das tarefas pedidas.

As subseções que se seguem tratam justamente desses dois momentos, apresentados como Fase 1 e 2 de avaliação, respectivamente.

5.1 Fase 1 de Avaliação – Clareza e Completude das etapas descritas

O Capítulo 4 apresentou todas as etapas do processo proposto para a construção de recursos lingüísticos aplicáveis em uma ferramenta de auxílio à escrita científica que, como observado, contém 11 etapas. Cada uma delas é composta por diretrizes/passos (na maioria, apresentadas no formato de manuais), que devem fornecer informações suficientes para o cumprimento dos passos descritos.

Portanto, a motivação principal da primeira fase de avaliação foi analisar a *Clareza e a Completude dos manuais de construção de recursos lingüísticos* contidos nas etapas E1', E2', E3 e E5, responsáveis, respectivamente, pela anotação manual dos componentes da estrutura esquemática existentes em cada seção de artigo científico, pela aplicação manual de uma rubrica para a verificação da adequação (qualidade) de um artigo científico, pela anotação manual de marcadores discursivos e, por fim, pela anotação manual das estratégias retóricas também existentes em cada seção de um dado artigo científico. Essas etapas foram escolhidas, pois possuem conteúdo especializado, com muitas informações lingüísticas da área de Análise de Gêneros e de Lingüística Textual.

Avaliar a clareza e a completude desses manuais implica em observar se a informação neles contida é adequadamente apresentada ao público-alvo de nosso projeto de pesquisa, de forma que possam entender e cumprir com êxito as tarefas descritas nos manuais em avaliação.

Para esta avaliação foram escolhidas quatro pessoas, cujo perfil se assemelha ao do público-alvo do processo proposto quanto ao conhecimento científico e de língua inglesa. No entanto, somente três delas puderam nos auxiliar nessa primeira fase. No caso, trata-se de três alunos do programa de pós-graduação em Ciências da Computação da USP-São Carlos (um do segundo ano de mestrado, o outro do terceiro e o último nos últimos três meses de seu doutorado). A opção por escolher mais de uma pessoa da mesma área para realizar uma mesma tarefa é motivada pelo desejo de se assegurar maior confiabilidade quanto à avaliação dos resultados obtidos nas tarefas descritas pelos manuais.

Além das pessoas apresentadas, utilizaram-se também três manuais de identificação de diferentes recursos lingüísticos: um para a identificação dos componentes da estrutura esquemática e das estratégias retóricas de um resumo (Apêndice 2), uma vez que a identificação desses recursos lingüísticos foi feita simultaneamente, outro para a avaliação da qualidade desses resumos (Apêndice 4) e o último para a identificação de marcadores discursivos (Apêndice 3). Vale ainda dizer que manuais para a anotação dos componentes da estrutura esquemática e de estratégias retóricas das seções Introdução, Metodologia, Resultados, Conclusão e Discussão de artigos científicos podem ser encontrados, respectivamente, nos Apêndices 1, 5, 6, 7 e 8.

Para identificar esses recursos lingüísticos foram utilizados cinco resumos (que compõem 46 sentenças ao todo), da área de Ciências da Computação retirados de www.sciencedirect.com, qualis "A" na Capes. A seção Resumo foi escolhida por ser a menor seção de um artigo científico e que, portanto, pôde contribuir para agilizar o processo de avaliação.

Vale dizer que essas pessoas receberam os resumos prontos para a anotação, isto é, não precisaram compilar e formatá-los.

A identificação desses recursos foi realizada da seguinte forma: os componentes da estrutura esquemática e as estratégias retóricas foram identificadas por meio de siglas. Os marcadores discursivos foram destacados em negrito, e os novos, isto é, os não contidos no manual de anotação, mas que apareceram nos resumos, foram anotados ao final da página do resumo em análise. Quanto a avaliação da qualidade dos resumos, esta foi feita conforme os critérios sugeridos pelo manual de rubricas, com os valores sendo atribuídos ao final de cada resumo analisado.

Um exemplo de resumo anotado pode ser observado na Figura 5.1.

Link: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TYF-4KWK15W-1&_user=972067&_handle=V-WA-A-W-B-MsSAYVA-UUW-U-AAZDWEVDDC-AAZVYDCCDC-AUZUAVEWZ-B-U&_fmt=summary&_coverDate=10%2F31%2F2006&_rdoc=3&_orig=browse&_srch=%23toc%235617%232006%23998299985%23635649!&_cdi=5617&view=c&_acct=C000049650&_version=1&_urlVersion=0&_userid=972067&md5=89a3e813461623bf8e894e200019d498

CORRECTIVE FEEDBACK AND PERSISTENT LEARNING FOR INFORMATION EXTRACTION

Aron Culotta, Trausti Kristjansson, Andrew McCallum and Paul Viola

LAC-CNR To **successfully** embed statistical machine learning models in real world applications, two post-deployment capabilities must be provided: (1) the ability to solicit user corrections and (2) the ability to update the model from these corrections.

COT-FOP We refer to the former capability as corrective feedback and the latter as persistent learning.

LAC-CNR **While** these capabilities have a natural implementation for simple classification tasks **such as** spam filtering, we argue that a more careful design is required for structured classification tasks.

COT-FOP One example of a structured classification task is information extraction, **in which** raw text is analyzed to **automatically** populate a database.
PRO-APM/PRO-APP In this work, we augment a probabilistic information extraction system with corrective feedback and persistent learning components to assist the user in building, correcting, and updating the extraction model.
MET-CMM We describe methods of guiding the user to incorrect predictions, suggesting the most informative fields to correct, and incorporating corrections into the inference algorithm.
PRO-APM/PRO-AMP We **also** present an active learning framework that minimizes not only how many examples a user must label, but **also** how difficult each example is to label.
MET-CMM We **empirically** validate each of the technical components in simulation and quantify the user effort saved.
COC-AC We conclude that more efficient corrective feedback mechanisms lead to more effective persistent learning.
Rubricas:
1) Baixo: lacuna intercalada com contexto; propósito intercalado com metodologia; resultados ausentes.
2) Alto: apesar de existirem duas sentenças para propósitos (o estudo possui dois objetivos).
3) Alto: apesar de não serem mostrados os resultados (o resumo já foi penalizado na rubrica 1).

Figura 5.1: Exemplo de resumo anotado na primeira fase de avaliação por um dos colaboradores.

As siglas MET-CMM, COC-AC, e assim por diante, localizadas no início de cada sentença, fazem referência aos componentes da estrutura esquemática e às estratégias retóricas que cada sentença está desempenhando. A primeira parte de cada sigla, isto é, a parte à esquerda do hífen, como por exemplo, MET-, COC-, referem-se, respectivamente, ao componente da estrutura esquemática Metodologia e Conclusão. A segunda parte das siglas, que aparecem do lado direito do hífen, como por exemplo, -CMM e –AC, respectivamente, *Citar materiais e métodos* e *Apresentar conclusões*, referem-se à estratégia retórica apresentada pela sentença. Em negrito, estão as palavras que exercem o papel de marcadores discursivos nesse resumo, como *also, automatically, such as, in which, while e succesfully*. Ao final desse resumo, pode ser encontrada a averiguação de sua qualidade, sob o título de Rubricas. Os valores

"Baixo", "Alto" e "Baixo" indicados correspondem à avaliação feita pelo anotador desse texto, segundo os critérios sugeridos pelo manual que consultou sobre qualidade/adequação de um resumo científico. À frente desses valores, está a justificativa de atribuição de tal valor. Nesse resumo, entretanto, não houve a sugestão de possíveis marcadores discursivos ausentes no manual de anotação.

Além disso, essas pessoas ficaram livres, isto é, não receberam nenhum tipo de questionário para identificarem eventuais dificuldades ou falhas ao realizarem as etapas descritas nos manuais entregues. Por exemplo, trechos de texto que acharem confusos, a existência de termos técnicos/específicos empregados de maneira confusa ou com falta de informações, se há informação insuficiente para realizar uma dada tarefa, etc.

Essa opção pela liberdade na avaliação da clareza e completude dos manuais se deveu ao fato de, segundo alguns especialistas, cada pessoa ter uma visão de mundo diferente das outras e essa visão influenciar o modo como cada uma delas interpreta o (con)texto em que vive. Assim, poderiam ser explicadas as diferentes interpretações possíveis a um mesmo texto por diferentes pessoas, ou até mesmo pela mesma pessoa em diferentes momentos. Portanto, essa liberdade na resposta visa justamente aproveitar essa diferença de interpretações que poderão surgir sobre os manuais e, conseqüentemente, gerar diferentes apontamentos de falhas e sugestões de melhora. Essas, por sua vez, podem se tornar importantes contribuições para a proposta do nosso projeto.

Enfim, foi dentro desse contexto descrito, que foi observada a possibilidade dos três colaboradores concluírem com êxito, ou não, os processos escolhidos para avaliação e quais as dificuldades que essas pessoas sentiram ao realizar esses processos.

O prazo estabelecido para os colaboradores foi de 15 dias tanto para a realização da reprodução das tarefas contidas nos manuais quanto para os comentários sobre os manuais utilizados. Acreditamos que o tempo foi mais do que suficiente, uma vez que as atividades pedidas foram realizadas antes do prazo estipulado.

A seguir, são apresentados os resultados dessa primeira fase de avaliação.

5.2 Resultados da Fase 1 de Avaliação

Retomando a motivação principal dessa fase que é avaliar a *Clareza e a Completude* dos manuais de construção de recursos lingüísticos contidos nas etapas escolhidas, E1', E2', E3 e E5, e observando os resultados das tarefas realizadas pelos

três colaboradores dessa fase de avaliação, podemos dizer que a mesma foi atingida com sucesso. Esse sucesso pode ser explicado pela boa qualidade do material por eles produzido. Essa qualidade nos possibilita dizer que os conceitos apresentados nos manuais foram lidos e compreendidos a ponto de poderem ser identificados nos cinco resumos por eles analisados, conforme as especificações pedidas pelos manuais em avaliação. Isso tudo, de maneira semelhante, ou seja, as três pessoas realizaram as especificações dos manuais, exatamente como descrito e da mesma maneira.

Para avaliar essa possibilidade de replicação das tarefas contidas nos manuais, foi utilizada a estatística *Kappa*.

Kappa é um método estatístico, que foi utilizado na análise de discurso e de diálogo pela primeira vez em 1995[1], por Isard e Carletta, para avaliar a replicabilidade de um esquema de anotação. Segundo a literatura, a estatística *Kappa* tem sido muito utilizada por diferentes pesquisadores como teste para tarefas de classificação nas quais alguns ou vários anotadores (ou juízes) têm como função atribuir classes a um grupo de itens.

Esses tipos de estatísticas, que visam medir o grau de concordância ou discordância entre os anotadores, como é o caso da estatística *Kappa*, possibilitam: 1) descobrir problemas de etiquetação surgidos durante o processo de classificação de sentenças, bem como 2) servir de teste de qualidade e abrangência do conjunto de etiquetas adotado, 3) do manual de anotação consultado e 4) do córpus em treinamento.

Os itens considerados no cálculo do *Kappa* são: o número de pessoas (juízes) que marcaram o córpus; o número de itens sendo classificados e o número de classes utilizadas, os quais aparecerem representados pela fórmula K= P (A) – P (E)/ 1- P (E). Onde:

P(A) é a proporção de vezes que os anotadores concordaram.

P(E) é a proporção de vezes que os juízes concordam aleatoriamente.

Depois de aplicado esse método estatístico, o valor de K é obtido, e poderá apresentar:

Concordância completa, quando o K=1;

Concordância aleatória, quando k=0 e

Máxima discordância, quando o k= -1.

[1] *AAAI 1995 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, March 27-29 1995, Stanford University, Palo Alto, CA, USA.

Essa estatística pode ser calculada de maneira automática. Para o cálculo com 2 juízes (ou anotadores) há vários programas Web disponíveis[2]. Utilizamos um pacote escrito em Perl[3] chamado *Kappa*[4] para vários juízes. Esse pacote contém 3 arquivos: o kappa2.pm, o kappaDiagnosis.pl e um exemplo de como a tabela com as anotações dos textos deve ser submetida a esse programa, em formato .txt, da maneira como é mostrada abaixo.

	J1	J2	J3	J4
1	MAT	MAT	MAT	MAT
2	MAT	MAT	MAT	MAT
3	MAT	MAT	MAT	MAT
4	MAT	MAT	MAT	MAT
5	MAT	MAT	PRO	MAT

Os números 1,2,3,... correspondem ao número de sentenças analisadas.
J1, J2, J3 e J4 indicam o número e quais os juízes ou anotadores que classificaram, no caso, os textos.
MAT MAT MAT MAT correspondem às etiquetas utilizadas pelos juízes no momento de classificação das sentenças.

Depois de baixados os dois pacotes, de se ter instalado o pacote Perl (cuja instalação é automática) e de se ter colocado os dados da anotação no formato a pouco apresentado e de os ter salvado em arquivo .txt, realiza-se a nomeação desse arquivo, que poderia ser, por exemplo, "arquivo_com_dados.txt". Em seguida, é necessária apenas a execução da seguinte linha de comando no *prompt* do DOS:

C:\kappa> perl kappaDiagnosis.pl arquivo_com_dados.txt > saída.txt.

Como resultado desse comando é obtido um arquivo no diretório *Kappa*, intitulado saída, com formato .txt (isto é, texto sem formatação) com todas as informações da análise realizada. O valor de *k* desse texto vai indicar a taxa de concordância existente entre os anotadores; *P(A)* a proporção de vezes que os anotadores concordaram entre si; *P(E)* a proporção de vezes que é esperado os juízes concordarem aleatoriamente; *N* número total de sentenças analisadas e entre parênteses,

[2] Veja um em: http://faculty.vassar.edu/lowry/kappa.html
[3] Para tal, é necessário instalar um pacote *Perl* no computador, onde se calculará o *Kappa*. Um endereço de *site* sugerido para se baixar tal pacote (diretório compactado), nomeado por *Windows AS Package,* é o http://www.activestate.com/store/download.aspx?prdGUID=81fbce82-6bd5-49bc-a915-08d58c2648ca.
[4] Pode ser feito *download* a partir de: http://coteia.icmc.usp.br/coteia/mostra.php?ident=102.5

as classes utilizadas na classificação. Além disso, é mostrada a quantidade de vezes que esses juízes concordaram comparando-se duas a duas as classes utilizadas.

Segundo Orwin (1994), valores para *k* obtidos em uma dada tarefa menores que 0.40 são ruins, valores entre 0.4-5.9 são regulares, os que ficam entre 0.6-0.74 são bons e aqueles maiores que 0.75 são excelentes.

Nessa primeira etapa de nosso estudo, o valor de *k* obtido para a tarefa de identificação dos componentes da estrutura esquemática (etapa E1' do processo proposto) contidas nos 5 resumos analisados foi K=0.835. Quanto às estratégias retóricas (etapa E5), a tarefa mais difícil de ser realizada entre as pedidas, uma vez que possui o maior conjunto de categorias possíveis, 22, o valor de *k* obtido foi K=0.779. Valores estes que, se enquadrados dentro do espectro de avaliação delimitado por Orwin (1994), nos indicam um excelente resultado em relação ao grau de concordância entre os anotadores e a replicabilidade do manual utilizado. Em outras palavras, indica um excelente resultado quanto à adequação do manual em termos de clareza e completude para a realização da tarefa, aplicação do modelo teórico descrito no manual e entendimento por parte dos anotadores das categorias e das atividades a serem realizadas.

Quanto à avaliação da anotação dos marcadores discursivos (etapa E3), esta não foi realizada por nenhum método estatístico. Foi realizada apenas uma comparação visual de listas de marcadores discursivos, destacados em vermelho, nos resumos analisados por cada juiz. Ao comparar as três listas, vimos que os marcadores discursivos identificados são praticamente os mesmos. Não há marcador discursivo algum identificado de maneira inadequada. Há apenas alguns marcadores que não foram destacados nos cinco resumos analisados. A hipótese para tal é uma possível distração na tarefa de identificação desses elementos. Isso, porque a palavra, por exemplo, *this,* aparece identificada nos dois primeiros resumos.

Os marcadores "*not only…but also"* (Adição), *"empirically, formally"* (Modo), *"mainly"* (Intensidade) entre outros, foram sugeridos pelos anotadores para serem inseridos na lista de marcadores discursivos contida no manual de anotação. Mais um fator que indica que o conceito sobre marcadores discursivos apresentado no manual de anotação foi compreendido e, portanto, pôde ser replicado.

Em relação ao manual de rubrica (etapa E2'), percebemos que todos os textos ganharam valores "Alto" ou "Baixo", conforme a ausência ou presença dos três critérios contidos nos manuais. Quanto ao primeiro critério desse manual, que se refere à

caracterização, organização de desenvolvimento do resumo, o valor de *k* obtido foi K=0.659, o que representa que houve apenas uma discordância na anotação das sentenças (Tabela 5.1). Quanto ao segundo critério que trata do *balanceamento entre os componentes de um resumo* o valor de k obtido foi k=1, ou seja, os anotadores concordaram em todos os momentos quanto à classificação das sentenças analisadas sob esse segundo critério (Tabela 5.2). Em relação ao terceiro e último critério, referente à avaliação da *coerência entre os componentes* do resumo, o valor obtido foi k= 0.444, o qual demonstra que eles discordaram duas vezes ao longo das análises das sentenças (Tabela 5.3).

Pelo fato das rubricas terem sido aplicadas apenas a cinco resumos e, portanto, uma discordância entre os anotadores ser suficiente para fazer o valor de *k* cair consideravelmente. Assim, decidimos trazer, a seguir, três tabelas de comparação das anotações feitas com essas rubricas. Dessa maneira, o bom trabalho realizado pelos colaboradores pode ser observado com uma melhor perspectiva e o fato desse manual também ter atingido seu objetivo de estar claro e completo, ter sido também atingido.

	A	**H**	**L**
1	Alto	Alto	Alto
2	Alto	Alto	Alto
3	Alto	Alto	Baixo
4	Alto	Alto	Alto
5	Baixo	Baixo	Baixo

Tabela 5.1: Tabela de comparação da classificação feita pelos três colaboradores em cinco resumos da área de Ciências da Computação. Esses resumos foram analisados sob o critério de *caracterização, organização e desenvolvimento de um resumo.* Observa-se que há apenas um único momento, apresentado na linha 3, em que uma das três pessoas discordou das outras duas quanto a classificação do resumo de número 3. Vale dizer, que a coluna com números de 1 a 5 referem-se aos resumos analisados, a linha com as letras A, H, L refere-se aos anotadores dos resumos e as colunas com valores "Alto" e "Baixo" são as classificações dadas por esses anotadores aos resumos. Nota-se também nessa tabela que os dois primeiros anotadores concordaram em todos os cinco momentos possíveis quanto à classificação dos resumos analisados, o que geraria um valor de k=1, o melhor valor que pode ser obtido em dada tarefa avaliada pela estatística *Kappa*.

	A	**H**	**L**
1	Alto	Alto	Alto
2	Alto	Alto	Alto
3	Alto	Alto	Alto
4	Baixo	Baixo	Baixo
5	Baixo	Baixo	Baixo

Tabela 5.2: Tabela de comparação da classificação feita pelos três colaboradores em cinco resumos da área de Ciências da Computação. Esses resumos foram analisados sob o critério de avaliação do *balanceamento entre os componentes de um resum*o. Observa-se que as três pessoas concordaram em todos os momentos quanto a classificação dos resumos sob o critério proposto, o que gerou um valor de k=1. Vale dizer, que a coluna com números de 1 a 5 referem-se aos resumos analisados, a linha com as

letras A, H, L refere-se aos anotadores dos resumos e as colunas com valores “Alto” e “Baixo” são as classificações dadas por esses anotadores aos resumos.

	A	H	L
1	Alto	Alto	Baixo
2	Alto	Alto	Alto
3	Baixo	Baixo	Baixo
4	Alto	Baixo	Baixo
5	Baixo	Baixo	Baixo

Tabela 5.3: Tabela de comparação da classificação feita pelos três colaboradores em cinco resumos da área de Ciências da Computação. Esses resumos foram analisados sob o critério de avaliação da coerência entre os componentes de um resumo. Observa-se que há aqui dois momentos (apresentado nas linhas 1 e 4 em tom mais escuro), nos quais um dos anotadores discordou dos outros dois quanto a classificação do resumo analisado. Vale dizer, que a coluna com números de 1 a 5 referem-se aos resumos analisados, a linha com as letras A, H, L refere-se aos anotadores dos resumos e as colunas com valores “Alto” e “Baixo” são as classificações dadas por esses anotadores aos resumos.

Um outro produto dessa primeira fase de avaliação foram os apontamentos de eventuais falhas/inadequações dos manuais utilizados, bem como a apresentação de sugestões de melhoria, segundo o ponto de vista de cada um dos três colaboradores.

Uma mesma sugestão foi feita, algumas vezes, por mais de um dos colaboradores e referem-se, de maneira geral, a:

- Sobre o Manual de Componentes Esquemáticos e de Estratégias Retóricas

Diferenciação entre duas categorias aparentemente semelhantes e que, portanto, causam dúvidas quanto à sua identificação nos textos, por exemplo, as estratégias retóricas 14 “Comentar/Discutir os resultados” e 15 “Descrever os resultados” e as de número 12 “Citar/descrever materiais e métodos” e 13 “Listar critérios ou condições”.

- Detalhamento dos exemplos trazidos nos quadros, uma vez que não trazem exemplos de sentenças com categorias que serão utilizadas pelos anotadores. São categorias para se aplicar em seções Metodologia e o manual é sobre categorização/anotação de resumos.

- Alteração da ordem de duas seções presentes no manual de anotação da estrutura retórica. A seção VI.1 viria antes da III. Outro colaborador sugeriu que a seção VI viesse antes da V.

- Elaboração de um quadro resumo com os principais pontos a serem considerados no processo de anotação dos textos.

- Na estratégia retórica “Citar a ausência ou falta de pesquisa anterior” adicionar também “ausência de estudos na disciplina”.

- Na estratégia retórica “Apresentar mais propósitos” adicionar também “ou mais detalhes sobre o propósito principal”.

Sobre o Manual de Marcadores Discursivos

- Dar mais detalhes no manual de como esses marcadores discursivos devem ser anotados. A sugestão é de que sejam utilizadas etiquetas, como as utilizadas na identificação de componentes da estrutura esquemática e de estratégias retóricas. Nessa etapa de avaliação o pedido para que os marcadores discursivos fossem identificados com a cor vermelha foi feito apenas no e-mail enviado e não no manual.
- Apresentação de exemplos de sentenças com marcadores discursivos destacados, com comentário do tipo de função que o mesmo está exercendo na frase.
- Organização alfabética das entradas da lista de marcadores discursivos facilitar a busca por termos.
- Citar exemplos de como um marcador discursivo pode dar dicas quanto à função retórica da frase em que ele se encontra inserido.
- O terceiro parágrafo desse mesmo manual que se refere à tarefa de como proceder para utilizar a lista de marcadores discursivos do manual e o procedimento a ser adotado quando for encontrado um marcador discursivo que não consta na mesma está longo e confuso: *“A partir do marcador discursivo identificado no artigo científico, é utilizado o recurso de busca do Microsoft Word (Word > Editar > Localizar) para encontrá-lo na lista abaixo. Caso o marcador discursivo não seja encontrado na lista abaixo, as funções retóricas que organizam a lista de marcadores abaixo (Contraste/Oposição, Comparação, Adição, etc.) associada ao contexto de uso desse marcador, ou seja, a sentença no qual ele aparece podem auxiliar na classificação do mesmo e em sua posterior inserção seguinte lista”*.

Sobre o Manual de Rubricas

- Um parágrafo explicando o que é e para que serve a rubrica: “Com o intuito de avaliar a qualidade... foram propostos alguns critérios agrupados nas rubricas apresentadas”... uma vez que tal explicação só foi feita no e-mail enviado aos possíveis colaboradores.
- (Rubrica 1) o valor baixo é atribuído quando as condições descritas (todas ou pelo menos uma) não são satisfeitas.
- Se um resumo não tem conclusão, mas está ok, ele não deve ser considerado um bom resumo?
- Quando se diz que o propósito deve ser apresentado em apenas uma linha, esse é o principal ou todos os propósitos?

- (Rubrica 3) E se não houver Lacuna?
- (Rubrica 3) É atribuído valor Baixo se duas condições não forem satisfeitas?
- Se não houver conclusão, o que é bastante comum, o *abstract* é avaliado como Baixo em todas as Rubricas?

Depois de se ter constatado a clareza e a completude dos manuais para a reprodução do processo proposto, o próximo passo foi avaliar a consistência da anotação de um material produzido, conforme as diretrizes trazidas pelos manuais. Além disso, avaliamos o tempo gasto, em média, pelas pessoas na realização das tarefas propostas, e, a partir do resultado obtido, tentamos estipular um tempo a ser gasto para a confecção de recursos lingüísticos.

5.3 Fase 2 de Avaliação – Consistência na anotação dos recursos lingüísticos produzidos e estimativa do tempo gasto na confecção desses recursos

Relembrando, duas são as motivações para a realização dessa segunda fase de avaliação: *Avaliar a consistência na anotação dos recursos lingüísticos produzidos,* segundo as diretrizes contidas nas etapas E1' (identificação dos componentes da estrutura esquemática), E2' (avaliação da adequação textual), e E5 (identificação das estratégias retóricas). Para tanto, utilizaremos a estatística *Kappa.* E e*laborar uma estimativa do tempo gasto na construção dos recursos lingüísticos produzidos nessa segunda fase*.

Vale dizer, que nas duas motivações dessa segunda fase de avaliação, o conhecimento científico e de língua inglesa possuído por nossos colaboradores será considerado, pois desejamos saber se esses conhecimentos podem influenciar tanto na qualidade, como no tempo gasto na produção dos recursos lingüísticos desejados.

Nesse contexto, convidamos seis colaboradores com perfis de conhecimento científico e de inglês o mais semelhante possível ao tipo de conhecimento do público-alvo de nosso processo, os quais são apresentados na Tabela 5.4.

	Graduado	**Mestrado**	**Doutorado**	**Pesquisador Sênior**
Inglês Intermediário	1 pessoa	2 pessoas	1pessoa	-
Inglês Avançado			1pessoa	1 pessoa

Tabela 5.4: Perfil dos colaboradores da segunda fase de avaliação.

Em geral, pode-se classificar o conhecimento de inglês que uma pessoa tem em três níveis: Básico, Intermediário e Avançado. Em nosso estudo, entretanto, optamos por convidar apenas colaboradores com níveis intermediários e avançados de língua inglesa. Isso, porque para se construir os recursos lingüísticos em inglês, é necessário um conhecimento, no mínimo, intermediário desse idioma, para ocorrer um entendimento adequado do conteúdo trazido pelos artigos científicos em inglês que serão anotados, posteriormente. Assim, para nós, um aluno intermediário é capaz de ler um artigo científico em inglês para depreender a idéia geral ou informação específica que necessite encontrar no texto, além de produzir textos com inadequações em nível gramatical e de vocabulário específico da área, necessitando assim, de auxílio especializado de seu orientador para produzir um artigo científico adequado. Em contrapartida, as pessoas de nível avançado são capazes de compreender toda a informação contida em um artigo científico de maneira rápida, bem como de não precisar de auxílio de outra pessoa para produzir um artigo científico com poucas inadequações do ponto de vista da "boa" escrita científica, isto é, que corresponda com as expectativas da comunidade acadêmica. Uma questão se coloca, entretanto: por que um pesquisador acadêmico com alto grau de conhecimento de língua inglesa poderia se interessar por nosso trabalho? Três são os motivos que podemos destacar a princípio, entre todos os possíveis:

1) o pesquisador pode ser um professor que queira que seus orientandos melhorem sua habilidade de escrita científica e, portanto, pode usar nosso processo para gerar um tipo de "ambiente didático" para esses alunos.

2) o(a) pesquisador(a) pode também utilizar o processo para gerar uma base com exemplos de diferentes formas de se dizer uma mesma idéia, por exemplo, o que incrementaria o conhecimento de língua por ele(a) já possuído quanto as idiossincrasias lingüísticas (vocabulário, expressões-padrão, colocados, etc.) da comunidade acadêmica da qual faz parte e portanto, produzir artigos científicos mais elaborados e diversificados lingüisticamente.

3) e por último, o pesquisador pode ser um professor que trabalha com ensino-aprendizagem de escrita científica e pode, portanto, gerar um ambiente de auxílio à escrita científica em inglês personalizado para seus alunos, utilizando-o dentro e fora da sala de aula. Esses alunos podem, inclusive, contribuir na construção de tal ferramenta, com a construção de recursos lingüísticos para serem nela incrementados.

Além desse conhecimento de língua, nossos colaboradores possuem os seguintes níveis de conhecimento científico:

Lingüística: 1 formada e 1 recém-doutor; Ciências da Computação: 1 do segundo ano de mestrado e 1 do terceiro ano de doutorado; Engenharia de Produção: 1 do primeiro ano de mestrado e 1 pesquisador sênior. Essas duas pessoas trabalham também com ensino de escrita científica.

Para cada colaborador dessas duplas foi entregue um grupo de 15 resumos em inglês, específicos da grande área em que atuam. Assim, a primeira dupla de lingüistas recebeu 15 resumos (87 sentenças ao todo) do periódico *International Journal of Corpus Linguistics*, acessível pelo endereço http://www.benjamins.com/cgi-bin/t_seriesview.cgi?series=IJCL. A segunda dupla formada por cientistas da computação recebeu 15 resumos (96 sentenças ao todo) do periódico *Science Direct*, acessível pelo endereço http://www.sciencedirect.com. E a última dupla recebeu 15 resumos (135 sentenças ao todo) da área de Engenharia de Produção do periódico *Emerald*, acessível pelo endereço http://puck.emeraldinsight.com. Esses três periódicos on-line de onde os resumos foram retirados possuem classificação "A" na Capes.

Vale dizer, que também nessa fase de avaliação, os anotadores já receberam seus resumos prontos para a realização da tarefa.

Além desses resumos, esses colaboradores receberam dois manuais (um para a anotação dos componentes da estrutura esquemática e estratégias retóricas e outro para a avaliação da qualidade dos resumos), utilizados na etapa anterior de avaliação. Foi estipulado um prazo de 20 dias para completar a tarefa de identificação dos recursos lingüísticos especificados pelos manuais.

Não foi ao acaso que escolhemos duplas de colaboradores da mesma área. O motivo para tal condição é a possibilidade de se aplicar a estatística *Kappa* para avaliar a consistência com que os colaboradores realizaram a anotação dos resumos específicos de suas áreas de diferentes áreas. E a condição de se usar essa estatística é de se ter, pelo menos, dois colaboradores realizando a mesma tarefa.

5.4 Resultados da Fase 2 de Avaliação

Conforme mencionado anteriormente, o método para avaliar a consistência da tarefa foi a estatística *Kappa*. Os resultados obtidos e algumas observações serão mostrados nas tabelas que se seguem.

	Computação	*Lingüística*	*Engenharia de Produção*
Estruturas Esquemáticas	K=0.899	K=0.829	K=0.799
Número de Sentenças	96	87	135

Tabela 5.5: Identificação dos componentes da estrutura esquemática

Conforme pode ser observado na Tabela 5.5, o valor de concordância entre as duplas pode ser interpretado como excelente, segundo a escala de Orwing (1994). O que implica dizer que as diretrizes contidas no manual referente à anotação dos componentes da estrutura esquemática contêm uma boa explicação tanto da tarefa a ser realizada quanto do modelo de componentes da estrutura esquemática adotado, inspirado no trabalho de Swales (1990). O maior valor de *k* obtido foi da dupla de cientistas da computação, conforme destacado. Esse valor indica o quanto eles concordaram entre si na classificação de cada sentença dos resumos.

	Computação	*Lingüística*	*Engenharia de Produção*
Estratégias Retóricas	K=0.769	K=0.798	K=0.722
Número de Sentenças	96	87	135

Tabela 5.6: Identificação das estratégias retóricas

Já na tarefa de identificação das estratégias retóricas (Tabela 5.6), que conforme já mencionado pode ser considerada a tarefa mais difícil do processo proposto por essa dissertação por possuir o maior conjunto de categorias (22), o resultado obtido é excelente para as duplas de cientistas da computação e de lingüistas e boa para a dupla de engenheiros. Isso porque, para Orwing (1994), o valor excelente é atribuído para valores de *k* maiores que 0.75. Esses ótimos valores de k obtidos também podem ser justificados pelo conteúdo do manual. Acreditamos também que se as duplas tivessem tido um tempo para se "acostumarem" com os manuais e também para praticarem com outros resumos antes de realizar a anotação desses 15 resumos recebidos, esse valor de *k*

poderia ter sido mais alto. O valor destacado na tabela é o mais alto valor obtido nessa tarefa entre as três duplas avaliadas.

	Computação	*Lingüística*	*Engenharia de Produção*
Dimensão 1	K=0.856	K=1.000	K=1.000
Número de Sentenças	96	87	135

Tabela 5.7: Avaliação da qualidade dos resumos segundo o critério de caracterização, organização e desenvolvimento de um resumo, Rubrica 1.

Nessa tarefa tivemos a boa surpresa de duas duplas, a de lingüistas e de engenheiros conforme destaque da Tabela 5.7. Eles atingiram o grau máximo de concordância entre si na realização da tarefa. Não menos importante é o resultado de avaliação da dupla de cientistas da computação, que também tiveram um excelente resultado na avaliação.

	Computação	*Lingüística*	*Engenharia de Produção*
Dimensão 2	K=0.813	K=0.722	K=0.779
Número de Sentenças	96	87	135

Tabela 5.8: Avaliação da qualidade dos resumos segundo o critério de balanceamento entre os componentes de um resumo, Rubrica 2.

Nessa tarefa, os maiores destaques foram as duplas de engenheiros e cientistas da computação que obtiveram um valor excelente na realização de suas tarefas. A dupla de lingüistas obteve um valor considerado bom. O quadro em destaque da Tabela 5.8 é o da dupla que obteve o maior grau de concordância entre si nessa tarefa.

	Computação	*Lingüística*	*Engenharia de Produção*
Dimensão 3	K=1.000	K=0.732	K=0.785
Número de Sentenças	96	87	135

Tabela 5.9: Avaliação da qualidade dos resumos segundo o critério de coerência entre os componentes de um resumo, Rubrica 3.

Nessa última tarefa também houve uma surpresa agradável com a presença do valor k=1, que significa haver uma total concordância entre os cientistas da computação

quanto à avaliação da coerência entre os componentes dos resumos por eles analisados (Tabela 5.9).

Em suma, essas boas avaliações obtidas nas tarefas há pouco avaliadas, mostram que um dos objetivos do processo aqui apresentado – que pessoas da comunidade acadêmica consigam construir os recursos lingüísticos necessários na geração de suas próprias ferramentas de auxílio à escrita científica – foi atingido em três áreas do conhecimento: Ciências da Computação, Lingüística e Engenharia de Produção. Isso, porque a descrição das tarefas realizadas, bem como dos conceitos e termos lingüísticos apresentados estão descritos em uma linguagem de fácil acesso para essa comunidade. Portanto, de fácil entendimento pelas pessoas representantes das três comunidades acadêmicas citadas, conforme mostraram os excelentes valores obtidos nas avaliações das tarefas realizadas. Como trabalho futuro, indicamos a avaliação com pessoas de outras áreas do conhecimento para ver o seu desempenho quanto à realização das tarefas avaliadas.

Em relação à segunda motivação dessa fase de avaliação, elaborar uma estimativa do tempo a ser gasto na construção de recursos lingüísticos, a Tabela 5.10 mostra quanto tempo cada um de nossos colaboradores gastou na realização das tarefas atribuídas a eles nessa segunda fase:

Perfil da pessoa	**Nível de Inglês**	**Tempo gasto nas tarefas**
1. Formada	Inglês Intermediário	06h00min
2. Recém-doutora	*Inglês Avançado*	08h00min
3. Segundo ano mestrado	Inglês Intermediário	05h30min
4. Terceiro ano doutorado	Inglês Intermediário	04h15min
5. Primeiro ano mestrado	Inglês Intermediário	06h30min
6. Pesquisador sênior	*Inglês Avançado*	01h45min a 02h00min

Tabela 5.10: Apresentação do tempo gasto pelos colaboradores na execução das tarefas da fase 2 de avaliação e o nível de conhecimento científico de cada uma delas.

Em destaque, estão o maior e o menor tempo gasto na execução das tarefas: 02h00min e 08h00min. Vale também dizer que esse tempo foi estipulado pelos próprios colaboradores. Foram eles quem nos deram essa informação, não os acompanhamos de perto na execução das tarefas a ponto de ser possível cronometrar o tempo que eles

gastaram. Com relação ao colaborador 6, a princípio pode-se dizer que esse pouco tempo gasto na execução das tarefas pode ser explicado, por exemplo, pelo alto nível de conhecimento que possui, tanto de língua estrangeira quanto científico, uma vez que se trata de um professor que além de ter publicado diferentes artigos científicos em revistas internacionais, também trabalha com disciplinas sobre escrita científica em língua inglesa na pós-graduação. Assim, por possuir um inglês fluente aliado a um bom entendimento das características do discurso científico, fez com que as tarefas fossem por ele realizadas em um tempo baixíssimo se comparado com o tempo gasto em média pelas outras pessoas. Poderíamos deduzir, então, que quanto mais consciência das idiossincrasias da escrita científica de uma comunidade acadêmica e de inglês uma pessoa tiver, menor poderá ser o tempo gasto na execução das tarefas, isto é, na produção de recursos lingüísticos.

Ao fazermos uma média do tempo gasto por essas pessoas na anotação dos componentes esquemáticos, das estratégias retóricas e da avaliação da qualidade de 15 resumos de suas respectivas áreas (Lingüística, Engenharia de Produção e Computação) obtivemos 05h29min, isto é, para a confecção de recursos lingüísticos da seção "Resumo", o usuário de nosso processo levou, em média, 05h29min.

No entanto, não podemos considerar esse tempo médio estimado para a construção de todas as outras seções de uma futura ferramenta de auxílio à escrita, isto é, uma base de casos com 15 exemplos de Introdução, mais 15 de Metodologia, e assim por diante, perfazendo um total de 90 seções para serem analisadas.

Isso devido ao fato de ser preciso fazer mais avaliações de tempo gasto para a anotação dos recursos lingüísticos requeridos, e de também:

- Levar em consideração, que a complexidade de análise das seções de artigo científicos é diferente: resumos são mais fáceis de anotar que metodologias, pois são mais estruturados, isto é, as estratégias e componentes esquemáticos aparecem bem mais definidos que em outras seções.
- O tempo gasto com a análise pode variar muito de pessoa para pessoa, dependendo do nível de inglês e da consciência da estrutura de um texto científico, que a pessoa anotadora dos textos possui.
- O tempo gasto para a anotação dos recursos lingüísticos pode variar: o número de sentenças, além da complexidade entre as seções podem influenciar no tempo gasto. Em 22 seções "Conclusões" do SciPo-Farmácia, por exemplo, foram encontradas 173

sentenças. Ao passo que na seção “Resultados”, um montante semelhante de textos, 26, possui muito mais sentenças: 1429.

Uma sugestão feita é que sejam escolhidos artigos científicos curtos - *letters*, que possuem um tamanho pequeno (de 4 a 6 páginas) e possuem uma padronização maior sobre o tamanho de texto contido em cada seção que o constitui.

6. Contribuições, Limitações e Trabalhos Futuros

6.1 Considerações Iniciais

O processo de gerar conhecimento novo e de o agregar à longa cadeia, construída por todos os pesquisadores de uma área, pressupõe a escrita de investigações realizadas, em revistas acadêmicas de língua inglesa.

Infelizmente, publicar artigos nesta língua é uma dificuldade comum entre pesquisadores não-nativos do inglês e, geralmente, a recusa de submissões se dá muito mais por problemas de escrita do que por problemas relacionados ao conteúdo científico em si.

Uma maneira de melhorar essa situação seria possibilitar a esses pesquisadores um acesso indexado da informação contida em bons artigos científicos de uma dada área. Esse acesso poderia possibilitar um contato com os componentes da estrutura esquemática das seções, por exemplo. E podendo, inclusive, auxiliar esses pesquisadores na produção de um primeiro rascunho, cujo conteúdo apresentasse adequadamente o trabalho relatado. Isso é possível com o uso de ferramentas de auxílio à escrita científica dependentes de domínio, isto é, que trabalham com uma base de dados formada por artigos científicos autênticos de uma dada área. As ferramentas de suporte apresentadas no Capítulo 2 desta pesquisa apresentam essa característica, que possibilita que as idiossincrasias lingüísticas da comunidade científicas, contidas nos artigos científicos da base dessas ferramentas, possam ser facilmente recuperadas por seus usuários. Além, de possibilitar um contato do pesquisador com um material lingüístico adequado às suas necessidades, ou seja, com estruturas e vocabulários pertinentes à área na qual ele precisa escrever.

Nesse contexto, surgiu o objetivo deste projeto de pesquisa: *formalizar um processo para a construção de recursos lingüísticos aplicáveis em ferramentas Web de suporte à escrita científica em inglês*. Essa formalização culminou em uma seqüência de 11 etapas (passos), que dita a ordem e quais atividades devem ser realizadas para se obter uma ferramenta Web de suporte à escrita científica.

As partes automáticas desse processo foram desenvolvidas por um aluno de mestrado da área de Ciências da Computação do NILC-ICMC-USP, Luiz Carlos

Genovês Jr., sob a mesma orientação. A facilidade de tornar genérica a ferramenta SciPo-Farmácia só foi possível devido ao excelente trabalho de programação de um outro aluno de mestrado, também do NILC-ICMC-USP, Lucas Antiqueira. Um terceiro trabalho, ainda futuro, poderá automatizar as tarefas (etapas) descritas no processo aqui proposto, criando-se, então, um *Ambiente Web Gerador de Ferramentas Computacionais de Suporte à Escrita Científica em inglês*.

Os produtos obtidos na elaboração deste processo são divididos em Contribuições, Limitações e Sugestões de trabalhos futuros, e apresentados a seguir.

6.2 Contribuições

6.2.1. Contribuições para a Lingüística de Córpus

- Divulgação da estatística Kappa, sistema estatístico comumente utilizado por cientistas da computação na avaliação da qualidade de uma dada tarefa, na comunidade de Lingüística de Corpus, enquanto avaliador da qualidade de anotação de um córpus. Conforme pudemos observar na literatura revisada (Myrahayuni, 2002; Motta-Roth, 1998; 1995; Oliveira, 2003; Biasi-Rodrigues & Jucá, 2004; Yang & Allison, 2003; Silva, 1999) esses trabalhos comentam a pertinência ou não de modelos teóricos sobre componentes esquemáticos de artigos científicos, com o auxílio de um córpus anotado com tal modelo. Esta pesquisa, entretanto, avalia a qualidade da anotação realizada com o modelo escolhido, no caso, componentes esquemáticos e estratégias retóricas de artigos científicos em inglês, antes de tecer comentários sobre a adequação ou não do modelo retórico estudado. Prática esta que aumenta a possibilidade de se obter resultados mais confiáveis, com embasamento em dados empiricamente avaliados e que, portanto, podem assegurar melhor uma generalização.
- Proposta de um processo para gerar córpus de textos científicos com seus componentes retóricos (componentes da estrutura esquemática e as estratégias retóricas) anotados que poderão ser utilizados em pesquisas de diferentes naturezas de investigação lingüística.
- O Córpus Met, produto gerado durante esta pesquisa (com seções Metodologia de artigos científicos em inglês para que dele fossem retirados recursos lingüísticos necessários à implementação da seção "Metodologia" do SciPo-Farmácia), também pode ser tornar objeto de pesquisa de futuras investigações. Por exemplo: 1) fonte de dados para estudos

terminológicos na área de especialidade que ele representa; 2) poderá também ser utilizado por professores da área de tradução para o ensino de técnicas e procedimentos de tradução da escrita científica em inglês na área de Farmácia; 3) fonte de material autêntico para alunos autodidatas/pesquisadores que queiram investigar o funcionamento da linguagem de especialidade contida nesse córpus; entre outros

6.2.2. Contribuições para o ESP (English for Specifics Purposes)

- Possibilitar que um material gerado a partir de linguagem em uso (córpus) e direcionado a propósitos específicos (no caso, produção textual de artigos científicos para a comunidade acadêmica internacional) possa ser utilizado em sala de aula. Contribuindo assim, enquanto recurso/material a ser utilizado no ensino-aprendizagem de inglês com propósitos específicos.
- Divulgação das teorias estudadas (por exemplo, a dos componentes esquemáticos, das estratégias retóricas e dos marcadores discursivos) e elaboração de um modo (processo) que possibilite uma aplicação das mesmas por pesquisadores de diferentes áreas, além do ESP e Lingüística. Tal situação fez com que temas complexos que há anos vêm sendo discutidos por lingüistas (por exemplo, modelo de componentes esquemáticos de seções de artigos científicos) e, portanto, ficando restritos aos especialistas nas áreas desses temas pudessem fazer parte da prática de pesquisadores de diferentes áreas, auxiliando-os no ensino-aprendizagem de um tipo textual como o artigo científico, tão importante e tão produzido pela comunidade acadêmica em geral.
- Auxílio na divulgação do potencial de córpus eletrônicos para a descoberta de informações lingüísticas até então não pensadas ou não abordadas corretamente. Tais córpus, quando utilizados com prudência e sabedoria, podem se tornar interessantes instrumentos para o ensino-aprendizagem de línguas, despertando o interesse de aprendizes pela investigação, busca de conhecimento sobre a língua em estudo.

6.2.3. Contribuições para o PLN (Processamento de Língua Natural)

- Para que a elaboração do processo pudesse ser concretizada, foi necessária a parceria entre lingüista e cientistas da computação. Portanto, os bons resultados deste trabalho podem servir como incentivo para a criação de mais pesquisas interdisciplinares e que promovam

este tipo de parceria. O caráter interdisciplinar do trabalho realizado proporcionou pontos de vista sobre um mesmo objeto (a língua) que se complementaram, enriquecendo assim as experiências dos envolvidos.

6.2.3. Outras Contribuições

- Disponibilização via Web do processo proposto em uma Web colaborativa como a do Projeto Plonetarium para tornar público e mais facilmente acessível os resultados deste trabalho:http://gen-writingtool.incubadora.fapesp.br/portal. Essa flexibilidade de acesso via Web de qualquer lugar e a qualquer momento, contribui para a promoção de sua divulgação para a comunidade acadêmica em geral, um dos objetivos pontuais desta pesquisa.
- Implementação da seção "Metodologia" do SciPo-Farmácia, por meio da extração de recursos lingüísticos do córpus Met.
- Possibilitar que pesquisadores de diferentes áreas confeccionem recursos lingüísticos para ferramentas de auxílio à escrita científica personalizadas para a área em que atuam, ao mesmo tempo em que podem adquirir noções sobre organização lingüística e retórica adequadas a esse gênero textual - o artigo científico.

6.3 Limitações

- A dificuldade de se encontrar voluntários para participarem das fases de avaliação do processo proposto, o que pode interferir em uma possível generalização que possa ser feita a partir dos resultados obtidos nas duas fases de avaliação, que envolveram a identificação de componentes esquemáticos, estratégias retóricas e da avaliação da qualidade de resumos em inglês. Apesar dos excelentes resultados obtidos, seria interessante ainda, realizar uma avaliação mais completa, que englobasse todas as etapas do processo proposto e com um número maior de pessoas de diferentes áreas.
- Mais modelos referentes à estruturação retórica de artigos científicos poderiam ter sido investigados e sugeridos no formato de manuais para os usuários do processo, os quais teriam um número maior de opções de modelo para escolherem para anotar os recursos lingüísticos de seu córpus. Por exemplo, se o usuário fosse da área de Humanas, trabalhando com pesquisa teórica, poderia primar por escolher um modelo que tivesse componentes da estrutura esquemática e estratégias retóricas mais

recorrentes em artigos científicos dessa área. No entanto, o modelo por nós escolhido é mais apropriado para pesquisas experimentais (Weissberg, 1999), portanto, esse tipo de adaptação necessária deverá ser feito com base nas características/necessidades apresentadas pelo córpus. Um exemplo de proposta de um modelo para resumos da área de Lingüística que poderia ter sido utilizado é o de Ramos (2003) que analisa um córpus de 75 resumos com um modelo adaptado de Swales (1990) para a Lingüística.

6.4 Sugestões de Trabalhos Futuros

- Os autores citados e investigados sobre marcadores discursivos, componentes esquemáticos, estratégias retóricas e expressões formulaicas podem se tornar ponto de partida para que outros estudos sejam realizados no contexto de nossa lingua materna, por exemplo. Esse ponto de partida é necessário para que possamos contribuir, de alguma forma, para o princípio de esclarecimento de questões que envolvem temas tão complexos.
- Contribuiu para a abertura de um campo de pesquisa que vise avaliar a relevância didático-pedagógica do processo proposto no ensino-aprendizagem de escrita científica em inglês. Assim como a investigação de questões relacionadas ao papel do professor e dos alunos que fazem uso desse tipo de recurso para sua instrução formal.
- Poderá ser avaliado com pesquisadores áreas do conhecimento diferentes das já avaliadas, a fim de se verificar o desempenho dessas pessoas quanto à realização das tarefas necessárias, aproveitando os resultados obtidos em uma melhor adequação da descrição do processo contido na disponibilização via Web.
- Quanto à produção da ferramenta de suporte genérica, o Scientific Writing, algumas implementações futuras podem ser sugeridas. Uma delas seria a implementação de um extrator automático de termos, que produzisse como saída uma lista de candidatos a termos específicos da área a qual o córpus pertence, os quais seriam, em seguida, submetidos automaticamente a um concordanciador, que retornaria tais termos em seu contexto de uso. Além desses termos, esse concordanciador poderia disponibilizar ao usuário, a possibilidade de se verificar, por meio de concordâncias, regências verbais, por exemplo. Seria interessante criar, por exemplo, criar um glossário nesse ambiente com as palavras mais importantes e mais usadas no ambiente acadêmico, além de *links* para dicionários *on-line*. Ou ainda, uma forma que permitisse que o Scientific Writing fosse utilizado enquanto material de ensino à distância de

escrita científica, dada sua natureza Web de disponibilização e acesso. Interessante também citar a possibilidade de surgir investigações com base nos artigos científicos produzidos com o auxílio desse ambiente de suporte à escrita, para a partir de uma análise sobre a estrutura e os erros detectados nesses textos sejam elaborados novos recursos ou modificados os recursos já existentes para que tais inadequações diminuam ou deixem de existir.

7 Referências

ALUISIO, S. M.; OLIVEIRA JÚNIOR, O. N. 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*, v.19, p.141-147, 1996. Disponível em: <http://www.cica.es/sepln96/sepln96.html>. Acesso em: agosto de 2006.

ALUÍSIO, S.M. *Ferramentas de auxílio à escrita de artigos científicos em inglês como língua estrangeira*. 1995. 216 f. Tese (Doutorado em Ciências - Física Aplicada, Subárea Física Computacional) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 1995.

ALUISIO, S.M. et al. 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*, Madison, Wisconsin. Los Alamitos, CA: IEEE Computer Society, 2001, v. 1, p. 257-260.

ALUÍSIO, S.M.; GANTENBEIN, R.E. Towards the application of systemic functional linguistics in writing tools. In: *Proceedings of international conference on computers and their applications*, Arizona, 1997, p. 181-185, 1997a.

ALUÍSIO, S.M.; GANTENBEIN, R.E. Educational tools for writing scientific papers. In: *Simpósio brasileiro de informática na Educação*, 8, 1997, nov. 18-20: São José dos Campos. Anais do VIII SBIE. São José dos Campos: ITA, 1997b, p. 239-253.

ALUÍSIO, S.M.; OLIVEIRA JUNIOR, O.N. A case-based approach for developing writing tools aimed at non-native english users. In: *Lecture Notes in Artificial Intelligence, 1010,*1995, p. 121-132.

ALUÍSIO, S.M *et al.* Evaluating scientific abstracts with a genre-specific rubric. In: *International Conference on Artificial Intelligence in Education - AIED*, 12, 2005, Amsterdã. Anais do XII ICAIE, Amsterdã, 2005, p. 18-22.

ARISTÓTELES. *Arte retórica e arte poética*. Rio de Janeiro: Tecnoprint, 1991.

ATKINS, S.; CLEAR, J.; OSTLER, N. Corpus design criteria. *Journal of Literary and Linguistic Computing*, Oxford, v.7, n.1. p.1-16, 1992.

AUGUSTO-NAVARRO, E. H. *Gênero discursivo e aspectos pragmáticos: i*mplicações para o ensino da correspondência em língua estrangeira (inglês) via correio eletrônico. 2002. Tese (Doutorado em Lingüística) – Departamento Lingüística e Língua Portuguesa, Universidade Estadual Paulista, Araraquara, 2002.

BAKHTIN, M. M. *Estética da criação verbal*. São Paulo: Martins Fontes, 1997.

BALDO, A. Gêneros discursivos ou tipologias textuais? *Revista Virtual de Estudos da Linguagem – ReVEL*, v. 2, n.2, 2004.

BARRASS, R. *Os cientistas precisam escrever:* guia de redação para cientistas, engenheiros e estudantes. São Paulo: Universidade de São Paulo, 1979.

BAZERMAN, C. *Shaping written knoweledge: t*he genre and activity of the experimental article in science. Madison: University of Wisconsin, 1988.

BERBER-SARDINHA, A.P. Lingüística de Corpus: histórico e problemática. *D.E.L.T.A.*, São Paulo, v.2, n.16, p. 323-367, 2000a.

_____. Computador, corpus e concordância no ensino da léxico-gramática da língua estrangeira. In:_____. LEFFA, V. (Org.). *As palavras e sua companhia: o* léxico na aprendizagem das línguas. Pelotas, RS, 2000, p. 45-72, 2000b.

_____. Lingüística de Corpus. São Paulo: Manole, 2004.

_____. A influência do tamanho do corpus de referência na obtenção de palavras chaves usando o programa WordSmith Tools. *The Especialist*, São Paulo, v. 26, n. 2, p. 183-204, 2005.

_____. *Usando o WordSmith Tools na investigação da linguagem*, LAEL, PUC-SP, São Paulo, 1999.

BHATIA, V. *Analysing Genre:* language use in professional settings. London: Longman, 1993.

BIASI-RODRIGUES, B. *Estratégias de condução de informações em resumos de dissertações.* 1998. Tese (Doutorado em Lingüística), v. I e II, Universidade Federal de Santa Catarina, Florianópolis, 1998.

BIASI-RODRIGUES, B; JUCÁ, D. C. N. Análise de mecanismos retóricos em resumos acadêmicos e em seções de introduções. In: CAVALCANTE, M.M.; BRITO, M.A.P. (Org.). *Gêneros textuais e referenciação*. Fortaleza, 2004, v 1, CD-Rom.

BIBER, D. *Dimensions of register variation: a* cross-linguistic comparison. Cambridge: Cambridge University Press, 1995.

_____. Using register-diversified corpora for general language studies. *Computational Linguistics*, Cambridge, MIT Press, v. 19, n.2, p. 219-41, 1993.

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus linguistics:* investigating language structure and use. Cambridge: Cambridge University Press, 1998.

BIBER, D. et al. *Longman Grammar of Spoken and Writen English.* London: Longman, 1999.

BRETT, P. A genre analysis of the results section of sociology articles. *English for Specific Purposes*, v. 13, n. 1, p. 47-59, 1994.

CALDEIRA, S.M.A. *et al.* Writing tools for non-native users of English. In: *Proceedings of the XVIII Latin-American Informatics Conference*, Spain, 1992, p. 224-231.

CARLETTA, J. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, v. 22, n. 2, p. 249-254, 1996.

CASELI, H.M.; FELTRIM, V.D.; NUNES, M.G.V. *TagAlign:* uma ferramenta de pré-processamento de textos. São Carlos: ICMC-USP, 2002. (Relatório técnico, NILC-TR-02-09).

CHURCH, K. W.; MERCER, R. L. Introduction to the special issue on Computational Linguistics using large corpora. *Computational Linguistics,* v. 19, n. 1, p. 1-24, 1993.

CORACINI, M. J. U*m fazer persuasivo:* o discurso subjetivo da ciência. São Paulo: EDUSC, 1991.

De OLIVEIRA, M.C.F. et al. A discussion on human-computer interfaces for writing support tools. In: *Proceedings of the XII International Conference of the Chilean Computer Science Society*, Santiago, Chile, p. 223-233, 1992.

DEYES, T. *Discourse, Science and Scientific Discourse*: the raw material of comprehension in ESP. São Paulo, Pontifícia Universidade Católica de São Paulo, 1982. Brazilian ESP Project. Working Paper 6.

ELLIS, R. S.L.A. *Research and Language Teaching*. New York Oxford University Press, 1997.

FELTRIM, V.D. *Uma abordagem baseada em corpus e em sistemas de critica para a construção de ambientes Web de auxilio à escrita acadêmica em português*. 2004. 181f. Tese (Doutorado em Ciências da Computação), Instituto de Ciências Matemáticas e de Computação, São Carlos, 2004.

FELTRIM, V.D., *et al.* A construção de uma ferramenta de auxílio à escrita de resumos acadêmicos em português. In: *Proceedings of ENIA' 2003*, Campinas: SBC, 2003, p. 2399-2404.

FLOWERDEW, L. An analysis of the problem-solution pattern in an apprentice and professional corpus of technical writing from a systemic-functional perspective. *TESOL Quarterly*, v. 37, n. 3, 2003, p. 489-511.

_____. The argument for using English specialized corpora to understand academic and professional language. In: CONNOR, U.; UPTON, T. (Ed.), *Discourse in the professions:* perspectives from corpus linguistics. Amsterdam: John Benjamins, 2005.

_____. Problems in writing for publication in English: the case of Hong Kong. *Journal of Second Language Writing*, v. 8, 1999, p. 243-248.

FONTANA, N.M., *et al.* Computer assisted writing: applications to English as a Foreign Language. *Computer Assisted Language Learning (CALL)*, v. 6, n. 2, 1993, p. 145-161.

FONTANA N.M. *Summarizing strategies in L1 and L2*. MA Dissertation. University College of North Wales, Bangor. 1989.

FORATTINI, O. P. A língua franca da ciência. *Revista de Saúde Pública,* v.31, n. 1, 1997, p.3-8.

FRASER, B. Discourse markers across language. In: *Pragmatics and Language Learning*, 1993, International Conference on Pragmatics and Language Learning, v.4, 1993.

_____. Pragmatic markers. *Journal of Pragmatics* v. 6, n. 2, p. 167-90, 1996.

_____. Towards a theory of discourse markers. In: FISCHER, K. (Ed.), *Approaches to Discourse Particles*, *Elsevier Press*, 2005.

_____. What are discourse markers? *Journal of Pragmatics*, v. 31, p. 931-952, 1999.

GENOVES JUNIOR, L.C. Avaliação automática da qualidade de escrita para resumos científicos em inglês. 2007. Dissertação (Mestrado em Ciências da Computação) – Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, São Carlos, 2007 (em fase de conclusão).

GENOVES JUNIOR, L.C *et al. A two-tiered approach to detecting English article usage:* an application in scientific paper writing tools, 2007. (submetido à ACL).

GRANJER, S.; TRIBBLE, C. Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In: GRANGER, S. (Org.), *Learner English on computer.* New York: Longman, 1998, p. 199-209.

HALLIDAY, M. A. K., ANGUS, M., STREVENS, P. *The Linguistic sciences and language teaching*. London : Longman, 1965.

HALLIDAY, M.A.K.; MARTIN, J.R. *Writing Science:* literacy and discursive power. London: The Falmer Press, 1993.

HALLIDAY, M.A.K.; MATTHIESSEN, C.M.I.M. *Constructing Experience through Meaning:* a language-based approach to cognition. London: Cassell, 1999.

HINKEL, E. Tense, aspect and the passive voice in L1 and L2 academic texts. *Language Teaching Research*, v.8, p. 5-29, 2004.

HOEY, M. *Patterns of lexis in text*. Oxford: Oxford University Press, 1993.

HUCKIN, T.N.; OLSEN, L.A. *Technical Writing and Professional Communication for Nonnative Speakers of English*. McGraw-Hill, 1991.

IDE, N., BREW, C. Requirements, tools, and architectures for annotated corpora. In: *Proceedings of data architectures and software support for large corpora.* European Language Resources Association, Paris, 2000, p.1-5.

IDE, N; BONHOMME, P; ROMARY, L. *XCES*: an XML-based encoding standard for linguistic corpora. In: *Second International Conference on Language Resources and Evaluation (LREC)*. Athens, 2000. Disponível em: <http://www.cs.vassar.edu/~ide/papers/xces-lrec00.pdf>. Acesso em: janeiro de 2007.

JACOBI-BLASZKOWSKI, C.C. *Lingüística de Corpus e ensino de espanhol a brasileiros: descrição de padrões e preparação das atividades didáticas*. 2000. 122f. Dissertação (Mestrado em Lingüística Aplicada e Estudos da Linguagem). Pontifícia Universidade Católica de São Paulo, São Paulo, 2000.

JAMES, K. *Foreign language learning*. Bangor (UK), Department of Linguistics, University of Wales, 1989. (Lecture Notes)

JAMES, K. The writing of theses by speakers of English as a Foreign Language: a case study. In : *Common ground: shared interest in ESP and communication studies*, WILLIAMS et al (Ed.), ELT Documents 117, Oxford: Pergamon Press, 1984, p. 99-113.

JOHNS, A.M.; DUDLEY-EVANS, T. English for Specific Purposes: international in scope, specific in purpose. *TESOL Quarterly*, v. 25, n. 2, 1991, p. 297-314.

JOHNS, T. *Should you be Persuaded*: two examples of Data-Driven Learning Classroom Concordancing*. English Language Research Journal,* v. 4, 1991, p. 1-16.

JOHNS, T. Whence and whiter classroom concordancing? In: BONGAERTS, T. *et al.* (Ed.), *Computer applications in language learning.*, Dordrecht: Foris, 1988, p. 9-27.

JORDAN, R. R. *English for academic purposes:* a guide and resource book for teachers. New York: Cambridge University Press, 1997.

KANOKSILAPATHAM, B. Rhetorical structure of biochemistry research articles. *English for Specifics Purposes*, v.24, 2005, p. 269-292.

KAUFFMAN, C.H. *O corpus do jornal:* variação lingüística, gêneros e dimensões da imprensa diária escrita. 2005. Dissertação (Mestrado em Lingüística Aplicada e Estudos da Linguagem), LAEL, Pontifícia Universidade Católica de São Paulo – PUC-SP, São Paulo, 2005.

KENNEDY, G.D. *An introduction to Corpus Linguistics*. London: Longman, 1998.

KUKICH, K. Beyond automated essay scoring. *IEEE Intelligent Systems*, v. 15, n. 5, 2000, p. 22-27.

LA PORTE, R. *Scientific publication:* evolution to the Internet, 1998.

MARCUSCHI, L. A. *Gêneros textuais:* o que são e como se constituem. Recife: UFPE, 2000.

MARCUSCHI, L.A. Gêneros textuais: definição e funcionalidade. In: DIONÍSIO, A.P.; MACHADO, A.R.; BEZERRA, M.A. (org.). *Gêneros textuais & ensino*. Rio de Janeiro: Editora Lucerna, p. 19-36, 2002.

MARQUIAFÁVEL, V.S.; GENOVÊS JUNIOR, L.C.; ALUISIO, S.M. Um processo semi-automático para a geração de ferramentas de suporte à escrita científica em inglês. *Fourth Workshop em tecnologia da Informação e da Linguagem Humana*, *TIL' 2006,* 27 e 28 de Outubro, Ribeirão Preto, 2006.

McENERY, A.M.; WILSON, A. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

MIRAHAYUNI, N. K., *Investigating textual structure in native and non-native english research articles*: strategy differences between english and indonesian writers. 2002. 345 f. Tese (Doctor of Philosophy) – Department of Linguistics, School of Modern Language Studies, University of New South Wales, Australia, 2002.

MOTTA-ROTH, D. Escritura, gêneros acadêmicos e construção do conhecimento. In: *Letras,*v.17, UFSM, Santa Maria: Palloti, 1998, p. 93-110.

_____. *Gêneros discursivos acadêmicos, construção de conhecimento e pluralidade de acesso:* a publicação acadêmica impressa e eletrônica e sua relação com os processos discursivos na construção do conhecimento científico. Santa Maria, UFSM, 2000. (Relatório do Projeto Integrado – Bolsa de produtividade em pesquisa, CNPq nº 350389/98-5)

_____. *Rhetorical and disciplinary cultures: a genre based study of academic book reviews in linguistics, chemistry and economics*. Florianópolis, 1995, 311 f. Tese (Doutorado em Letras), Programa de Pós-Graduação em Inglês, Santa Catarina, Universidade Federal de Santa Catarina, 1995.

_____. A importância do conceito de gêneros discursivos no ensino de redação acadêmica**.** *Intercâmbio*, PUC-SP, v.8, p.119-128, 1999.

NARCHI, NZ; SECAF, V. Códigos de ética profissional e a pesquisa: direitos autorais e do ser humano. *Revista Paulista de Enfermagem*, v. 21, n. 3, 2002, p. 227-33.

NARITA, M. 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*, p. 1187-1191, 2000a.

_____. Corpus-based English Language Assistant to Japanese Software Engineers. In: *Proceedings of Machine Translation and Multilingual Applications in the New Millennium, MTMA' 2000*, 2000b.

NWOGU, K. Discourse variation in medical texts: schema, theme and cohesion on professional and journalistic accounts. In: *Systematic Linguistics*, v.2, University of Nottingham, England, 1990. (Monographs)

OLIVEIRA Jr., O.N.; CALDEIRA, S.M.A.; FONTANA, N. Chusaurus: a writing tool resource for non-native users of english. In: BAEZA-YATES, R.; MANBER, U. (ed.) *Computer Science:* Research and Application, New York: Plenum Press, 1992, p. 63-72.

OLIVEIRA, F.M. *A configuração textual da seção de metodologia em artigos acadêmicos de Lingüística Aplicada*. 2003. 136f. Dissertação (Mestrado em Letras), Curso de Pós-graduação em Letras, Universidade Federal de Santa Maria, Santa Maria, Rio Grande do Sul, 2003.

OLIVEIRA, S. L. *Tratado de metodologia científica*. São Paulo: Pioneira, 2001.

ORWIN, R.G. Evaluating coding decisions. In: COOPER, H.; HEDGES, (Ed.), The *handbook of research synthesis*. University of Birmingham: ELR Journal, Birmingham , v. 1, 1994. p. 79-116.

OZTURK, I. The textual organization of research article introductions in applied linguistics: variability within a single discipline, *English for Specific Purposes*, v. 25, 2006. (Article in Press)

PAIZAN, D.C. *O uso da linguagem da Internet na produção de um módulo de leitura de inglês instrumental*. 2001. 177f. Dissertação (Mestrado em Letras, Lingüística e Língua Portuguesa. UNESP-Araraquara, fevereiro de 2001.

PARDO, M.R. *Critérios de Construção e Organização de um Corpus de Especialidade:* o Corpus Técnico-Científico de Ortodontia. 2004. 156f. Dissertação (Mestrado em Letras), Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2004.

PARDO, T.A.S. *SENTER: u*m segmentador sentencial automático para o português do Brasil. São Carlos: ICMC-USP, 2006, 6p. (Relatório Técnico).

POSSAMAI, V. *Marcadores textuais do artigo científico em comparação português e inglês:* um estudo sob a perspectiva da tradução. 2004. 165f. Dissertação (Mestrado em Teorias do Texto e do Discurso) – Departamento, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

QUIRK, R. et al. *A comprehensive grammar of the English language*. Londres: Longaman, 1985.

QUIRK, R. On corpus principles and design. In: SVARTVIK (ed.), 1992, p. 457-469.

RAMOS, W.C., E*quacionamento das fases lingüística e representacional de um programa computacional de auxílio à escrita de abstracts em inglês*. 2004. Dissertação (Mestrado em Lingüística e Língua Portuguesa) - Universidade Estadual Paulista Júlio de Mesquita Filho, 2004.

RENOUF, A. *Explorations in Corpus Linguistics*. Rodopi, 1984.

SAMRAJ, B. An exploration of genre set: research article abstracts and introductions in two disciplines. *English for Specific Purposes*, v.24, p. 141-156, 2005.

SANTOS, V.B.M.P. Estabelecendo as diferenças entre os termos registro e gênero. *English for Specific Purposes*, v. 19, n. 1, p. 1-40, 1996.

SCHUSTER, E. *et al.* Enhancing the writing of scientific abstracts: a two-phased process using software tools and human evaluation. *Anais do ENIA*, 2005, p. 962-971.

SCOTT, M. *WordSmith Tools Version 3*. Oxford: Oxford University Press, 1998.

SECAF, V. *Artigo científico***:** do desafio à conquista. São Paulo: Reis Editorial, 2ª ed., 2001.

SEVERINO, A. J. *Metodologia do trabalho científico*. São Paulo: Editora Cortez, 1996.

SHARPLES, M.; PEMBERTON, L. Representing writing: external representations and the writing process. In: HOLT, P.O.; WILLIAMS, N. (ed.). Computers and writing: state of the art. *Intellect*, Oxford, 1992, p.319-336.

SILVA, M.H.B., PELIZZONI, J.M. & ALUISIO, S.M. Uma abordagem híbrida baseada em críticas e casos para a construção de ferramentas colaborativas de ensino de escrita de artigos científicos. In: *Anais do IX Simpósio Brasileiro de Informática na Educação*, 1998.

SILVA, L.F. *Análise de gênero*: uma investigação da seção de resultados e discussão em artigos científicos em química. 1999. 111 f. Dissertação (Mestrado em Letras), Curso de Pós-graduação em Letras, Universidade Federal de Santa Maria, Rio Grande do Sul, 1999.

SINCLAIR, J. Corpus, concordance, collocation . Oxford: Oxford University Press, 1991.

SOLER, V. Analyzing adjectives in scientific discourse: an exploratory study with implications for spanish speakers at advanced university level. *English for Specific Purposes*, v.21, p. 145-165, 2002.

SWALES, J. *Aspects of article introduction*. Birmingham, UK, The University of Aston, Language Studies Unit, 1981.

_____. Genre and engagement. *Revue Belge de Philologie et d'Histoire* , v. 71, p. 687-698, 1993.

_____. Non-native speaker graduate students and their introductions: global coherence and local management. In: CONNOR, U. & JOHNS, A.M. (ed.), Coherence in Writing: Research and Pedagogical Perspectives, *TESOL Quarterly* , Alexandria, 1990, p. 187-207.

_____. Rethinking genre: another look at discourse community effects. In: *Colóquio: Rethinking Genre*, Carleton University, Ottawa, 1992.

_____. *Genre Analysis:* English in academic and research settings. Cambridge: Cambridge University Press, 1990.

SWALES, J.; FEAK, C.B. *English in Today's Research World:* a writing guide. Michigan: The University of Michigan Press, 2003.

SWALES, J. *et al*. Consider this: the role of imperatives in scholarly writing. *Applied Linguistics*, v.19, p. 97-121, 1998.

SWALES, J; LEE, D. A Corpus-Based EAP course for NNS doctoral students: moving from available specialized corpora to self-compiled corpora, *International Journal of Corpus Linguistics*, v.11, n..2, p. 256-257, 2006.

TAGNIN, S. E. O. *Convencionalidade e Produção de Texto: um dicionário de Colocações Verbais Inglês/Português; Português/Inglês*. Tese de Livre-Docência. Universidade de São Paulo, São Paulo, 1998.

TAGNIN, S. E. O. *Expressões idiomáticas e convencionais*. São Paulo: Ática, 1989.

TAVARES, L.S.L. *Uma análise da estrutura retórica de um gênero em inglês:* a comunicação em VHF. 2004. Dissertação (Mestrado em Letras), Programa de Pós-graduação em Letras, Pontifica Universidade Católica do Rio, Rio de Janeiro, 2004.

TELINE, M.F. *Avaliação de Métodos de Extração Automática de terminologia para textos em Português*. 2004. Dissertação (Mestrado em Ciências da Computação), ICMC-USP, São Carlos, Fevereiro 2004.

TOGNINI-BONELLI, E. From a reliable source: uses and function of the adjective real. *Dialogue Analysis IV, papers from the 4th Conference, Basel 1992*. Tubingen: Max Niemeyer Verlag, pp. 429-436, 1993.

TRIBBLE, C. Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching. In: MELIA, J.; LEWANDOWSKA-TOMASZCZYK, B. (ed.). *Proceedings of Practical Applications in Language Corpora, PALC' 97*, Lodz: Lodz University Press, University of Lodz, Poland, 1997.

THURSTUN, J.; CANDLIN, C.N. Concordancing and the teaching of the vocabulary of academic english, *English for Specific Purposes*, v.17, n.3, pp. 267-280, 1998.

UPTON, T. Understanding direct mail letters as a genre. *International Journal of Corpus Linguistics*, v. 7, n. 1, 2002, p. 65-85.
WEISSBERG, R.; BUKER, S. *Writing up research:* experimental research report writing for students of english. Prentice Hall, 1990.

WOLFSON, N. *Perspectives*: sociolinguistics and TESOL. New York: Newbury House Publishers, 1989.

YANG, R.; ALLISON, D. Research articles in Applied Linguistics: moving from results to conclusions. *English for Specific Purposes*, v. 22, 2003, p. 365-385.

Apêndice 1: Manual de anotação das estruturas esquemáticas e estratégias retóricas da seção Metodologia

As orientações abaixo descrevem o esquema de anotação para a seção Materiais e Métodos do corpus de artigos científicos que compõe o projeto SciPo-Farmácia.

1. Artigo científico: estrutura

Uma característica comum a praticamente todos os textos científicos, que descrevem pesquisa experimental, é o tipo de organização que sua estrutura esquemática deve seguir. Essa estrutura pode ser apresentada como Introdução, Desenvolvimento e Conclusão, sendo que o Desenvolvimento pode ser subdividido em Materiais e Métodos e Resultados, ou ainda Materiais e Métodos, Resultados e Discussão. O objetivo desse tipo de estruturação é guiar o leitor e fazer com que ele siga, na leitura ou escrita do texto, o movimento do fluxo da informação a ser transmitida que parte do geral-para-específico na Introdução e chega ao específico-para-geral, na Conclusão, conforme pode ser observado na figura abaixo.

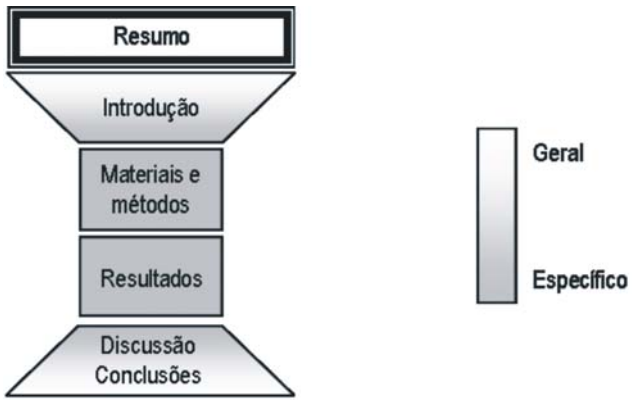


Figura 1: Movimento Geral-Específico-Geral presente na estrutura global do texto científico (Weissberg & Buker, 1990)

2. O que é a seção Materiais e Métodos?

Normalmente, após a Introdução, é apresentada uma segunda seção de texto chamada de Materiais e Métodos (ou apenas Metodologia), principalmente em trabalhos que envolvem pesquisas experimentais.

Esta seção tem como objetivo prover ao leitor uma explicação minuciosa, rigorosa e exata de toda a ação desenvolvida no método (caminho) do trabalho de pesquisa e os materiais utilizados em cada passo. Esta seção serve aos seguintes propósitos do leitor:

1) Entender cada etapa do experimento realizado;
2) Analisar mais criticamente os resultados obtidos;
3) Entender como a metodologia utilizada influenciou os resultados;
4) Reproduzir, se assim desejar, os resultados apresentados.

Para que tais propósitos sejam atingidos, qualquer informação que possa ter tido influência na aquisição dos dados devem ser mencionada no texto, tais como: condições ambientais, marca dos produtos, eventuais alterações nas metodologias geralmente utilizadas etc. Portanto, o objeto de estudo e os equipamentos utilizados também devem ser citados.

3. Categorias para a anotação retórico-manual do texto

As categorias escolhidas para realizar a anotação retórica dos textos foram inspiradas no modelo de estruturação retórica proposto por Swales (1990). A seguir, na figura 2, temos uma lista de siglas que correspondem as possíveis seções que podem ser encontradas na seção metodologia de um artigo

experimental. Para cada uma dessas categorias foi elaborada uma sigla que a representasse, facilitando-se assim, o processo de anotação do corpus.

Siglas das categorias	Descrição das siglas
MAT-LIST	Listagem dos materiais utilizados no estudo
MAT-FONT	Detalhamento da Fonte dos materiais utilizados
MAT-INFO	Fornecimento de informações a respeito dos materiais
PRO-DOC	Detalhamento dos procedimentos utilizados para a execução correta da metodologia
PRO-DET	Detalhamento dos procedimentos utilizados
PRO-JUST	Fornecimento de justificativa sobre os procedimentos
EQU	Equipamentos utilizados
PAD	Procedimentos de análise dos dados
RES	Resultados da Pesquisa

Figura 2: Note, na figura acima, que a sigla da categoria é composta sempre por letras que compõem uma dada categoria, de forma a facilitar a memorização e fácil identificação do significado da categoria que deverá ser empregada nas sentenças do corpus.

3.1. O modelo de Swales (1990)

A análise dos Movimentos, desenvolvida por Swales, visa representar artigos científicos em termos de sua organização textual hierárquica construída por seções distintas. Cada seção pode ser dividida em movimentos e cada movimento pode ser subdividido em passos. De acordo com o modelo de Swales, a seção metodologia possui quatro movimentos.

O movimento consiste em uma unidade textual funcional, utilizada com algum propósito retórico identificável. Os movimentos podem variar em tamanho, mas normalmente possuem, no mínimo, uma proposição (Mauranen, 1993:225).

No caso desse estudo, os movimentos serão identificados ao longo de cada sentença.

Vale lembrar que essas estruturas (categorias) visam a descrever de maneira geral as funções retóricas que podem ser encontradas em um córpus. Mas é bem possível que outras funções possam ser encontradas no córpus em análise e acrescentadas a esse modelo. Aliás, esse modelo é um ponto de partida para que seja aprimorado segundo as características que o córpus em estudo/análise apresentar.

3.2 Detalhamento de cada um dos movimentos (categorias)

Nessa seção, são apresentados os movimentos (ou categorias) e a forma como esses podem aparecer no texto.

Vale salientar que abaixo aparecem trechos de sentenças sublinhados para indicar quais elementos foram importantes para o julgamento/categorização das sentenças segundo os movimentos e passos propostos.

Movimento 1- Descrição de Materiais – Sigla MAT

Esse movimento engloba uma grande variedade de materiais utilizados nas pesquisas**,** abrangendo desde as substâncias naturais, órgãos ou tecidos animais ou humanos, às substâncias químicas (por exemplo, *cell lines*, anticorpos, plasmídeos, enzimas, nucleotídeos, microssomos, membranas, soro, proteínas, genes, *transporons*, DNAs).

O **Movimento 1** pode ser realizado via **Passo 1, Passo 2 ou Passo 3:**

Passo 1: Listagem dos materiais, itemizando os materiais e as substâncias utilizadas no estudo.

Passo 2: Detalhamento da fonte dos materiais, identificando como esses itens são obtidos, tais como, por compra, por doação, etc.

Passo 3: Fornecimento de informações sobre os materiais utilizados, incluindo a descrição, propriedades ou características dos materiais.

A realização do Movimento 1 via Passos 1-3 é ilustrada a seguir:

Movimento 1, Passo1: Listagem dos materiais
MAT-LIST Bacterial strains used in this study and their origin are listed in Table 3.
MAT-LIST/MAT-FONT Materials—CO, ^13C^18O, K^13C^15N, K^13CN, and D2^18O were purchased from Icon (Mt. Marion, NY), and KC^15N was from Cambridge Isotopes.
*Interessante notar que essa sentença possui duas funções retóricas: a de informar o modo como foram obtidos os materiais, daí MAT-FONT, como também itemizar/listar os materiais utilizados no estudo.
Movimento 1, Passo 2: Detalhamento da fonte dos materiais
MAT-FONT COS-7 cells were obtained from S.Brandt (Vanderbilt University, Nashville, Tenn).
MAT-FONT Microsomes derived from samples of human renal cortex were obtained from the Human Cell Culture Center (Laurel, MD), from the International Institute for the Advancement of Medicine (Scranton, PA), and from Dr. Barbara Haehner-Daniels (Indiana University, Indianapolis, IN).
MAT-FONT GSH was from Roche Molecular Biochemicals.
Movimento 1, Passo 3: Fornecimento de informações sobre os materiais
MAT-INFO Antisense riboprobe for RNase protection assay contains the murine mdm2 cDNA fragment spanning from nt+264 to nt +3 (R).
MAT-INFO/MAT-FONT Catalase was purchased from Sigma (Type C-40, specific activity 17,890 Sigma units/mg, assayed at pH 7.0 and 25°C by the rate of decrease in absorbance of a 10.3 mM solution of hydrogen peroxide at 240 nm) and from Roche Molecular Biochemicals (Catalog No. 106 810, specific activity 75,080 Sigma units/mg).
Exemplo interessante que apresenta as três funções retóricas descritas acima:
MAT-FONT/MAT-INFO/MAT-LIST Disodium hydrogen orthophosphate monohydrate, sodium dihydrogen orthophosphate dihydrate, copper(II) sulfate pentahydrate, ferrous sulfate heptahydrate, ferrous ammonium sulfate hexahydrate, ferric ammonium sulfate dodecahydrate, ferric chloride hexahydrate, cobalt(II) chloride hexahydrate, nickel(II) chloride hexahydrate, chromium(III) potassium sulfate dodecahydrate, aluminum ammonium sulfate dodecahydrate, and zinc sulfate heptahydrate were Analar-grade chemicals from BDH.

Movimento 2: Descrição de procedimentos experimentais – Sigla PRO

Esse movimento indica que disciplinas, como, por exemplo, a bioquímica, são disciplinas bem estabelecidas e seus procedimentos, métodos e técnicas são freqüentemente protocolados. Esse segundo movimento pode ser realizado por meio de três passos:
Movimento 2, Passo 1: Documentação de procedimentos estabelecidos, relata um processo experimental que já foi realizado por pesquisadores anteriores. Como resultado dessa padronização dos procedimentos experimentais, a referência simples ao nome específico do método ou procedimento utilizado para conduzir a pesquisa já é suficiente.

Às vezes, certos procedimentos são únicos ou não-ortodoxos para um estudo em particular. Nesses casos, aconselha-se a utilização do Passo 2.
Movimento 2, Passo 2: Detalhamento dos procedimentos, é utilizado para fornecer descrição detalhada dos procedimentos a fim de permitir replicações futuras da pesquisa.

O Movimento 2 também pode ser realizado via Passo 3:
Movimento 2, Passo 3: Fornecimento de informações sobre os procedimentos promovendo assim justificativa para a escolha das técnicas ou procedimentos, comentários e/ou observações realizados durante o experimento.

Exemplos:
Movimento 2, Passo 1: Documentação de procedimentos estabelecidos
PRO-DOC The syd2 mutant was identified by screening in 3rd chromosome EMS lethal lines (bq; st (3)EMS/TM6B, TB) obtained from Charles Zuker (UCSD) as described previously (R).
PRO-DOC/PRO-DET hIDO was expressed and purified as a fusion protein to a hexahistidyl tag as detailed elsewhere (18).
* Além de apresentar detalhes importantes para a correta execução da metodologia do estudo apresentado, essa frase também possui a função de relatar procedimentos que já foram realizados por outro(s) pesquisador(es), daí PRO-DOC

PRO-DOC All resonance Raman measurements were made using the instrumentation described previously (19).
PRO-DOC Fig.1 comparatively depicts two process flow diagrams where (i) the upper represents the conventional approach of discrete mechanical cell disruption followed by aqueous two-phase extraction; and (ii) the lower one represents the integrated process adopted for this study.
* Essa sentença é bastante interessante, pois dá margem a diferentes interpretações. Nesse caso, concluiu-se que o fato de apresentar um fluxograma tradicional, mesmo sem referência, é uma forma de buscar um documento anterior.
PRO-DOC Protein content and G3PDH activity in the disrupted suspension and in phase samples was estimated, following appropriate dilution and centrifugal clarification, using methods described by Gilchrist [6].

Movimento 2, Passo 2: Detalhamento dos procedimentos
PRO-DET Proteins in both fractions were precipitated by the addition of 4 volumes of cold acetone, collected by centrifugation, and resuspended in electrophoresis sample buffer.
PRO-DET An excess of timobesone acetate was added to 3 ml of the surfactant-water or surfactant- propylene glycol –water solution being investigated.
PRO-DET After 1 min of sonication, the suspension was equilibrated for 4 days with rotary mixing in a 25ºC water bath.
PRO-DET Each study was conducted in duplicate or triplicate.
PRO-DET Visual estimates of the volumes of top and bottom phases and solids, were made in graduated centrifuge tubes and used to estimate the volume ratio (Vr=volume of the top phase/volume of the bottom phase).

Movimento 2, Passo 3: Fornecimento de informações sobre os procedimentos
Importante estar atento para o fato de que a categoria PRO-JUST só será anotada nas sentenças que possuem a função por ela desempenhada de maneira explícita, como por exemplo, por meio de um verbo. Aquelas que não tiverem essa marca explícita deverão ser analisadas segundo outra função que possam desempenhar. Vale dizer que, todos os textos da seção metodologia sempre buscam pelo que é mais conveniente para o estudo e essa conveniência não pode ser considerada sempre como uma justificativa, pois levaria o anotador a marcar todas as sentenças da seção metodologia como PRO-JUST.
PRO-JUST They were referred to as Cre-Mate mice, since the nature of the gene targeted for conditional ablation in the epidermis was irrelevant for that study.
PRO-JUST The electronic absorption spectrum of the samples was recorded before and after every experiment to confirm sample purity and stability.
PRO-JUST The dispensed mixtures were collected for pH measurements at the end of the experiments to detect changes caused by mixing with the alkaline stock solutions of peroxynitrite.
PRO-JUST For Fe^2^+CO complexes, laser power was kept at <2 milliwatts to avoid ligand photodissociation.
PRO-DET/PRO-JUST To eliminate excess hydrogen peroxide, peroxynitrite was treated with manganese dioxide.

Movimento 3: Detalhando o equipamento – Sigla EQU
Fornece informação detalhada e relativa ao ambiente de aparelhos e instrumentos utilizados em tarefas específicas de um dado experimento. Os aparelhos mais comuns em procedimentos experimentais são: microscópios, câmeras, spectrofotômetro, citômetro, hemacytometer, etc...
Esse terceiro movimento não apresenta passos.
Uma ressalva importante a ser considerada é a de que essa categoria só será identificada na sentença se a mesma possuir informações sobre o equipamento, tais como modelo, marca, etc. Caso contrário, se ele for apenas citado, então a sentença não deverá ser marcada como EQU, como acontece na seguinte sentença:

PRO-DET Before use, hIDO was gel-filtered through a Sephadex G-25 column eluted with 100 mM phosphate buffer (pH 7.4) containing 100 miM EDTA

Exemplos:
EQU Ultraviolet and visible absorbance measurements were made with a Cary 3 double beam spectrophotometer equipped with a Cary temperature controller from Varian (Sugar Land, Texas).
EQU Images were recorded through a Hamamatsu C-2400 New vicon camera using a 10 x objective and brightfield optics. Video images were digitized at a rate of 6 frames/min as described above.
EQU Simulations were carried out using the Gepasi software, version 3.2 (26, 27).

Movimento 4: Procedimentos de Análise dos Dados – Sigla PAD

Nesse movimento encontra-se relatados os procedimentos de avaliação dos resultados da pesquisa, por exemplo:
PAD The t-test was used to statistically compare the individual ratios from two given strains.
PAD The data were fitted to the Michaelis–Menten Equation 1 by using a non-linear least squares approach and the kinetic constants ± SE.
PAD The initial rate of cysteine oxidation was calculated from data obtained during the first 50% of the reaction, when loss was linear with time.
PAD Simulations were carried out using the Gepasi software, version 3.2 (26, 27).

Movimento 5: Resultados da Pesquisa – Sigla RES

Em sentenças que possuem esse tipo de movimento, encontra-se o objetivo de se relatar muito brevemente o resultado da pesquisa. Maiores detalhes desse resultado podem ser encontrados em outra seção do artigo, específica para esse tipo de relato, a seção Resultados e Discussões.

Exemplos:

RES Measurement of hIDO activity after resonance Raman experiments showed <10% loss of activity due to laser-induced damage to the protein.
RES This activity is comparable to that of native human (89 mol/min/mol; Ref. 18) and rabbit IDO (~108 mol/min/mol; Ref. 11).

4. Processo de anotação – o que é?

Consiste na identificação da função retórica de cada sentença do texto utilizando-se para isso 4 siglas (MAT-LIST, EQU, PAD, etc...), representativas desse papel retórico, as quais serão colocadas no início de cada sentença.

5. Antes do processo de anotação - orientações

- Importante ler o texto antes da anotação, uma vez que a interpretação de determinadas sentenças só se torna possível após uma visão geral do texto.
- Não oriente sua leitura para o entendimento da pesquisa relatada, mas sim para o entendimento da estrutura de argumentação construída pelo autor.
- Não anotar o título ou os subtítulos do texto. Utilize-os apenas como dica/ponteiro do conteúdo que se encontra abaixo dos mesmos.

5.1 Dicas para a identificação dos movimentos

Para se identificar a função retórica de uma dada sentença, aspectos lexicais e gramaticais podem ser de grande ajuda. Exemplos:

1) PAD *The initial rate of cysteine oxidation was calculated from data obtained during the first 50% of the reaction, when loss was linear with time.*

Essa sentença foi classificada como PAD (Procedimentos de Análise de Dados), pois as palavras grifadas acima nos remetem à idéia de que se está descrevendo a maneira como a taxa de dada substância (*cysteine oxidation)* foi obtida.

2) *MAT-FONT Whole antibodies 31154 (human IgG) and 31127 (horse IgG) were obtained from PharMingen.*

Essa sentença recebeu a classificação de MAT-FONT, pois descreve a origem das substâncias que foram utilizadas no experimento. Por meio do verbo "were obtained" seguido da preposição "from" chega-se em PharMingen, nome do laboratório que forneceu as substâncias.

3) PRO-DOC PMNs were isolated from human peripheral blood according to a published procedure (11).

Nota-se que a função dessa sentença é a de relatar o tipo de procedimento utilizado e também de informar que o mesmo já foi realizado por pesquisadores anteriores, como pode ser notado pelo grupo de palavras "published procedure", cujo link indicado pelo número 11 leva à referência desse procedimento. Como resultado da padronização dos procedimentos experimentais, a referência simples ao nome específico do método ou procedimento utilizado para conduzir a pesquisa já é suficiente para ser citado no texto, sem necessidade de maiores detalhes a respeito do procedimento.

4) EQUSamples were analyzed by a FACSCalibur flow cytometer (Becton Dickinson) equipped with CELLQUEST software.

Nessa sentença o que prevalece é a descrição dos equipamentos utilizados para a análise das amostras: um "FACSCalibur flow cytometer (Becton Dickinson)" equipado com um "CELLQUEST software".

6. Durante o processo de anotação - orientações

O processo de anotação (ou classificação) deve ser feito para cada sentença do corpus, que receberá uma classificação (ou categoria). Entretanto, pode haver sentenças que apresentam características de mais de uma categoria, ou seja, sentenças nas quais os papéis argumentativos se sobrepõem, como por exemplo, sentenças que relatam ao mesmo tempo o procedimento e os equipamentos utilizados. Nesses casos, deve-se identificar a sentença, por meio de uma barra entre as categorias possíveis. Exemplo:

PRO/EQU Cells were subjected to centrifugation in a Ficoll Hypaque density gradient (Amersham Pharmacia) to further purify PMNs. PRO/PAD The PMNs then were removed by filtration, and the supernatants were analyzed by HPLC analysis.

Exemplos retirados do texto Met_02 do corpus.

Note ainda que sentenças consecutivas do texto podem receber a mesma classificação. É comum anotar sentenças consecutivas com a mesma categoria, desde que juntas preencham os critérios de uma dada categoria. Por exemplo: é possível marcar mais de uma sentença como MAT (Materiais) se juntas, elas compõem a lista de materiais utilizados na pesquisa.

MAT Whole antibodies 31154 (human IgG) and 31127 (horse IgG) were obtained from PharMingen. MAT Bovine catalase was obtained from Sigma. MAT All assays were carried out in PBS (10 mM phosphate/160 mM sodium chloride, pH 7.4). MAT Commercial protein solution samples were dialyzed into PBS as necessary. MAT Indigo carmine, isatin sulfonic acid, HOCl, H2O2, vinylbenzoic acid, and 4-carboxybenzaldehyde were obtained from Aldrich.

Exemplo de frases consecutivas com mesma categoria (extraídas do corpus):

Se não for possível atribuir nenhuma categoria do esquema utilizado a uma dada sentença, anote-a com um identificador qualquer (por exemplo "?") e descreva, em uma folha a parte, a dificuldade sentida em classificá-la e a função que ela apresenta. Anote também as possíveis

dificuldades na classificação de trechos, bem como com a própria categoria que está sendo utilizada. Ex:

? Thiols, at an initial concentration of 1 mM, together with additives as indicated, were equilibrated to 37°C in a shaking water bath.

Importante:

*Qualquer tipo de dúvida é muito importante e deve ser anotada e levada para discussão com o grupo, pois visa a uma melhor caracterização/adequação de uma categoria problemática.

*Não se esqueça de anotar todas as sentenças do texto e suas eventuais dificuldades, que serão discutidas numa reunião com o grupo de anotadores.

*Anote também quais foram os critérios que utilizou para identificar as funções retóricas e, posteriormente, anotar o texto.

7. Alterações realizadas no corpus

Com base em objetivos e necessidades computacionais, foram feitas algumas alterações no corpus que devem ser levadas em consideração pelo anotador na tarefa de anotação do texto.

1) As letras gregas foram substituídas por sua forma por extenso: ε foi substituído por épsilon, α por alfa, e assim por diante.

2) Os números subscritos de fórmulas, por exemplo, H_2O_2, foram substituídos por H2O2, C_{18} substituído por C18, etc..

3) Os números sobrescritos, por exemplo, o 7 de -10^7, foram reescritos com a adição de acento circunflexo, -10^7.

4) sinais gráficos também foram trocados por sua forma por extenso, por exemplo, ≥ <maior ou igual>, <menor ou igual>, <figura 1>, etc..

8. Texto anotado – texto Met_01 do corpus

INHIBITION OF COPPER-CATALYZED CYSTEINE OXIDATION BY NANOMOLAR CONCENTRATIONS OF IRON SALTS

Munday R, Munday CM, Winterbourn CC.

MAT-FONT D- and L -cysteine, D -penicillamine, cysteamine, homocysteine, desferrioxamine mesylate (DFO), apotransferrin, Tris, glycylglycine, and Pipes were purchased from Sigma.
MAT-FONT GSH was from Roche Molecular Biochemicals.
MAT-FONT Disodium hydrogen orthophosphate monohydrate, sodium dihydrogen orthophosphate dihydrate, copper(II) sulfate pentahydrate, ferrous sulfate heptahydrate, ferrous ammonium sulfate hexahydrate, ferric ammonium sulfate dodecahydrate, ferric chloride hexahydrate, cobalt(II) chloride hexahydrate, nickel(II) chloride hexahydrate, chromium(III) potassium sulfate dodecahydrate, aluminum ammonium sulfate dodecahydrate, and zinc sulfate heptahydrate were Analar-grade chemicals from BDH.
MAT-FONT Chelex 100 resin was a product of Bio-Lab.
MAT-INFO/MAT-FONT Catalase was purchased from Sigma (Type C-40, specific activity 17,890 Sigma units/mg, assayed at pH 7.0 and 25°C by the rate of decrease in absorbance of a 10.3 mM solution of hydrogen peroxide at 240 nm) and from Roche Molecular Biochemicals (Catalog No. 106 810, specific activity 75,080 Sigma units/mg).

MEASUREMENT OF THIOL OXIDATION

PRO-DET All reactions were conducted in 25 mM buffer which, except where indicated, was at pH 7.25.

PRO-DET Buffers and reagents were made up in new plastic containers, and all contact with glass was avoided.
PRO-DET Thiols, at an initial concentration of 1 mM, together with additives as indicated, were equilibrated to 37°C in a shaking water bath.
PRO-DET The oxidation reaction was initiated by addition of an aqueous solution of cupric sulfate.
PRO-DOC/EQU At intervals, samples were removed for analysis of remaining thiol by the 5,5'-dithiobis (2-nitrobenzoic acid) (DTNB) method of Ellman [28], using a Shimadzu UV-1601 spectrophotometer (épsilon 412 = 14,100).
PAD The initial rate of cysteine oxidation was calculated from data obtained during the first 50% of the reaction, when loss was linear with time.
PRO-DET Where indicated, metal contamination of buffers was eliminated by passage through a column of Chelex resin, washed and adjusted to pH 7.25 [27].
PRO-DET When used, DFO was added at a concentration of 20 miM.
PRO-DET Ferrous salts were added as aqueous solutions.
PRO-JUST Ferric chloride and ferric ammonium sulfate were added as solutions in 0.01 N hydrochloric acid, and 0.01 N sulfuric acid, respectively, to avoid hydrolysis.
PRO-DET For treatment with apotransferrin, buffer (40 ml) was dialyzed against apotransferrin (30 mg in 5 ml of purified buffer, contained in a dialysis bag) with stirring for 72 h.

Referências

Mauranen, A. (1993) Contrastive ESP rhetoric: Metatext in Finnish-English economic texts. English for Specific Purposes,12, 3-22.

Swales, J.M. Genre Analysis: English in Academic and Research Settings. Cambridge Applied Linguistics series, 1990.

Weissberg, R.; Buker, S. *Writing up Research: Experimental Research Report Writing for Students of English*. Prentice Hall, 1990.

Apêndice 2 - ***Manual de Anotação das Estruturas Esquemáticas e Estratégias Retóricas de Abstracts***

As orientações abaixo descrevem o esquema de anotação manual das estruturas esquemáticas e estratégias retóricas para a seção "Resumo" de córpus de artigos científicos em inglês.

I. Artigo científico: estrutura

Uma característica comum a praticamente todos os textos científicos, que descrevem pesquisa experimental, é o tipo de organização que sua estrutura esquemática deve seguir. Essa estrutura pode ser apresentada como Introdução, Desenvolvimento e Conclusão, sendo que o Desenvolvimento pode ser subdividido em Materiais e Métodos e Resultados, ou ainda Materiais e Métodos, Resultados e Discussão. O objetivo desse tipo de estruturação é guiar o leitor e fazer com que ele siga, na leitura ou escrita do texto, o fluxo da informação a ser transmitida que parte do geral-para-específico na Introdução e chega ao específico-para-geral, na Conclusão, conforme pode ser observado na figura abaixo.

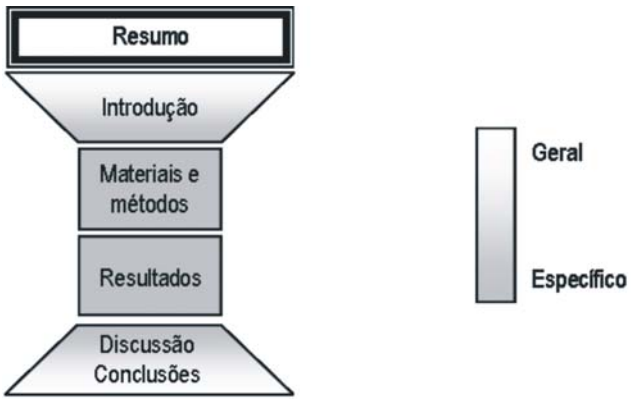


Figura 1: Movimento Geral-Específico-Geral presente na estrutura global do texto científico (Weissberg & Buker, 1990)

II. O que é a seção "Resumo"?

O Resumo corresponde à primeira seção de um artigo científico. Em geral aparece depois do título e antes da introdução. Em algumas áreas do conhecimento pode aparecer grafado como *summary*. No resumo estão contidas informações que demonstram, de forma breve, uma prévia do estudo que foi realizado. É importante, portanto, que seja elaborado depois de concluído o trabalho, pois conterá informações das outras seções do texto. Essa seção é muito importante num artigo, pois muitos leitores se limitam a ler o título e o resumo de um trabalho para decidirem se o artigo lhes interessa. Dessa forma, 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. Vale ressaltar ainda que, com o crescimento do uso de repositórios on-line de trabalhos acadêmicos, o resumo passa a ter um papel ainda mais importante, já que tais repositórios muitas vezes disponibilizam apenas o resumo de um trabalho. Os resumos de quase todas as áreas de estudo são escritos de uma maneira muito similar. 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 sobre o tipo de informação que deve ser incluída em um bom resumo e da ordem na qual tais informações devem aparecer. Assim, o Quadro 1 contém as informações típicas que, segundo Weissberg & Buker (1990:186), podem ser encontradas, em geral, em resumos de quaisquer áreas do conhecimento.

Contexto – conhecimento aceito pela comunidade científica **Lacuna** – problema de Pesquisa, necessidade **Propósito** – principal atividade da pesquisa **Metodologia** – alguma informação sobre a metodologia utilizada **Resultados** – os resultados mais importantes obtidos **Conclusão** – conclusões, recomendações, contribuições, etc.

Quadro 1: Informações retóricas ou estruturas esquemáticas típicas de resumos.

II.1 O modelo de Weissberg & Buker (1990)

A análise dos Movimentos, desenvolvida por Weissberg & Buker (1990), visa representar artigos científicos em termos de sua organização textual hierárquica construída por seções distintas. Cada seção pode ser dividida em estruturas esquemáticas (Contexto, Lacuna, Propósito, Metodologia, Resultados e Conclusão) e cada Estrutura Esquemática, por sua vez, pode ser subdividida em Estratégias Retóricas.

A estrutura esquemática consiste em uma unidade textual funcional, utilizada com algum propósito retórico identificável. Essas estruturas realizam-se lingüisticamente em um texto de diferentes formas, ou seja, por diferentes estratégias retóricas, conforme a a estrutura esquemática na qual se encontram contidas.

III Processo de anotação – o que é?

Consiste na identificação da função retórica de cada sentença do texto utilizando-se, para isso, siglas representativas desse papel retórico, as quais serão colocadas no início de cada sentença.

IV Antes do processo de anotação - orientações

- Importante ler o texto antes da anotação, uma vez que a interpretação de determinadas sentenças só se torna possível após uma visão geral do texto.
- Ao anotar um texto que não seja da área em que atue, não oriente sua leitura para o entendimento dos pormenores da pesquisa relatada, mas sim para o entendimento da estrutura de argumentação construída pelo autor.
- Não anote o título ou os subtítulos do texto; utilize-os apenas como dica/ponteiro do conteúdo que se encontra abaixo dos mesmos.

V Durante o processo de anotação - orientações

O processo de anotação (ou classificação) deve ser feito para cada sentença do corpus[1], que receberá uma classificação (ou categoria). Entretanto, pode haver sentenças que apresentam características de mais de uma categoria, ou seja, sentenças nas quais os papéis argumentativos se sobrepõem, como por exemplo, sentenças que relatam ao mesmo tempo o procedimento e os equipamentos utilizados. Nesses casos, deve-se identificar todas as categorias possíveis por meio de uma barra.

Exemplo:

PRO/EQU Cells were subjected to centrifugation in a Ficoll Hypaque density gradient (Amersham Pharmacia) to further purify PMNs. PRO/PAD The PMNs then were removed by filtration, and the supernatants were analyzed by HPLC analysis.

[1] Pode ser definido, grosso modo, como uma coleção de textos ou partes de textos de uma determinada língua, escolhidos segundo determinados critérios. No caso desse manual, os textos escolhidos visam fazer parte da base de casos de uma ferramenta de auxílio à escrita científica.

Exemplos retirados do texto Met_02 do córpus Met composto por seções "Metodologia" de artigos científicos da área de Ciências Farmacêuticas. As siglas PRO, EQU e PAD significam, respectivamente: Procedimentos, Equipamentos e Processamento de Análise de Dados.

Note ainda que sentenças consecutivas do texto podem receber a mesma classificação. É comum anotar sentenças consecutivas com a mesma categoria, desde que juntas preencham os critérios de uma dada categoria. Por exemplo, é possível marcar mais de uma sentença como MAT (Materiais) se juntas elas compõem a lista de materiais utilizados na pesquisa.

MAT Whole antibodies 31154 (human IgG) and 31127 (horse IgG) were obtained from PharMingen.
MAT Bovine catalase was obtained from Sigma.
MAT All assays were carried out in PBS (10 mM phosphate/160 mM sodium chloride, pH 7.4).
MAT Commercial protein solution samples were dialyzed into PBS as necessary.
MAT Indigo carmine, isatin sulfonic acid, HOCl, H2O2, vinylbenzoic acid, and 4-carboxybenzaldehyde were obtained from Aldrich.

Exemplo de sentenças consecutivas com mesma categoria (extraídas do córpus Met):

Se não for possível atribuir nenhuma categoria do esquema utilizado a uma dada sentença, anote-a com um identificador qualquer (por exemplo "?") e descreva, em uma folha a parte, a dificuldade sentida em classificá-la e a função que ela apresenta. Anote também as possíveis dificuldades na classificação de trechos, bem como com a própria categoria que está sendo utilizada. Por exemplo:

? Thiols, at an initial concentration of 1 mM, together with additives as indicated, were equilibrated to 37°C in a shaking water bath.

Importante:

*Qualquer tipo de dúvida é muito importante e deve ser anotada e levada para discussão com o grupo, pois visa a uma melhor caracterização/adequação de uma categoria problemática.

*Não se esqueça de anotar todas as sentenças do texto e suas eventuais dificuldades, que serão discutidas numa reunião com o grupo de anotadores.

*Anote também quais foram os critérios que utilizou para identificar as funções retóricas e, posteriormente, anotar o texto.

VI. Categorias para a anotação retórico-manual do texto

As categorias escolhidas para realizar a anotação retórica dos textos foram inspiradas no modelo de estruturação retórica proposto por Weissberg & Buker (1990). A seguir, no Quadro 2, temos uma lista de siglas que correspondem às possíveis seções que podem ser encontradas na seção "Resumo" de um artigo experimental. Na verdade, existem 6 categorias que correspondem às estruturas esquemáticas de Weissberg & Buker (1990): Contexto (COT), Lacuna (LAC), propósito (PRO), Metodologia (MET), Resultado (RES) e Conclusão (COC). Essas estruturas esquemáticas (que correspondem as três primeiras letras das siglas contidas na Figura 2) podem se realizar de diferentes formas na língua, que são as estratégias retóricas, representadas pelo restante de letras das siglas abaixo.

Essas siglas foram elaboradas para cada uma dessas categorias, a fim de facilitar o processo de anotação do córpus.

Siglas das categorias	Descrição das siglas
COT-FOP	Familiarizar termos, objetos e processos
COT-RPA	Citar resultados de pesquisas anteriores
COT-AHI	Apresentar hipóteses
COT-DPT	Declarar proeminência do tópico
LAC-CPD	Citar problemas/dificuldades
LAC-CNR	Citar necessidades/requisitos
LAC-APA	Citar a ausência ou falta da pesquisa anterior
PRO-AMP	Apresentar mais propósitos
PRO-APP	Apresentar o propósito principal
PRO-APM	Apresentar o propósito com a metodologia
PRO-APR	Apresentar o propósito com os resultados
MET-CMM	Citar/Descrever materiais e métodos
MET-LCC	Listar critérios ou condições
RES-CDR	Comentar/Discutir os resultados
RES-DR	Descrever os resultados
RES-IR	Indicar os resultados
COC-ACP	Apresentar contribuições/valor da pesquisa
COC-AR	Apresentar recomendações
COC-AC	Apresentar conclusões

Figura 2: Note, na figura acima, que a sigla da categoria é composta sempre por letras que compõem uma dada categoria, de forma a facilitar a memorização e fácil identificação do significado da categoria que deverá ser empregada nas sentenças do corpus.

VI.2 Detalhamento das categorias

Nessa seção, são apresentadas as estruturas a serem identificadas nos textos de forma mais detalhada e com exemplos autênticos retirados de um córpus formado por artigos científicos da área de Ciências Farmacêuticas. As definições das estruturas esquemáticas abaixo apresentadas tiveram suas definições baseadas no trabalho de Feltrim (2004).

Vale lembrar que essas estruturas (categorias) visam descrever de maneira geral as funções retóricas que podem ser encontradas em um córpus. Mas é bem possível que outras funções possam ser encontradas no córpus em análise e acrescentadas a esse modelo. Aliás, esse modelo é um ponto de partida para que seja aprimorado segundo as características que o córpus em estudo/análise apresentar.

Estrutura Esquemática 1- Contexto – Sigla COT

São sentenças que apresentam conhecimento já reconhecido em uma 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. Porém, há casos nos quais elas podem acontecer de sentenças de contexto aparecer no meio do texto e, nesses casos, uma sugestão para auxiliar na identificação de sentenças de Contexto é usar o seguinte teste: se a sentença poderia aparecer no início do texto e ela não contém material próprio da pesquisa apresentada pelo autor, então anote como Contexto.

Sentenças de Contexto podem conter citações. Em geral, essas citações são "pioneiras" da área, de 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.

O Contexto pode ser realizado das seguintes maneiras, ou seja, por meio das seguintes estratégias retóricas:

Estratégia Retórica - Familiarizar termos, objetos e processos – Sigla - COT-FOP
Problems caused by the presence of adventitious metals in buffers and reagents are well recognized in studies of metal-catalyzed oxidation reactions.
In most cases, metal contamination leads to an increase in rate, and chelating agents are inhibitory.
Estratégia Retórica - Citar resultados de pesquisas anteriores – Sigla - COT-RPA
Recent studies have suggested that antibodies can catalyze the generation of previously unknown oxidants including dihydrogen trioxide (H2O3) and ozone (O3) from singlet oxygen (1O.2) and water.
To identify the signaling mechanisms, we evaluated patterns of cross-desensitization between SAA and other leukocyte chemoattrctants.
Estratégia Retórica - Apresentar hipóteses – Sigla - COT-AHI
Given that neutrophils have the potential both to produce 1O.2 and to bind antibodies, we considered that these cells could be a biological source of O3.
Estratégia Retórica - Declarar proeminência do tópico – Sigla - COT-DPT
Antioxidants can modulate the expression of immune and inflammatory genes, and pyrrolidine dithiocarbamate (PDTC) is a frequently used antioxidant to inhibit the transcription factor NF-kappaB.
Oxidative injury is implicated in the development of chronic lung disease in preterm infants with respiratory distress.

Estrutura Esquemática 2- Lacuna – Sigla LAC

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 discursivos de contraste, como h*owever*, *although*, *but*, *in contrast*, *on the other hand*, 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 contém 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.

A Lacuna pode ser realizada das seguintes maneiras, ou seja, por meio das seguintes estratégias retóricas:

Estratégia Retórica - Citar problemas/dificuldades – Sigla - LAC-CPD

However , direct evidence of a causal role is limited and the source of reactive oxidants has not been identified.
Experiments in many laboratories have been limited by the availability of the enzyme, and its production required at least a week of work to complete its purification.
Estratégia Retórica - Citar a ausência ou falta da pesquisa anterior – Sigla - LAC-APA
Although DNA codon optimization is a standard molecular biology strategy to overcome poor gene expression, to date no public software exists to facilitate this process.
Estratégia Retórica - Citar necessidades/requisitos - Sigla - LAC-CNR
To make a colonic delivery system practical for medical use, in vitro testing methods need to be established in order to determine the specifications of the preparations.

Estrutura Esquemática 3- Propósito – Sigla PRO

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 que podem existir propósitos secundários, principalmente quando se tratam de teses e dissertações. Tanto a sentença que apresenta o propósito principal, como as sentenças que detalham o propósito e/ou que introduzem 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 Propósito pode ser realizado das seguintes maneiras, ou seja, por meio das seguintes estratégias retóricas:

Estratégia Retórica - Apresentar mais propósitos – Sigla - PRO-AMP
We also show that activation provokes the influx of an enormous concentration of ROS into the endocytic vacuole.
We also observed that expression of IDO by immunogenic mouse tumor cells prevents their rejection by preimmunized mice.
Estratégia Retórica - *Apresentar o propósito principal* – Sigla - PRO-APP
We show *here* that this simple scheme, which for many years has served as a satisfactory working hypothesis, is inadequate.

These observations suggest that cells expressing IDO inhibit T cell responses in vivo.
Estratégia Retórica - Apresentar o propósito com a metodologia – Sigla - PRO-APM
To directly evaluate the hypothesis that cells expressing IDO inhibit T cell responses, we prepared IDO-transfected cell lines and transgenic mice overexpressing IDO and assessed allogeneic T cell responses in vitro and in vivo.
We have used resonance Raman spectroscopy to characterize the heme environment of purified recombinant human indoleamine 2,3-dioxygenase (hIDO).

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.

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:

Estratégia Retórica 11 - Apresentar o propósito com os resultados – Sigla - PRO-APR
Here we show that K+ crosses the membrane through large-conductance Ca2+-activated K+ (BKCa) channels.

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, tome nota do identificador do texto para que ele seja posteriormente revisado.

Estrutura Esquemática 4- Metodologia – Sigla MET

Sentenças descrevendo a metodologia utilizada para a realização da pesquisa devem ser marcadas como Metodologia. Sentenças de Metodologia geralmente aparecem *após* o Propósito principal, mas isso não é uma regra. Marque como Metodologia apenas as sentenças relacionadas a 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*. Por métodos entendemos métodos/técnicas/abordagens/etc.
- Sentenças indicando trabalhos nos quais a pesquisa é baseada 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.
- Procedimentos utilizados para a *avaliação*/comprovação dos resultados, como *estudo de caso* e *testes empíricos*.

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.

A Metodologia pode ser realizada das seguintes maneiras, ou seja, por meio das seguintes estratégias retóricas:

Estratégia Retórica - Citar/Descrever materiais e métodos – Sigla - MET-CMM
Vinylbenzoic acid, an orthogonal probe for ozone detection, is oxidized by activated neutrophils to 4-carboxybenzaldehyde in a manner analogous to that obtained for its oxidation by ozone in solution.

The distal pocket of Fe3+ hIDO was explored further by an exogenous heme ligand, CN; again, binding of L-Trp introduced strong H-bonding and/or steric interactions to the heme bound CN.
Estratégia Retórica - Listar critérios ou condições – Sigla - MET-LCC
The radical was unambiguously identified by its EPR parameters (g = 2.0113; line width = 5.5 G) and by experiments with bicarbonate labeled with 13C.
Genes encoding mediators of inflammation and host defense, including CD11c, CD14, CD54, FcR1, FcR, CD120b, TLR5, IL-4R, CCR1, p47phox, p40phox, IL-8, CXCL1, Nramp1, and calgranulins A and B, were up-regulated constitutively in unstimulated XCGD patient PMNs.

Estrutura Esquemática 5- Resultado – Sigla RES

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 (RES) sentenças que:

- Descrevem um artefato. 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" os resultados de experimentos.
- Descrevem ou "indicam" os resultados de avaliações.
- Comentam/discutem os resultados da pesquisa.

Perceba que sentenças que apenas indicam a existência de resultados também devem ser marcadas como Resultado. O Resultado pode ser realizado das seguintes maneiras, ou seja, por meio das seguintes estratégias retóricas:

Estratégia Retórica - Discutir os resultados – Sigla - RES-CDR

These effects are attributable to inhibition of copper-catalyzed oxidation by adventitious iron.

In addition , adoptive transfer of alloreactive donor T cells yielded reduced numbers of donor T cells when injected into IDO-transgenic recipient mice.

Estratégia Retórica - Descrever os resultados – Sigla - RES-DR

In purified buffer at pH 7.25, containing 0.4 M copper, cysteine was oxidized at a rate of 32 M/min.

IDO inhibitor treatment triggered extensive inflammation at the maternal-fetal interface in susceptible mating combinations, which was characterized by complement deposition and hemorrhagic necrosis.

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 (COC) 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/contribuições" podem aparecer em outro contexto. Nesses casos, você deve usar o 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.

Estrutura Esquemática 6- Conclusão – Sigla COC

Podem ocorrer casos de sentenças que têm 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 (COC), assim como sentenças que enfatizam a importância teórica do estudo no avanço do estado da arte em uma área de pesquisa específica.

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 àqueles resultados, fazendo um movimento do tipo "Resultado – Conclusão – Resultado – Conclusão...".

Pode acontecer do texto não conter nenhuma sentença de Conclusão. A Conclusão pode ser realizada das seguintes maneiras, ou seja, por meio das seguintes estratégias retóricas:

Estratégia Retórica - Apresentar contribuições/valor da pesquisa – Sigla - COC-ACP

Our results offer an explanation for the conflicting literature reports of the effects of chelating agents and catalase on cysteine oxidation, and emphasize the need for buffer purification or addition of DFO in studies concerned with the oxidation or cytotoxicity of this thiol.

Importantly , in addition to contributing to the understanding of nitrosoperoxocarboxylate decomposition pathways, this is the first report unambiguously demonstrating the formation of the carbonate radical anion at physiological pHs by direct EPR spectroscopy.

Estratégia Retórica - Apresentar recomendações - Sigla - COC-AR
The exceptional sensitivity of copper-catalyzed cysteine oxidation to iron makes this an attractive system for monitoring the iron content of buffers, and may also have application for determining the free iron content of physiological fluids.
Moreover , it can be used to optimize any other genes of interest and is freely available online at http://www.vectorcore.pitt.edu/upgene.html.
Estratégia Retórica - Apresentar conclusões - Sigla - COC-AC
Remarkably , microbial killing and digestion were abolished when the BKCa channel was blocked, revealing an essential and unexpected function for this K+ channel in the microbicidal process.
We show that it is the proteases, *thus* activated, that are primarily responsible for the destruction of the bacteria.

O texto a seguir é um resumo que possui sua estrutura esquemática e respectivas estratégia retóricas anotadas. A primeira identificada pela primeira parte da sigla (lado esquerdo do hífen da sigla) e a segunda pelo lado direito da sigla. Trata-se de um resumo da área de Ciências Farmacêuticas e foi retirado da Base de Casos do ambiente SciPo-Farmácia (http://www.nilc.icmc.usp.br/scipo-farmacia/).

Resumo - Caso ab_01
Link:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=14990354&itool=iconabstr
Inhibition of copper-catalyzed cysteine oxidation by nanomolar concentrations of iron salts
Munday R, Munday CM, Winterbourn CC.
COT-FOP Problems caused by the presence of adventitious metals in buffers and reagents are well recognized in studies of metal-catalyzed oxidation reactions.
COT-FOP In most cases, metal contamination leads to an increase in rate, and chelating agents are inhibitory.
PRO-APR In the present study, *however* , the rate of copper-catalyzed oxidation of cysteine was found to be increased by buffer purification with Chelex resin or by addition of micromolar concentrations of the specific iron chelator desferrioxamine (DFO).
RES-CDR These effects are attributable to inhibition of copper-catalyzed oxidation by adventitious iron.
RES-DR In purified buffer at pH 7.25, containing 0.4 M copper, cysteine was oxidized at a rate of 32 M/min.
RES-DR Addition of iron salts to this buffer caused a dose-related decrease in this rate, up to a maximum of 85%.
RES-DR A 50% decrease in rate was recorded at an iron concentration of only 11 nM.
RES-DR Other transition metals were without effect.
RES-CDR Similar effects of purification or addition of DFO on the rate of cysteine oxidation were seen in Tris, glycylglycine, Mops, and Pipes buffers.
RES-CDR Catalase decreased the rate of cysteine oxidation, *but* the sensitivity to iron was similar in the presence and absence of catalase.
RES-CDR Copper-catalyzed oxidation of cysteamine and reduced glutathione was much less sensitive to inhibition by iron.
COC-ACP Our results offer an explanation for the conflicting literature reports of the effects of chelating agents and catalase on cysteine oxidation, and emphasize the need for buffer purification or addition of DFO in studies concerned with the oxidation or cytotoxicity of this thiol.
COC-AR The exceptional sensitivity of copper-catalyzed cysteine oxidation to iron makes this an attractive system for monitoring the iron content of buffers, and may also have application for determining the free iron content of physiological fluids.

Apêndice 3: Manual para anotação de Marcadores Discursivos de artigos científicos

Orientações para a Anotação Manual de Marcadores Discursivos

I) O que são Marcadores Discursivos?

Segundo Fraser (1996; 2005:1) os marcadores discursivos constituem uma classe de palavras que são como "pistas lingüisticamente codificadas, as quais sinalizam as intenções comunicativas potenciais do falante (Schourup, 1999 apud Paizan, 2001)". Em outras palavras, os marcadores discursivos são conjunções, advérbios, locuções, etc., que servem para estabelecer uma relação lógica entre frases e idéias, bem como a função retórica de uma dada sentença, por exemplo. Em português essas palavras são chamadas de *articuladores* ou *conectores* (Schütz, 2006). O uso correto destas palavras confere solidez ao argumento, fluidez na leitura, trama textual adequada e conseqüentemente elegância ao texto. As funções que podem desempenhar em textos científicos podem ser contraste/oposição, adição, conseqüência/resultado, e assim por diante.

Ex. de marcador discursivo com função de Contraste: *However* , direct evidence of a causal role is limited and the source of reactive oxidants has not been identified.

Ex. de marcador discursive com função de Conclusão: We show that it is the proteases, *thus* activated, that are primarily responsible for the destruction of the bacteria.

Uma lista de marcadores discursivos retirados de artigos científicos pode ser encontrada a seguir. Tal lista pode ser utilizada como um tipo de referência na identificação de marcadores discursivos em um dado artigo, bem como para classificar o marcador discursivo encontrado segundo a função que o mesmo estiver desempenhando na sentença em que foi encontrado. Caso o marcador discursivo não seja encontrado na lista abaixo, ele deve ser adicionado a ela, com o auxílio das funções retóricas que organizam a lista de marcadores abaixo (Contraste/Oposição, Comparação, Adição, etc.) associada ao contexto de uso desse marcador, ou seja, a sentença no qual ele aparece podem auxiliar na classificação do mesmo e em sua posterior inserção seguinte lista.

II) Lista de Marcadores Discursivos

Função do Marcador Discursivo	**Tipo de Marcador Discursivo**
1.1 Relacionador de Mensagem	
Contraste/Oposição	1. (al)though 2. after (all) 3. alternatively 4. as opposed to 5. but 6. alternately 7. alternatively 8. conversely 9. despite 10. even though 11. for all that 12. for (my, his, ...) part 13. however 14. in contrast 15. despite (doing) this/that 16. in (the) face of

	17. in (the) light of 18. meanwhile 19. nevertheless 20. nonetheless 21. notwithstanding 22. on the contrary 23. in comparison (with/to this/that) 24. in contrast (with/to this/that) 25. in spite of (doing) this/that 26. instead (of (doing) this/that) 27. nevertheless 28. nonetheless 29. on the contrary 30. on the other hand 31. on the other hand 32. still 33. then again 34. though 35. unlike 36. whereas 37. while 38. yet 39. rather (than (do) this/that
1.2 Elaborativo	
Comparação	40. as well as 41. both 42. comparability 43. comparatively 44. comparing 45. either 46. equally 47. in comparison (with) 48. in the same way 49. likewise 50. likewise 51. similarly
Adição	52. above all 53. additionally 54. again 55. also 56. and (then) 57. as well as 58. at the same time 59. besides 60. both ... and 61. either ... or 62. equally important 63. further 64. furthermore 65. in addition (to) 66. indeed 67. jointly 68. last but not least 69. likewise 70. moreover

	71. neither ... nor 72. next 73. not only ... but also 74. not to mention 75. not to speak of 76. moreover 77. on top of that 78. or 79. plus 80. similarly 81. together with 82. what's more 83. subsequently 84. together 85. what is more
Adição de maneira mais específica	86. In particular 87. specifically 88. sspecially 89. principally
Reformulação da informação anterior	90. that is (i.e.) 91. (or) rather 92. in other words
Exemplificação	93. e.g. 94. for example 95. for instance 96. in another case 97. in particular 98. in this case 99. in this manner 100. including 101. namely 102. such as 103. take the case of 104. that is 105. the following example 106. to illustrate
Estruturação da informação em forma de lista	107. after that 108. finally 109. first 110. first of all 111. firstly 112. following 113. in the first place 114. initially 115. last 116. later 117. next 118. other 119. second 120. secondly 121. then
1.3 Inferencial	
Conseqüência/resultado	122. accordingly 123. accordingly 124. as

	125. as a consequence 126. as a result 127. because of this/that 128. consequently 129. due to 130. hence 131. in order that 132. in this/that case 133. now that 134. of course 135. that is because 136. that is why 137. the main reason 138. thereby 139. therefore 140. thus
Conclusão	141. after all 142. after that 143. all in all 144. all things considered 145. as a conclusion 146. as I have said 147. as we have seen 148. at last 149. evidently 150. finally 151. importantly 152. in brief 153. in conclusion 154. in other words 155. in short 156. in summary 157. interestingly 158. last(ly) 159. on the whole 160. otherwise 161. overall 162. relatively 163. significantly 164. so 165. then 166. thus 167. to conclude 168. to sum up 169. to summarize
1.4 Explicação	170. Because 171. for this/that reason 172. since 173. Indeed 174. Towards 175. While
2. Relacionador de Tópico	
Digressão	176. By the way 177. incidentally 178. before I forget

Reintrodução de um tópico	179. Speaking of X 180. with regards to 181. to return to my point
Especificação	182. a key feature 183. a major concern 184. above all 185. definitely 186. especially not 187. especially significant 188. even more 189. here 190. i this paper 191. in any event 192. in fact 193. in particular 194. in this report 195. in this study 196. indeed 197. least of all 198. let alone 199. most important(ly) 200. most of all 201. naturally 202. particularly 203. positively 204. primarily 205. principally 206. specifically 207. the basic cause 208. the chief factor 209. the key point 210. the main reason 211. unquestionably 212. valuable to note 213. without a doubt
Recomendação	214. Can be used for efficient 215. Need to
Tempo	216. after a while 217. afterward(s) 218. as time goes by 219. at last 220. at present 221. at the same time 222. at this point 223. biweekly 224. briefly 225. constantly 226. continuously 227. currently 228. daily 229. generally 230. here 231. immediately 232. in the meantime 233. lately

	234. later 235. meanwhile 236. more recently 237. normally 238. now 239. nowadays 240. occasionally 241. presently 242. previously 243. rapidly 244. recently 245. regularly 246. routinely 247. shortly (after) 248. simultaneously 249. since 250. so far 251. soon 252. temporarily 253. then 254. thereafter 255. thereupon 256. throughout 257. to date 258. typically 259. until 260. up until now 261. while 262. yet at the same time
Gradação das informações	263. Inasmuch 264. These 265. That 266. This
Restrição	267. Simply 268. Strictly 269. Apart
Argumento mais forte no sentido de determinada conclusão	270. Alternatively
Valoração/Destaque do trabalho	271. Significantly 272. Unlikely 273. Satisfactorily 274. Importantly 275. Notably 276. Remarkably 277. Useful 278. Valuable 279. Usefulness
Faz referência ao que foi dito anteriormente	280. Previously 281. Preliminarily 282. Respectively 283. Whose 284. Whereas 285. This 286. These results
Sumarização	287. Overall

	288. Collectively 289. Finally
Seqüência	290. Subsequently 291. Successively 292. Repeatedly 293. Sequentially 294. Randomly 295. Arbitrarily
Modo	296. Slowly 297. Kindly 298. Closely 299. Especially 300. Gradually 301. Quantitatively 302. Automatically 303. Stably 304. Carefully 305. Gently 306. Regularly 307. Directly 308. Originally 309. Freely 310. Essentially 311. Thoroughly 312. Generously 313. Differentially 314. Similarly 315. Separately 316. Selectively 317. Relatively 318. Freshly 319. Independently 320. Individually 321. Homogeneously 322. Spontaneously 323. Commercially 324. Conventionally
Imprecisão/incerteza	325. Possibly 326. Probably 327. Approximately 328. Apparently 329. Nonspecifically
Precisão	330. Exactly
Intensidade	331. Completely 332. Entirely 333. Highly 334. Partially 335. Increasingly Exhaustively 336. Vigorously 337. Tightly 338. Slightly 339. Moderately
Lugar	340. Nearly 341. Centrally

	342. Externally 343. Internally
Método /"por meio de"	344. Numerically 345. Statistically 346. Verbally 347. Visually 348. Fluorometrically 349. Thermally 350. Anaerobically 351. Intravenously 352. Intraperitoneally 353. Subcutaneously 354. Clinically

Referências **Schütz,** Ricardo. "Words of Connection (Conectivos)" English Made in Brazil <http://www.sk.com.br/sk-conn.html>. Acessado em 22 de novembro de 2006.

Apêndice 4: Rubrica utilizada na Avaliação Manual de Qualidade de Escrita Científica da Fase 2 de avaliação do processo semi-automático proposto.

A seguir serão apresentadas as três dimensões utilizadas na segunda fase de avaliação do processo semi-automático proposto por esta pesquisa. Esses três critérios são utilizados para analisar o resumo completo, diferentemente da rubrica investigada, na qual os dois primeiros critérios é que analisam o resumo todo e é atribuído um valor Alto ou Baixo. E a partir do terceiro critério os valores são atribuídos a cada sentença do texto em análise.

Dimensão 1 – Organização e Desenvolvimento de um texto: Esse critério é indicado para investigar a estrutura esquemática contida em uma dada seção de um artigo científico, no caso, a seção resumo. Ela objetiva tanto a identificação de componentes essenciais a essa seção em foco quanto à verificação da ordem que esses componentes devem aparecer no texto. Para tal avaliação são utilizados dois valores: Alto e Baixo. O valor Alto é atribuído quando os componentes principais da estrutura esquemática estão presentes e são apresentados em ordem lógica. Por exemplo, na seção "Resumo" a estrutura esquemática principal apresentaria a seguinte ordem: Propósito, Metodologia (se houver), Resultados principais e Conclusão. Como nem todos os resumos apresentam a mesma ordem proposta por esse modelo ideal de estrutura esquemática de um resumo considerado adequado as especificações dos pesquisadores sobre escrita científica, a ordem dos componentes presentes deve obedecer a uma lógica que satisfaça as expectativas do leitor, ou seja, deve conter uma ordem que apresente de maneira lógica as informações descritas. Assim, se houver uma Lacuna, esta deve ser seguida pelo Propósito. Se existir Contexto e Lacuna, a Lacuna deve aparecer depois do Contexto. Mas é possível também haver ciclos de Contexto e Lacuna. O valor Baixo é atribuído quando as condições descritas acima não forem satisfeitas.

Dimensão 2 – Balanceamento entre os componentes: Essa dimensão visa verificar o balanceamento do tamanho de cada uma das seções de um artigo científico, em separado. Por exemplo, os resumos, em geral, não devem ultrapassar um limite de 200 a 300 palavras, o que implica na imposição de algumas restrições ao uso de dadas estruturas esquemáticas utilizadas em resumos, como não supervalorizar a escrita de um contexto com várias sentenças. Para tal verificação são também utilizados os valores Alto e Baixo. O valor Alto é atribuído em resumos escritos em inglês na área de Ciências Farmacêuticas quando: 1) O Propósito existe e foi escrito em apenas uma sentença; 2) A Conclusão existe e foi escrita em apenas uma sentença; 3) Se existir Contexto, não deve ultrapassar 30% das palavras de um *abstract*. O valor Baixo é atribuído quando as condições descritas acima não forem satisfeitas.

Para se fazer a verificação do balanceamento das estruturas esquemáticas em outras seções de artigos científicos, que sejam ou não da área de Farmácia é interessante realizar um levantamento empírico com a área e a seção de artigo científico para a qual se deseja verificar o tipo de balanceamento da estruturação esquemática mais recorrente. Os artigos utilizados na extração de informação para a caracterização da dimensão anterior poderão ser utilizados nessa mesma etapa.

Dimensão 3 – Coerência entre os componentes

Essa rubrica visa a avaliar a coerência entre as estruturas esquemáticas de uma seção, ou seja, verificar se as estruturas estão relacionadas entre si de forma a contribuir com a coerência do texto. A coerência pode, a grosso modo, ser definida como o resultado de uma não-contradição entre os diversos segmentos de um texto, que devem estar encadeados logicamente. Cada segmento textual é pressuposto do seguimento que vem a seguir, que por

sua vez será pressuposto para o(s) que lhe sucederem, formando assim uma corrente, uma cadeia na qual todos os segmentos estejam concatenados de maneira harmônica. Quando um segmento está em contradição com um anterior, perde-se coerência textual. Para a verificação da coerência de uma dada seção de um artigo científico são também utilizados os valores Alto e Baixo. Abaixo serão apresentados os critérios que devem estar presentes no momento de verificação da coerência de resumos. Para as outras seções de um artigo científico, outros critérios de coerência devem ser criados, de modo que as características peculiares da seção a ser avaliada sejam consideradas.

O valor **Alto** é atribuído a coerência de um resumo quando:

- Se o Propósito estiver relacionado com a Lacuna, em uma relação de *fullfilment,* isto é, é observado o desejo de realizar alguma tarefa. Interessante notar que como a Lacuna não é um item obrigatório, quando não está presente, o Propósito é assumido como padrão.
- Se os Resultados principais estiverem relacionados com o Propósito, em uma relação de *accomplishment*, isto é, a intenção de realização, alcance.
- Se a Conclusão estiver relacionada com os resultados principais, em uma relação de *generalization*, isto é, observa-se a intenção de obtenção de idéias gerais a partir de instâncias.

Apêndice 5: Manual de anotação das estruturas esquemáticas e estratégias retóricas da seção "Resultados"

As orientações abaixo descrevem o esquema de anotação manual das estruturas esquemáticas e estratégias retóricas para a seção "Resultados" de córpus de artigos científicos em inglês.

III. Artigo científico: estrutura

Uma característica comum a praticamente todos os textos científicos, que descrevem pesquisa experimental, é o tipo de organização que sua estrutura esquemática deve seguir. Essa estrutura pode ser apresentada como Introdução, Desenvolvimento e Conclusão, sendo que o Desenvolvimento pode ser subdividido em Materiais e Métodos e Resultados, ou ainda Materiais e Métodos, Resultados e Discussão. O objetivo desse tipo de estruturação é guiar o leitor e fazer com que ele siga, na leitura ou escrita do texto, o movimento do fluxo da informação a ser transmitida que parte do geral-para-específico na Introdução e chega ao específico-para-geral, na Conclusão, conforme pode ser observado na figura abaixo.

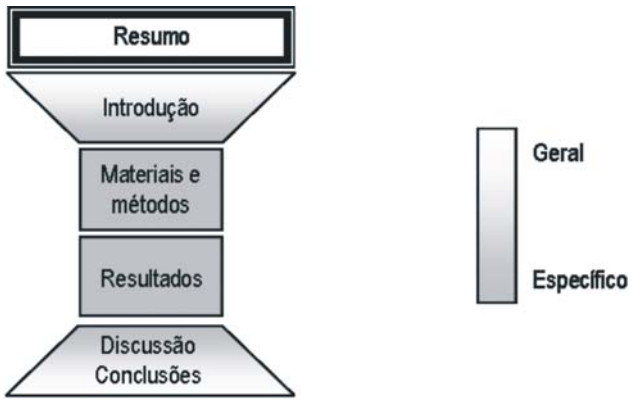


Figura 1: Movimento Geral-Específico-Geral presente na estrutura global do texto científico (Weissberg & Buker, 1990)

IV. O que é a seção "Resultados"?

A seção de Resultados deve conter uma exposição factual sobre o que foi observado, deve ser desenvolvida apoiada nas estatísticas, tabelas e gráficos elaborados no decorrer da análise dos dados, durante a investigação. Os resultados do trabalho devem ser apresentados numa ordem lógica – que pode ser diversa da ordem em que foi desenvolvida a investigação.

Os resultados de quase todas as áreas de estudo são escritos de uma maneira muito similar. Os tipos de informação incluídos e a ordem em que aparecem são muito convencionais, de modo que podem ser enunciados como modelos de resultados. Tais modelos objetivam guiar o escritor no sentido do tipo de informação que deve ser incluída em uma boa seção "resultado" e da ordem que

tais informações devem aparecer. Assim, O quadro 1, contém as informações típicas que, segundo Weissberg & Buker, (1990:186), podem ser encontradas, em geral, em resultados de quaisquer áreas do conhecimento.

Contexto – conhecimento aceito pela comunidade científica **Bibliografia/literatura –** Menção de trabalhos anteriores **Propósito** – principal atividade da pesquisa **Metodologia** – alguma informação sobre a metodologia utilizada **Resultados** – os resultados mais importantes obtidos

Quadro 1: Informações retóricas ou estruturas esquemáticas típicas de resultados.

III Processo de anotação – o que é?

Consiste, nesse caso, na identificação da função retórica de cada sentença do texto utilizando-se para isso siglas representativas desse papel retórico, as quais serão colocadas no início de cada sentença.

IV Antes do processo de anotação - orientações

- Importante ler o texto antes da anotação, uma vez que a interpretação de determinadas sentenças só se torna possível após uma visão geral do texto.
- Não oriente sua leitura para o entendimento da pesquisa relatada, mas sim para o entendimento da estrutura de argumentação construída pelo autor.
- Não anotar o título ou os subtítulos do texto. Utilize-os apenas como dica/ponteiro do conteúdo que se encontra abaixo dos mesmos.

V Durante o processo de anotação - orientações

O processo de anotação (ou classificação) deve ser feito para cada sentença do corpus, que receberá uma classificação (ou categoria). Entretanto, pode haver sentenças que apresentam características de mais de uma categoria, ou seja, sentenças nas quais os papéis argumentativos se sobrepõem, como por exemplo, sentenças que relatam ao mesmo tempo o procedimento e os equipamentos utilizados. Nesses casos, deve-se identificar a sentença, por meio de uma barra entre as categorias possíveis.

Exemplo:

PRO/EQU Cells were subjected to centrifugation in a Ficoll Hypaque density gradient (Amersham Pharmacia) to further purify PMNs. PRO/PAD The PMNs then were removed by filtration, and the supernatants were analyzed by HPLC analysis.

Exemplos retirados do texto Met_02 do córpus Met comporto por seções "Metodologia" de aritigos científicos da área de Ciências Farmacêuticas.

Note ainda que sentenças consecutivas do texto podem receber a mesma classificação. É comum anotar sentenças consecutivas com a mesma categoria, desde que juntas preencham os critérios de uma dada categoria. Por exemplo: é possível marcar mais de uma sentença como MAT (Materiais) se juntas, elas compõem a lista de materiais utilizados na pesquisa, por exemplo.

MAT Whole antibodies 31154 (human IgG) and 31127 (horse IgG) were obtained from PharMingen. MAT Bovine catalase was obtained from Sigma. MAT All assays were carried out in PBS (10 mM phosphate/160 mM sodium chloride, pH 7.4).

MAT Commercial protein solution samples were dialyzed into PBS as necessary. MAT Indigo carmine, isatin sulfonic acid, HOCl, H2O2, vinylbenzoic acid, and 4-carboxybenzaldehyde were obtained from Aldrich.

Exemplo de frases consecutivas com mesma categoria (extraídas do córpus Met):

Se não for possível atribuir nenhuma categoria do esquema utilizado a uma dada sentença, anote-a com um identificador qualquer (por exemplo "?") e descreva, em uma folha a parte, a dificuldade sentida em classificá-la e a função que ela apresenta. Anote também as possíveis dificuldades na classificação de trechos, bem como com a própria categoria que está sendo utilizada. Ex:

? Thiols, at an initial concentration of 1 mM, together with additives as indicated, were equilibrated to 37°C in a shaking water bath.

Importante:

*Qualquer tipo de dúvida é muito importante e deve ser anotada e levada para discussão com o grupo, pois visa a uma melhor caracterização/adequação de uma categoria problemática.

*Não se esqueça de anotar todas as sentenças do texto e suas eventuais dificuldades, que serão discutidas numa reunião com o grupo de anotadores.

*Anote também quais foram os critérios que utilizou para identificar as funções retóricas e, posteriormente, anotar o texto.

VI. Categorias para a anotação retórico-manual do texto

As categorias escolhidas para realizar a anotação retórica dos textos foram inspiradas no modelo de estruturação retórica proposto por Weissberg & Buker, (1990). A seguir, no Quadro 2, temos uma lista de siglas que correspondem as possíveis seções que podem ser encontradas na seção "Resultados" de um artigo experimental. Para cada uma dessas categorias foi elaborada uma sigla que a representasse para facilitar o processo de anotação do córpus.

Siglas das categorias	Descrição das siglas
CON-TOP	Familiarizar termos, objetos e processos
BLI-MTA	Mencionar trabalho anterior do autor
BLI-MTR	Mencionar trabalhos relacionados
BLI-CTA	Comparar trabalho anterior do autor
BLI-COT	Comparar outros trabalhos
PRO-CIP	Citar propósito
MET-CIM	Citar metodologia
RES-TRE	Topicalizar resultados
RES-LRE	Localizar resultados
RES-ARE	Apresentar resultados
RES-DRE	Discussão dos resultados
RES-ERR	Explicar razões dos resultados
RES-ERE	Especular resultados
RES-EED	Exemplificar explicação/discussão

Figura 2: Note, na figura acima, que a sigla da categoria é composta sempre por letras que compõem uma dada categoria, de forma a facilitar a memorização e fácil identificação do significado da categoria que deverá ser empregada nas sentenças do corpus.

3.1. O modelo de Weissberg & Buker (1990)

A análise dos Movimentos, desenvolvida por Weissberg & Buker (1990), visa a representar artigos científicos em termos de sua organização textual hierárquica construída por seções distintas. Cada seção pode ser dividida em estruturas esquemáticas (Contexto, Bibliografia/Literatura, Propósito e Resultados e cada Estrutura Esquemática, por sua vez, pode ser subdividida em Estratégias Retóricas.

A estrutura esquemática consiste em uma unidade textual funcional, utilizada com algum propósito retórico identificável. Essas estruturas realizam-se lingüisticamente em um texto de diferentes formas, ou seja, por diferentes estratégias retóricas, conforme a estrutura esquemática na qual se encontram contidas.

Vale lembrar que essas estruturas visam a descrever de maneira geral as funções retóricas que podem ser encontradas em um córpus. Mas é bem possível que outras funções possam ser encontradas no córpus em análise e acrescentadas a esse modelo. Aliás, esse modelo é um ponto de partida para que seja aprimorado segundo as características que o córpus em estudo/análise apresentar.

3.2 Detalhamento das categorias

Nessa seção, são apresentados as categorias e a forma como essas podem aparecer no texto. A definição das estruturas esquemáticas abaixo apresentadas tiveram suas definições baseadas no trabalho de Feltrim (2004).

As estratégias retóricas apresentadas abaixo são as diferentes maneiras que as estruturas esquemáticas apresentadas na Figura 2 podem ser encontradas.

Estratégia Retórica 1 - Familiarizar termos, objetos e processos – Sigla CON-TOP
Their effects on copper-catalyzed cysteine oxidation have *previously* been considered only in terms of complexing the copper and there is no consistency among published studies [12, 23 and 24].
Copper-catalyzed cysteine oxidation generates hydrogen peroxide which can oxidize *further* cysteine by a mechanism that involves direct reaction of the thiolate anion [11 and 33].
Estratégia Retórica 2 - Mencionar trabalho anterior do autor – Sigla BLI-MTA
We *previously* reported that mice immunized against *this* antigen reject a challenge of P815 cells injected intraperitoneally14.
We showed *previously* that rapid and uniform rejection of allogeneic concepti occurred when pregnant CBA mice were exposed to 20 mg/day of 1-methyl-tryptophan9.
Estratégia Retórica 3 - Mencionar trabalhos relacionados – Sigla BLI-MTR
Other workers have demonstrated the influence of medium composition on stability of protein expression, and the potential of medium re-feeding or *other* treatments to enhance cell-productivity of recombinant protein.
A hybridoma maintained stable antibody productivity in medium containing 5% v/v serum *but* lost production at 1.5% v/v serum (Ozturk and Palsson, 1990), *while* a recombinant NS0 myeloma clone expressing a humanised monoclonal antibody showed marked differences in long term stability of expression depending on the medium used for its cultivation (Castillo et al., 1999).
Estratégia Retórica 4 - Comparar trabalho anterior do autor – Sigla BLI-CTA
Using *this* expression vector, we successfully isolated homodimeric p66 RT [9,17,18].
However, the expression of p66 was not sufficient to routinely prepare very large quantities ofRT for structural studies.
Estratégia Retórica 5 - Comparar outros trabalhos – Sigla BLI-COT
Moreover, previous studies indicate that phagocytosis induces global changes in PMN gene expression (13, 14, 15).
Previous studies have demonstrated that the expression of chemokines and receptors for inflammatory molecules are regulated by cell redox status (26, 27).
Estratégia Retórica 6 - Citar propósito – Sigla PRO-CIP
We *therefore* tested whether 1MT treatment would prevent the growth of IDO-expressing P815B cells injected into P1A-immunized mice.
To identify constitutive differences in gene expression that underlie chronic inflammation in XCGD patients, we compared transcript levels in unstimulated PMNs from XCGD patients and healthy

control individuals (Table I, and supplemental Table II on the Journal of Immunology web site, which contains the complete set of microarray data for *these* experiments).
Estratégia Retórica 7 - Citar metodologia – Sigla MET-CIM
P815 tumor cells regularly produce progressive tumors when injected intraperitoneally into naive syngeneic DBA/2 mice, even though *they* are clearly immunogenic and express several antigens recognized by cytolytic T lymphocytes (CTLs).
For the in vivo experiments reported below, we selected three clones: clone 6, which expresses very high levels of IDO; clone 7, which has IDO activity similar to that of placenta; and clone 1, which was transfected with a control vector and does not express any IDO.
Estratégia Retórica 8 - Topicalizar resultados – Sigla RES-TRE
Inhibition of copper-catalyzed cysteine oxidation by added iron
Iron inhibition in *other* buffers
Estratégia Retórica 9 - Localizar resultados – Sigla RES-LRE
Figure 2 illustrates the staining of some sections, including a non-small-cell lung carcinoma (Fig.2c), where the staining of tumor cells was abolished by blocking with a synthetic peptide corresponding to the IDO C-terminal sequence, *further* demonstrating the specificity of the staining.
Fig.3 D shows that HDL, whether preincubated with SAA or simultaneously added to SAA, completely abolished SAA-induced FPRL1/293 cell migration.
Estratégia Retórica 10 - Apresentar resultados – Sigla RES-ARE
Addition of the selective iron chelator DFO increased the rate of copper-catalyzed cysteine oxidation to that seen in Chelex-treated buffer (Fig. 1).
Both concentrations caused the same enhancement of oxidation, indicating that on a molar basis, less DFO than copper was required for maximum effect.
Estratégia Retórica 11 - Discussão dos resultados – Sigla RES-DRE
If the buffer was pretreated with Chelex resin, and scrupulous attention was paid to avoiding all glass contact and using only new plasticware, the rate of oxidation was much faster than in unpurified buffer (Fig. 1).
However, *this* mode of action seems unlikely *because* it would be expected to give inhibition *rather than* enhanced oxidation.
Estratégia Retórica 12 - Explicar razões dos resultados – Sigla RES-ERR
MC57G fibrosarcoma cell lines (H-2b haplotype) were selected for *this* study *because* they elicit potent H-2Kb-specific T cell responses in vitro and do not express IDO constitutively (data not shown).
After electroporation to introduce rDNA containing CMV promoter elements linked to murine IDO cDNA sequences, we isolated a series of IDO-transfected MC57G clones and screened them for IDO gene transcription, protein expression, and enzyme activity (Fig.1).
Estratégia Retórica 13 - Especular resultados – Sigla RES-ERE
DFO *also* binds copper, *so* it could potentially act by removing the copper from the Cu-cysteine complex involved in the oxidation mechanism.
An alternative explanation for the Chelex effect, that chelating agent became detached from the resin and interacted with the copper, can *also* be excluded *because* the addition of 5 small mu M chelating agent (iminodiacetic acid) had little effect on the rate of cysteine oxidation (not shown).
Estratégia Retórica 14 - Exemplificar explicação/discussão – Sigla RES-EED
For example, genes encoding SFRS protein kinase 1 (SRPK1), lipoma HMGIC fusion partner-like 2 (LHFPL2), CBF1-interacting corepressor (CIR), oligodendrocyte lineage transcription factor 2 (RACK17), zinc finger protein 147, and zinc finger protein 254 were up-regulated only in normal cells (Fig.1).

For example, at a temperature of 38°C, DOT 4 5% and a seeding density of 0.27 × 106 cells, the average specific growth rate was 0.41 d?1 (data shown in Figure 4(d)).

O texto a seguir é uma seção “resultados” que possui sua estrutura esquemática e respectivas estratégia retóricas anotadas. A primeira identificada pela primeira parte da sigla (lado esquerdo do hífen da sigla) e a segunda pelo lado direito da sigla. Trata-se de um resultado de um artigo científico

da área de Ciências Farmacêuticas, retirado da Base de Casos do ambiente SciPo-Farmácia (http://www.nilc.icmc.usp.br/scipo-farmacia/).

Resultados - Caso resul_36

Link: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WPJ-4B8BMMK-2&_user=972067&_handle=B-WA-A-A-WU-MsSAYZA-UUW-AUEUDZBZWZ-AUYYBVVVWZ-VYDYYYVZW-WU-U&_fmt=full&_coverDate=03%2F31%2F2004&_rdoc=12&_orig=browse&_srch=%23toc%236992%232004%23999659998%23477805!&_cdi=6992&view=c&_acct=C000049650&_version=1&_urlVersion=0&_userid=972067&md5=9384091840bebf54083cd3ff1243fc3f

Development of a large-scale HPLC-based purification for the urease from Staphylococcus leei and determination of subunit structure

MING JIN, WILDYS ROSARIO, ELSIE WATLER, AND DAVID H. CALHOUN

RES-TER Cell lysis and clarification of the crude extract

RES-ARE We tested a variety of lysis techniques and found that S.leei is resistant to many routine cell lysis procedures, including sonication, a French Pressure Cell at 36,000 psi, a Niro homogenizer operating at a maximum pressure of 15,000 psi, or a Gaulin mill at 8000 psi.

RES-DRE *However*, we *finally* were successful with three cycles in a Dynomill with glass beads, which resulted in lysis of about 95% of the cells.

RES-DRE Examination of the cells microscopically indicated that each passage broke about a third of the cells.

RES-DRE *This* procedure is suitable for scale up and was used with cells from two 60-liter fermentation batches.

RES-TRE Purification of the urease of S. leei

RES-DRE The slightly turbid crude extract was clarified by centrifugation, cell debris was removed by pumping through a 0.2 lm hollow fiber filter, and the filtrate was *then* concentrated and the buffer was changed by dia- filtration using 10kDa pore size hollow fiber filter.

RES-DRE The retentate was used to purify the urease by sequential chromatography on Q-Sepharose, Poros HP2, Sephacryl S-300, and hydroxyapatite (Fig. 1, Table 1).

RES-ARE The procedure resulted in 98-fold purification with an 18% yield for *this* aliquot of the total crude extract.

RES-ARE The total urease protein obtained from the 86 ml aliquot of the crude extract was 0.42 mg (Table 1).

RES-ARE *Subsequently*, the 3.7-liter crude extract from 120-liter of cells yielded 16 mg of enzyme.

RES-ARE The purified enzyme is composed of three distinct subunits that by analogy to related microbial species *such as* S.xylosus are designated (Fig. 1) a, b, and c, with molecular weights of 65, 21, and 12 kDa, respectively.

RES-ARE *These* three subunits correspond to the products of the ureABC genes of S.leei (Lin et al., in preparation).

RES-DRE To confirm the identity of the urease, partial < Fig.1 > Table 1 amino acid sequence was determined by mass spectrometry and the amino acid sequence EPGDEKEVQLVEY was obtained and found to be 100% identical to a segment predicted by the ureB gene of S. leei.

RES-ARE Native molecular weight determination and in situ enzyme assay A Sephacryl S-300 molecular sieve column that separates proteins in the range of 10-1500 kDa was calibrated with proteins from 158 to 669 kDa (Fig. 2, upper).

RES-ARE The urease catalytic activity (Fig. 2, lower) and absorbance at 280nm (data not shown) eluted as a uniform peak corresponding to approximately 480 kDa.

RES-DRE The urease of Staphylococcus saprophyticus has subunits of 72.4, 20.4, and 13.9 kDa and a native molecular weight of 420kDa and was proposed to have an ðabcÞ4 stoichiometry [23].

RES-DRE The estimated 480 kDa molecular weight of the S.leei urease is consistent with an (abcÞ5 structure with a calculated molecular weight of 490kDa for the a (65 kDa), b (21 kDa), and c (12 kDa) subunits.

RES-DRE The crystal structure of the Klebsiella aerogenes urease [12] clearly indicates a 1:1:1 ratio of the abc subunits and it was proposed [20] on the basis of extensive sequence similarities among urease proteins that all ureases have equal numbers of each of their distinct subunits.
RES-DRE An in situ enzyme assay (Fig. 3) using the urease at various stages of purification (Fig. 1, Table 1) reveals slow and fast moving bands indicating heterogeneity in enzyme structure in a nondenaturing gel, similar to the enzyme from of S. saprophyticus [23].
RES-DRE Electrophoretic mobility in *this* type of gel depends on molecular weight, subunit aggregation, charge, or association with *other* proteins for the partially purified forms.
RES-DRE The observation that the protein elutes from the Sephacryl S-300 column as a single symmetrical peak without shoulders (Fig. 2, lower) indicates that, under *these* conditions, the enzyme is present as a single molecular weight species.
The two forms present for the in situ enzyme assay (Fig. 3) could represent two molecular weight forms < Fig.2 > < Fig.4 > present under the conditions of SDS gel electrophoresis.
RES-ERE *Alternatively*, the two forms could differ in charge, with the more electropositive form migrating faster.
RES-DRE The Km of the urease for urea is 1.66mM (Fig. 4) which is similar to that of H. pylori (0.3mM [4]) and lower than that of *other* Staphylococcus species (*e.g.*, 9.5mM for S. saprophyticus [23]).

Apêndice 6: Manual de anotação das estruturas esquemáticas e estratégias retóricas da seção "Discussão"

As orientações abaixo descrevem o esquema de anotação manual das estruturas esquemáticas e estratégias retóricas para a seção "Discussão" de córpus de artigos científicos em inglês.

V. Artigo científico: estrutura

Uma característica comum a praticamente todos os textos científicos, que descrevem pesquisa experimental, é o tipo de organização que sua estrutura esquemática deve seguir. Essa estrutura pode ser apresentada como Introdução, Desenvolvimento e Conclusão, sendo que o Desenvolvimento pode ser subdividido em Materiais e Métodos e Resultados, ou ainda Materiais e Métodos, Resultados e Discussão. O objetivo desse tipo de estruturação é guiar o leitor e fazer com que ele siga, na leitura ou escrita do texto, o movimento do fluxo da informação a ser transmitida que parte do geral-para-específico na Introdução e chega ao específico-para-geral, na Conclusão, conforme pode ser observado na figura abaixo.

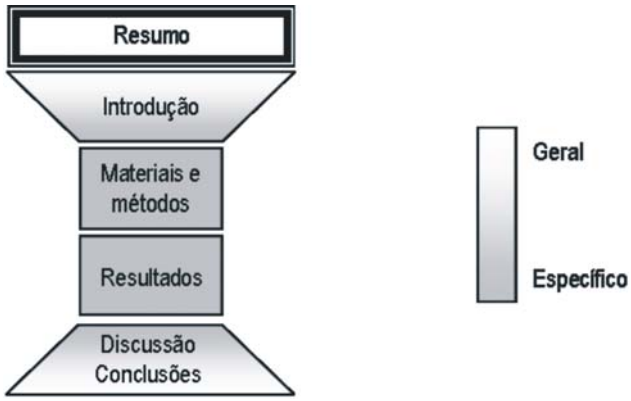


Figura 1: Movimento Geral-Específico-Geral presente na estrutura global do texto científico (Weissberg & Buker, 1990)

VI. O que é a seção "Discussão"?

Nas Discussões, os comentários sobre os resultados aparecem de forma mais densa e se faz muito mais relações do que foi encontrado com a área de pesquisa, pois compara estudos já realizados sobre o assunto, questiona trabalhos de outros autores, compara trabalhos anteriores de um mesmo autor, além de apresentar implicações e limitações da pesquisa realizada. O movimento das Discussões é o contrário do das Introduções, ou seja, parte de informações específicas (no caso os resultados da pesquisa realizada) e, com os comentários, passa a interpretar os dados obtidos na pesquisa no cenário da área científica que o trabalho se insere. Devido ao seu caráter argumentativo, a parte formal da Discussão é menos importante do que seu conteúdo. Por conta disso, ao escrever uma Discussão, o pesquisador deve se preocupar em identificar as metas mais relevantes da pesquisa, ressaltar o resultado principal, apresentar especulações sobre os resultados obtidos, discutir os resultados consistentes e inconsistentes com a literatura, etc.

A seção "Discussão" de quase todas as áreas de estudo é escrita de uma maneira muito similar. Os tipos de informação incluídos e a ordem em que aparecem são muito convencionais, de modo que podem ser enunciados como modelos de Discussão. Tais modelos objetivam guiar o escritor no sentido do tipo de informação que deve ser incluída em uma boa discussão e da ordem que tais informações devem aparecer. Assim, O quadro 1, contém as informações típicas que, segundo Weissberg & Buker, (1990:186), podem ser encontradas, em geral, na seção "Discussão" de quaisquer área do conhecimento.

Retomar Contexto – retomada de informações que contextualizam a pesquisa que foi realizada
Relacionar a pesquisa com a literatura – informações que remetem aos estudos realizados sobre o que foi realizado
Retomar Propósito – retomada do objetivo da pesquisa
Revisar Resultados mais importantes – os resultados obtidos na pesquisa são retomados para, a partir deles, ser construída a discussão
Retomar Metodologia – retomada dos métodos utilizados na pesquisa
Apresentar Conclusões – discussão sobre as interpretações do pesquisador a respeito dos resultados obtidos

Quadro 1: Informações retóricas ou estruturas esquemáticas típicas de discussão.

III Processo de anotação – o que é?

Consiste, nesse caso, na identificação da função retórica de cada sentença do texto utilizando-se para isso siglas representativas desse papel retórico, as quais serão colocadas no início de cada sentença.

IV Antes do processo de anotação - orientações

- Importante ler o texto antes da anotação, uma vez que a interpretação de determinadas sentenças só se torna possível após uma visão geral do texto.
- Não oriente sua leitura para o entendimento da pesquisa relatada, mas sim para o entendimento da estrutura de argumentação construída pelo autor.
- Não anotar o título ou os subtítulos do texto. Utilize-os apenas como dica/ponteiro do conteúdo que se encontra abaixo dos mesmos.

V Durante o processo de anotação - orientações

O processo de anotação (ou classificação) deve ser feito para cada sentença do corpus, que receberá uma classificação (ou categoria). Entretanto, pode haver sentenças que apresentam características de mais de uma categoria, ou seja, sentenças nas quais os papéis argumentativos se sobrepõem, como por exemplo, sentenças que relatam ao mesmo tempo o procedimento e os equipamentos utilizados. Nesses casos, deve-se identificar a sentença, por meio de uma barra entre as categorias possíveis.

Exemplo:

PRO/EQU Cells were subjected to centrifugation in a Ficoll Hypaque density gradient (Amersham Pharmacia) to further purify PMNs. PRO/PAD The PMNs then were removed by filtration, and the supernatants were analyzed by HPLC analysis.

Exemplos retirados do texto Met_02 do córpus Met comporto por seções "Metodologia" de aritigos científicos da área de Ciências Farmacêuticas.

Note ainda que sentenças consecutivas do texto podem receber a mesma classificação. É comum anotar sentenças consecutivas com a mesma categoria, desde que juntas preencham os critérios de uma dada categoria. Por exemplo: é possível marcar mais de uma sentença como MAT (Materiais) se juntas, elas compõem a lista de materiais utilizados na pesquisa, por exemplo.

MAT Whole antibodies 31154 (human IgG) and 31127 (horse IgG) were obtained from PharMingen. MAT Bovine catalase was obtained from Sigma. MAT All assays were carried out in PBS (10 mM phosphate/160 mM sodium chloride, pH 7.4). MAT Commercial protein solution samples were dialyzed into PBS as necessary. MAT Indigo carmine, isatin sulfonic acid, HOCl, H2O2, vinylbenzoic acid, and 4-carboxybenzaldehyde were obtained from Aldrich.

Exemplo de frases consecutivas com mesma categoria (extraídas do córpus Met):

Se não for possível atribuir nenhuma categoria do esquema utilizado a uma dada sentença, anote-a com um identificador qualquer (por exemplo "?") e descreva, em uma folha a parte, a dificuldade sentida em classificá-la e a função que ela apresenta. Anote também as possíveis dificuldades na classificação de trechos, bem como com a própria categoria que está sendo utilizada. Ex:

? Thiols, at an initial concentration of 1 mM, together with additives as indicated, were equilibrated to 37°C in a shaking water bath.

Importante:

*Qualquer tipo de dúvida é muito importante e deve ser anotada e levada para discussão com o grupo, pois visa a uma melhor caracterização/adequação de uma categoria problemática.

*Não se esqueça de anotar todas as sentenças do texto e suas eventuais dificuldades, que serão discutidas numa reunião com o grupo de anotadores.

*Anote também quais foram os critérios que utilizou para identificar as funções retóricas e, posteriormente, anotar o texto.

VI. Categorias para a anotação retórico-manual do texto

As categorias escolhidas para realizar a anotação retórica dos textos foram inspiradas no modelo de estruturação retórica proposto por Weissberg & Buker, (1990). A seguir, no Quadro 2, temos uma lista de siglas que correspondem as possíveis seções que podem ser encontradas na seção "Discussão" de um artigo experimental. Para cada uma dessas categorias foi elaborada uma sigla que a representasse para facilitar o processo de anotação do córpus.

Siglas das categorias	Descrição das siglas
RCO-PTP	Indicar proeminência do tópico para a área de pesquisa
RCO-TPC	Familiarizar termos, processos, conceitos
RPL-RP	Resumir a pesquisa
RPL-MTA	Mencionar trabalho anterior do autor
RPL-CPA	Comparar com pesquisa(s) anterior(es) do autor
RPL-PAA	Comparar com pesquisa(s) anterior(es) de outros autores
RPL-DL	Discutir a literatura
RPR-RPH	Recuperar propósito ou hipótese
RRI-PDP	Apresentar ponto mais dramático da pesquisa
RRI-AED	Apresentar especulações ou deduções
RRI-AR	Apresentar resultados
RRI-DR	Discutir resultados
RRI-ARI	Apresentar resultado(s) inesperado(s)
RM-CM	Comentar metodologia
ACO-ALP	Apresentar limitações da pesquisa
ACO-AIP	Apresentar implicações da pesquisa
ACO-ATF	Apresentar trabalhos futuros
ACO-MF	Mencionar financiadores
ACO-AC	Agradecer colaborações

Figura 2: Note, na figura acima, que a sigla da categoria é composta sempre por letras que compõem uma dada categoria, de forma a facilitar a memorização e fácil identificação do significado da categoria que deverá ser empregada nas sentenças do corpus.

3.1. O modelo de Weissberg & Buker (1990)

A análise dos Movimentos, desenvolvida por Weissberg & Buker (1990), visa a representar artigos científicos em termos de sua organização textual hierárquica construída por seções distintas. Cada seção pode ser dividida em estruturas esquemáticas, como as que podem ser encontradas na seção "Discussão" (Retomar Contexto, Relacionar a pesquisa com a literatura, Retomar Propósito, Revisar Resultados mais importantes, Retomar Metodologia, Apresentar Conclusões) que por sua vez, pode ser subdividida em Estratégias Retóricas.

A estrutura esquemática consiste em uma unidade textual funcional, utilizada com algum propósito retórico identificável. Essas estruturas realizam-se lingüisticamente em um texto de diferentes formas, ou seja, por diferentes estratégias retóricas, conforme a estrutura esquemática na qual se encontram contidas.

Vale lembrar que essas estruturas visam a descrever de maneira geral as funções retóricas que podem ser encontradas em um córpus. Mas é bem possível que outras funções possam ser encontradas

no córpus em análise e acrescentadas a esse modelo. Aliás, esse modelo é um ponto de partida para que seja aprimorado segundo as características que o córpus em estudo/análise apresentar.

3.2 Detalhamento das categorias

Nessa seção, são apresentados as categorias e a forma como essas podem aparecer no texto. A definição das estruturas esquemáticas abaixo apresentadas tiveram suas definições baseadas no trabalho de Feltrim (2004).

As estratégias retóricas apresentadas abaixo são as diferentes maneiras que as estruturas esquemáticas apresentadas na Figura 2 podem ser encontradas.

Estratégia Retórica 1 - Indicar proeminência do tópico para a área de pesquisa – Sigla RCO-PTP

Direct EPR detection of the carbonate radical anion is important to unravel mechanistic details of oxidative damage inflicted not only by peroxynitrite *but* also by *other* oxidizing species *such as* hydrogen peroxide and the hydroxyl radical.

Estratégia Retórica 2 - Familiarizar termos, processos, conceitos – Sigla RCO-TPC
It is essential that the volume of the vacuole be restricted for the requisite hypertonicity to develop.
The hazard of injurious effects from *these* enzymes in normal tissues is reduced by packaging them in an inactive form adsorbed on acidic proteoglycans, from which *they* are released and activated only by a combination of the unusual conditions of hypertonicity and alkalinity prevailing in the phagocytic vacuole.
Estratégia Retórica 3 - Resumir a pesquisa – Sigla RPL-RP
Our results have uncovered a *previously* unsuspected mechanism of antimicrobial activity in the phagocyte.
In brief, the O-2-generating system causes an influx of K+ into the phagocytic vacuole with an attendant rise in pH to the optimal level for the activity of the granule proteases.
Estratégia Retórica 4 - Mencionar trabalho anterior do autor – Sigla RPL-MTA
In an earlier study, we demonstrated by several methods that p67PHOX is able to bind directly to cytochrome b558 (10).
A direct interaction between p67PHOX and cytochrome b558 is in accord with the idea that p67PHOX regulates the transfer of electrons from NADPH to the flavin (18) *because* p67PHOX *then* would be in proximity to the flavin center, enabling it to perform a regulatory function in *this* part of the protein.
Estratégia Retórica 5 – Comparar com pesquisa(s) anterior(es) do autor – Sigla RPL-CPA
As reported by the previous study (Dong et al., 1995), gratuitous overproduction of LacZ reaching 30% total cell protein could result in cellular ribosome destruction.
Taken together, it could lead to an argument that the cell growth impairment might be likely attributed to the breakdown of cellular ribosome or/and detrimental overload of plasmid DNAs.
Estratégia Retórica 6 – Comparar com pesquisa(s) anterior(es) de outros autores – Sigla RPL-PAA
In conclusion , our data establish a direct interaction between p67PHOX and cytochrome b558, as demonstrated *previously* .
These outcomes with IDO-transfected cells recapitulate previous data showing that human macrophages expressing IDO blocked T cell cycle progression (7).
Estratégia Retórica 7 – Discutir a literatura – Sigla RPL-DL
One suggestion has been that gp91phox, the flavocytochrome b of the NADPH oxidase, is itself the channel35, 36, *although* contradictory evidence exists37, 38.
Early theories implicating oxygen radicals in tissue damage39 stemmed from the apparent toxicity of *these* agents against microbes, which are much tougher than human cells.
Estratégia Retórica 8 – Recuperar propósito ou hipótese – Sigla RPR-RPH

In accord with *this* hypothesis, we report in *this* work that the induction of proinflammatory cytokines by whole GBS cell walls, *as well as* secreted streptococcal products *such as* GBS-F, is entirely dependent on MyD88.
In *this* paper we have described a convenient procedure to prepare AOX protein from thermogenic A.maculatum spadices.
Estratégia Retórica 9 – Apresentar ponto mais dramático da pesquisa – Sigla RRI-PDP
Nevertheless, the data we report *here* are the first direct test of the hypothesis that genetic manipulations to enhance IDO expression in APCs lead to inhibition of T cell responses.
In summary, our results represent the first detection of the carbonate radical anion in aqueous solutions at physiological pHs.
Estratégia Retórica 10 – Apresentar especulações ou deduções – Sigla RRI-AED
One might question the need for such an elaborate activation system.
A possible explanation lies in the very large numbers of neutrophils that infiltrate sites of acute inflammation and the potential of their enzymes to damage autologous tissues if released from cells in a freely soluble, active, form.
Estratégia Retórica 11 – Apresentar resultados – Sigla RRI-AR
In preliminary studies we have found cathepsin G to be very sensitive to oxidation by H2O2 and to be inactivated at a greatly increased rate in phagocytosing neutrophils treated with azide to inhibit MPO.
We show that IDO-transfected tumor cells and tissue microenvironments with enhanced IDO activity inhibited T cell proliferation and reduced the number of T cells elicited over time.
Estratégia Retórica 12 – Discutir resultados – Sigla RRI-DR
Our demonstration that ROS generation and MPO activity are not themselves sufficient to kill key model target organisms is important, not only *because* of the insight it affords into normal immunity, *but* also *because* of the light it throws on pathological mechanisms.
Experiments with MPO were generally performed at what has been shown *here* to be unphysiologically low concentrations of enzyme and H2O2 and at too low a pH (ref.7).
Estratégia Retórica 13 – Apresentar resultado(s) inesperado(s) – Sigla RRI-ARI
In contrast to expectations, we found *here* that the deletion of TLR2 did not *significantly* alter the cellular response to whole GBS as compared with normal cells, suggesting a lesser role of peptidoglycan in streptococcal pathogenesis than might otherwise have been predicted.
It is difficult to assess whether *this* figure indicates that 80% of the protein had its iron chelated or, perhaps more likely, that the protein sample is not homogeneous *but* comprises a mixture of enzymes containing *either* zero, one or two iron atoms.
Estratégia Retórica 14 – Comentar metodologia – Sigla RM-CM
In *this* process, the reducing equivalents are generated by conversion of about one-third of the fructose to lactic acid and acetic acid.
The fermentation time decreased considerably from 136 to 92 h by using the fed-batch approach.
Estratégia Retórica 15 – Apresentar limitações da pesquisa – Sigla ACO-ALP
Currently we do not know whether GBS-F is secreted and immunologically relevant in vivo.
It is not clear why acetic acid was not produced.
Estratégia Retórica 16 - Apresentar implicações da pesquisa – Sigla ACO-AIP
Similar mechanisms of oxidative inactivation of degradative enzymes could explain the accelerated deposition of atheromatous material observed in MPO-deficient mice40, 41.
Detection and characterization of the radical as negatively charged at neutral pHs should contribute to the understanding of the roles of ubiquitous carbon dioxide in modulating the pathogenic mechanisms of peroxynitrite and *other* oxidizing intermediates.
Estratégia Retórica 17 – Apresentar trabalhos futuros – Sigla ACO-ATF
Further studies are required to elucidate the mechanism of enhanced reactivity.
Whether their sensitivity to oxidation can be modulated by interactions with substrates, or *other* protein components of the apoptotic machinery, warrants *further* investigation.
Estratégia Retórica 18 – Mencionar financiadores – Sigla ACO-MF
Estratégia Retórica 19 – Agradecer colaborações – Sigla ACO-AC

O texto a seguir é uma seção "Discussão" que possui sua estrutura esquemática e respectivas estratégia retóricas anotadas. A primeira identificada pela primeira parte da sigla (lado esquerdo do hífen da sigla) e a segunda pelo lado direito da sigla. Trata-se de um resultado de um artigo científico

da área de Ciências Farmacêuticas, retirado da Base de Casos do ambiente SciPo-Farmácia (http://www.nilc.icmc.usp.br/scipo-farmacia/).

Discussão - Caso disc_34

Link: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T1S-4BDC557-3&_user=972067&_handle=B-WA-A-A-AZ-MsSAYVW-UUW-AUEUDCZEZA-AUYYBBDDZA-VYDDYBAUV-AZ-U&_fmt=full&_coverDate=02%2F15%2F2004&_rdoc=12&_orig=browse&_srch=%23toc%234898%232004%23983919997%23478073!&_cdi=4898&view=c&_acct=C000049650&_version=1&_urlVersion=0&_userid=972067&md5=e7874ad2231196e44c79c931d227f1b6

Purification of the plant alternative oxidase from Arum maculatum: measurement, stability and metal requirement

CHARLES AFFOURTIT, ANTHONY L. MOORE

RPR-RPH In *this* paper we have described a convenient procedure to prepare AOX protein from thermogenic A.maculatum spadices.

RRI-AR The obtained sample is virtually pure, exhibits a high specific AOX activity and is exceptionally stable.

RRI-DR From the data shown in Fig.3A, it can be concluded that activity of *this* AOX sample is severely inhibited by the metalchelating agent 8-hydroxyquinoline.

RRI-DR *This* inhibited activity can be fully restored by ferric iron (Fig.3B), *but* not by ferrous iron, manganese or zinc.

RPL-PAA *These* observations support the notion that iron is essential for AOX catalysis and agree with the current belief that the enzyme's active site comprises a non-haem diiron centre [8–10].

RPL-DL To our knowledge, metal requirement experiments have not been performed *previously* with purified AOX protein or *indeed* with any system containing the plant AOX.

RPL-PAA *However*, our results may be compared with those obtained from studies into the homologous expression of the yeast AOX in P. anomala [11] and the heterologous expression of the trypanosome AOX in E.coli [6].

RPL-PAA *These* studies revealed that the presence of o-phenanthroline in the respective growth media results in the expression of inactive AOX protein.

RPL-PAA In E.coli, an active enzyme could be obtained when surplus ferrous iron was present *in addition* to the metal chelator when expression was induced [6], whilst the inactive AOX protein in P.anomala could be rendered functional by subsequent addition of ferrous, *but* not ferric iron [11].

RPL-PAA *These* observations seem to be in conflict with our findings.

RPL-DL It should be noted, *however*, that ferrous iron is very susceptible to autoxidation [36] and that it is *therefore* difficult to interpret data obtained from experiments involving Fe2 + that were performed under aerobic conditions at neutral pH.

RPL-PAA It is conceivable too that the apparent discrepancy is *due to* differences in experimental design, since our data were obtained using a sample that differs considerably from that of Minagawa et al.[11] and Ajayi et al.[6] regarding *both* its nature and physical state.

RRI-AED The observed unusual dose dependency of the extent to which 8-hydroxyquinoline-inhibited AOX activity is restored by ferric iron (Fig.3B) may be explained by potential heterogeneity of the sample.

RRI-AR Prior to the addition of iron, the sample exhibited approximately 20% of its 'non-chelatorinhibited' activity.

RRI-ARI It is difficult to assess whether *this* figure indicates that 80% of the protein had its iron chelated or, perhaps more likely, that the protein sample is not homogeneous *but* comprises a mixture of enzymes containing *either* zero, one or two iron atoms.

RRI-ARI Clearly, it is difficult to predict quantitatively the stimulatory effect of iron on such a heterogeneous mixture.

RRI-DR It should *also* be noted that interpretation of the trend in the data shown in Fig.3B is *further* complicated by experimental variation that is mainly *due to* a protein-independent O2-uptake rate caused by the combination of ferric iron and DQH2, which becomes more dominant with increasing metal concentrations (cf.legend to Fig.3B).

RRI-DR Irrespective of the shape of the reconstitution curve, *however*, it can be concluded that the plant AOX requires iron for activity.
RRI-DR *Finally*, a few comments should be made with respect to the minimum requirements for AOX activity.
RPL-PAA In agreement with previous work from Zhang et al.[20] and Hoefnagel et al.[35], we have shown that AOX activity is substantially increased by pyruvate and the detergent EDT-20.
RRI-AED *Interestingly*, it appears from our experiments that *either* compound alone does not affect the AOX *significantly*, *but* that only their combined presence results in a f5 times increase in activity.
RRI-AED *This* may indicate that the site at which pyruvate interacts with the AOX is obscured in the purified protein and only becomes accessible upon a detergent-induced conformational change.
ACO-ATF Exploiting the reliable spectrophotometric assay described in *this* paper, we are *currently* investigating the regulation of AOX by pyruvate and *other* organic acids in *further* detail.
ACO-ATF *Furthermore*, the effect of hydrogen peroxide on AOX catalysis is being studied *at present*.
ACO-AIP We anticipate that the outcomes of such studies will not only be of mechanistic relevance, *but* will *also* provide valuable information as to the in vivo regulation and role of the plant AOX.

Apêndice 7: Manual de anotação das estruturas esquemáticas e estratégias retóricas da seção "Conclusão"

As orientações abaixo descrevem o esquema de anotação manual das estruturas esquemáticas e estratégias retóricas para a seção "Conclusão" de córpus de artigos científicos em inglês.

VII. Artigo científico: estrutura

Uma característica comum a praticamente todos os textos científicos, que descrevem pesquisa experimental, é o tipo de organização que sua estrutura esquemática deve seguir. Essa estrutura pode ser apresentada como Introdução, Desenvolvimento e Conclusão, sendo que o Desenvolvimento pode ser subdividido em Materiais e Métodos e Resultados, ou ainda Materiais e Métodos, Resultados e Discussão. O objetivo desse tipo de estruturação é guiar o leitor e fazer com que ele siga, na leitura ou escrita do texto, o movimento do fluxo da informação a ser transmitida que parte do geral-para-

específico na Introdução e chega ao específico-para-geral, na Conclusão, conforme pode ser observado na figura abaixo.

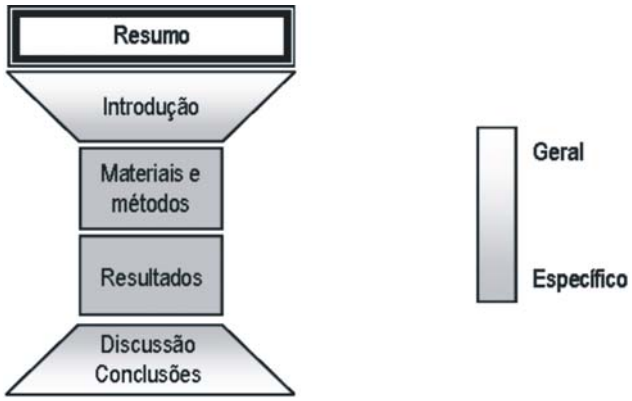


Figura 1: Movimento Geral-Específico-Geral presente na estrutura global do texto científico (Weissberg & Buker, 1990)

VIII. O que é a seção "Conclusão"?

A seção Conclusão é a última do texto e dessa forma deve tem a função de finalizar o assunto. Durante a conclusão do trabalho, a discussão deve ser uma consideração objetiva dos resultados apresentados na seção anterior e deve conduzir com naturalidade às suas principais conclusões. Deve-se fazer referência a qualquer esclarecimento adicional sobre os problemas levantados na seção Introdução e dizer como o trabalho se enquadra no conjunto das investigações precedentes. A conclusão deve proporcionar um resumo sintético, porém completo, da argumentação, das provas e os exemplos abordados nas duas primeiras partes do trabalho. Deve relacionar, em primeiro lugar, as diversas partes da argumentação, unir as idéias desenvolvidas. É por isso que se diz que, em certo sentido, a conclusão é uma volta à introdução. Além de desempenhar o papel de fecho de uma demonstração, a conclusão pode – e 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.

As Conclusões de quase todas as áreas de estudo são escritas de uma maneira muito similar. Os tipos de informação incluídos e a ordem em que aparecem são muito convencionais, de modo que podem ser enunciados como modelos de conclusões. Tais modelos objetivam guiar o escritor no sentido do tipo de informação que deve ser incluída em uma boa seção "resultado" e da ordem que tais informações devem aparecer. Assim, O quadro 1, contém as informações típicas que, segundo Weissberg & Buker, (1990:186), podem ser encontradas, em geral, em conclusões de quaisquer áreas do conhecimento.

Contexto	Familiarizar termos, objetos e processos
Propósito	Apresentar o propósito principal
Metodologia	Descrição dos materiais e métodos utilizados no experimento
Resultado	Menção dos resultados como gancho para a conclusão que fechará o texto.
Conclusão	Situação dos resultados encontrados dentro de um cenário mais amplo

Quadro 1: Informações retóricas ou estruturas esquemáticas típicas de conclusões.

III Processo de anotação – o que é?

Consiste, nesse caso, na identificação da função retórica de cada sentença do texto utilizando-se para isso siglas representativas desse papel retórico, as quais serão colocadas no início de cada sentença.

IV Antes do processo de anotação - orientações

- Importante ler o texto antes da anotação, uma vez que a interpretação de determinadas sentenças só se torna possível após uma visão geral do texto.
- Não oriente sua leitura para o entendimento da pesquisa relatada, mas sim para o entendimento da estrutura de argumentação construída pelo autor.
- Não anotar o título ou os subtítulos do texto. Utilize-os apenas como dica/ponteiro do conteúdo que se encontra abaixo dos mesmos.

V Durante o processo de anotação - orientações

O processo de anotação (ou classificação) deve ser feito para cada sentença do corpus, que receberá uma classificação (ou categoria). Entretanto, pode haver sentenças que apresentam características de mais de uma categoria, ou seja, sentenças nas quais os papéis argumentativos se sobrepõem, como por exemplo, sentenças que relatam ao mesmo tempo o procedimento e os equipamentos utilizados. Nesses casos, deve-se identificar a sentença, por meio de uma barra entre as categorias possíveis.

Exemplo:

PRO/EQU Cells were subjected to centrifugation in a Ficoll Hypaque density gradient (Amersham Pharmacia) to further purify PMNs. PRO/PAD The PMNs then were removed by filtration, and the supernatants were analyzed by HPLC analysis.

Exemplos retirados do texto Met_02 do córpus Met comporto por seções "Metodologia" de aritigos científicos da área de Ciências Farmacêuticas.

Note ainda que sentenças consecutivas do texto podem receber a mesma classificação. É comum anotar sentenças consecutivas com a mesma categoria, desde que juntas preencham os critérios de uma dada categoria. Por exemplo: é possível marcar mais de uma sentença como MAT (Materiais) se juntas, elas compõem a lista de materiais utilizados na pesquisa, por exemplo.

MAT Whole antibodies 31154 (human IgG) and 31127 (horse IgG) were obtained from PharMingen. MAT Bovine catalase was obtained from Sigma. MAT All assays were carried out in PBS (10 mM phosphate/160 mM sodium chloride, pH 7.4). MAT Commercial protein solution samples were dialyzed into PBS as necessary. MAT Indigo carmine, isatin sulfonic acid, HOCl, H2O2, vinylbenzoic acid, and 4-carboxybenzaldehyde were obtained from Aldrich.

Exemplo de frases consecutivas com mesma categoria (extraídas do córpus Met):

Se não for possível atribuir nenhuma categoria do esquema utilizado a uma dada sentença, anote-a com um identificador qualquer (por exemplo "?") e descreva, em uma folha a parte, a dificuldade sentida em classificá-la e a função que ela apresenta. Anote também as possíveis dificuldades na classificação de trechos, bem como com a própria categoria que está sendo utilizada. Ex:

? Thiols, at an initial concentration of 1 mM, together with additives as indicated, were equilibrated to 37°C in a shaking water bath.

Importante:

*Qualquer tipo de dúvida é muito importante e deve ser anotada e levada para discussão com o grupo, pois visa a uma melhor caracterização/adequação de uma categoria problemática.

*Não se esqueça de anotar todas as sentenças do texto e suas eventuais dificuldades, que serão discutidas numa reunião com o grupo de anotadores.

*Anote também quais foram os critérios que utilizou para identificar as funções retóricas e, posteriormente, anotar o texto.

VI. Categorias para a anotação retórico-manual do texto

As categorias escolhidas para realizar a anotação retórica dos textos foram inspiradas no modelo de estruturação retórica proposto por Weissberg & Buker, (1990). A seguir, no Quadro 2, temos uma lista de siglas que correspondem as possíveis seções que podem ser encontradas na seção "Conclusão" de um artigo experimental. Para cada uma dessas categorias foi elaborada uma sigla que a representasse para facilitar o processo de anotação do córpus.

Siglas das categorias	Descrição das siglas
COT-TOP	Familiarizar termos, objetos e processos
PRO-APP	Apresentar o propósito principal
MET-DMM	Citar/Descrever materiais e métodos
RES-DR	Descrever os resultados
RES-EER	Explicações/Especulações dos resultados
CON-ALP	Apresentar limitações da pesquisa
COM-AIP	Apresentar implicações da pesquisa
COM-AR	Apresentar recomendações
COM-TAA	Citações de trabalhos anteriores do autor
COM-TA	Citações de trabalhos anteriores
COM-ACP	Apresentar contribuições/valor da pesquisa

Figura 2: Note, na figura acima, que a sigla da categoria é composta sempre por letras que compõem uma dada categoria, de forma a facilitar a memorização e fácil identificação do significado da categoria que deverá ser empregada nas sentenças do corpus.

3.1. O modelo de Weissberg & Buker (1990)

A análise dos Movimentos, desenvolvida por Weissberg & Buker (1990), visa a representar artigos científicos em termos de sua organização textual hierárquica construída por seções distintas. Cada seção pode ser dividida em estruturas esquemáticas (Contexto, Propósito, Metodologia, Resultado e Conclusão, que por sua vez, pode ser subdividida em Estratégias Retóricas.

A estrutura esquemática consiste em uma unidade textual funcional, utilizada com algum propósito retórico identificável. Essas estruturas realizam-se lingüisticamente em um texto de diferentes formas, ou seja, por diferentes estratégias retóricas, conforme a estrutura esquemática na qual se encontram contidas.

Vale lembrar que essas estruturas visam a descrever de maneira geral as funções retóricas que podem ser encontradas em um córpus. Mas é bem possível que outras funções possam ser encontradas no córpus em análise e acrescentadas a esse modelo. Aliás, esse modelo é um ponto de partida para que seja aprimorado segundo as características que o córpus em estudo/análise apresentar.

3.2 Detalhamento das categorias

Nessa seção, são apresentados as categorias e a forma como essas podem aparecer no texto. A definição das estruturas esquemáticas abaixo apresentadas tiveram suas definições baseadas no trabalho de Feltrim (2004).

Estratégia Retórica 1 - Familiarizar termos, objetos e processos - Sigla COT-TOP

Productivity in many fungal fermentations is detrimentally affected by high broth viscosity, which leads to reduced oxygen mass transfer

Estratégia Retórica 2 - Apresentar o propósito principal - Sigla PRO-APP

In conclusion , we have detected chlorotyrosine in TA proteins and demonstrated that it is present in significantly higher amounts in preterm infants with respiratory distress than in control infants.

The study showed that Saccharomyces cerevisiae ATCC 36858 was able to produce fructose and ethanol and to utilize raffinose from the beet molasses media.

Estratégia Retórica 3 - Citar/Descrever materiais e métodos - Sigla MET-DMM

Maltodextrin feed was added *either* continuously *or* in 1.5-min pulses followed by 3.5 min of no carbon addition.

In both addition modes the same total amount of carbon was added.

Estratégia Retórica 4 - Descrever os resultados - Sigla RES-DR

Infants who subsequently developed chronic lung disease were *also* found to have higher chlorotyrosine levels at ~1 wk of age than those who did not.

The fructose and ethanol yields were above 93 and 59% of theoretical values, *respectively* , in beet molasses media with sugar concentrations below 242.0 g/L.

Estratégia Retórica 5 - Explicações/Especulações dos resultados - Sigla RES-EER

Even at a total sugar concentration of 276.2 g/L, the produced syrup contained 69% fructose, and, *thus* , it was richer in this carbohydrate than the ordinary 55% HFCS.

The increase in the Et/Ac ratio was the result of an increase in the ethanol yield combined with a decrease in the acetate yield.

Estratégia Retórica 6 - Apresentar limitações da pesquisa - Sigla CON-ALP

Further studies are required to intensify the biomass loading and increase the yield beyond the current 73% without compromise to the quality of fractionation and product purity.

Estratégia Retórica 7 - Apresentar implicações da pesquisa - Sigla COM-AIP

These findings, if proven to be widely applicable, could lead to significant productivity increases in filamentous fungal fermentations used to produce recombinant protein and could potentially benefit the initial stages of downstream processing.

It is possible that the optimum ATPE conditions for lysozyme purification from tobacco extract might be *further* improved using PEG with different molecular masses or other PEG-salt systems.

Estratégia Retórica 8 - Apresentar recomendações - Sigla COM-AR

Further use of this selective biomarker assay should be a valuable tool for establishing whether neutrophil oxidants *indeed* have a causal role in chronic lung disease as well as for monitoring effectiveness of intervention strategies.

Studies of G6PD partitioning in two-phase aqueous mixed (nonionic/cationic) micellar systems in the presence of other proteins, *as well as* in real fermentation broths, should be performed in the future to study the effect of the other components present in the system on the G6PD partitioning behavior.

Estratégia Retórica 9 - Citações de trabalhos anteriores do autor - Sigla COM-TAA

We have shown in our previous study [17] that *although* it has poor skin permeability, acetaminophen (paracetamol) can be delivered into the systemic blood by using dermal patches containing glyceryl oleate, PEG-40 stearate, tetraglycol, isopropyl myristate and water.

Our previous work had established that a model NADPH-dependent reaction could be carried out efficiently by engineered E. coli cells under glucose fed-batch conditions in the absence of cell division (13).

Estratégia Retórica 10 - Citações de trabalhos anteriores - Sigla COM-TA

For example , Alred et al.(1994) reported a G6PD partition coefficient value of about 0.75 in the EO20PO80/ Dextran T500 two-phase aqueous polymer system.

In addition , increases in the G6PD partition coefficient from about 0.005 to about 0.03, using unbound triazine dyes as affinity ligands in two-phase aqueous PEG/phosphate systems, were reported by Bhide et al.(1995) and by Wang et al.(1992).

Estratégia Retórica 11 - Apresentar contribuições/valor da pesquisa - Sigla COM-ACP

In conclusion , two-phase aqueous mixed (nonionic/ cationic) micellar systems can be considered as a new promising alternative for the purification of G6PD.

In view of *these* numerous successful applications of the CPE technique, which employs water as the predominant component along with small amounts of nonionic or zwitterionic surfactants, it should be evident that CPE represents an attractive alternative to conventional organic-solvent-based extractions.

O texto a seguir é uma seção "Conclusões" que possui sua estrutura esquemática e respectivas estratégia retóricas anotadas. A primeira identificada pela primeira parte da sigla (lado esquerdo do hífen da sigla) e a segunda pelo lado direito da sigla. Trata-se de um resultado de um artigo científico da área de Ciências Farmacêuticas, retirado da Base de Casos do ambiente SciPo-Farmácia (http://www.nilc.icmc.usp.br/scipo-farmacia/).

Conclusão - Caso conc_44

Link: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TFJ-44JJJJ5-1&_user=972067&_handle=B-WA-A-A-AC-MsSAYVW-UUA-AUYYDYUZZU-AUYZBZAVZU-VVCDAUUYA-AC-U&_fmt=full&_coverDate=08%2F28%2F2002&_rdoc=4&_orig=browse&_srch=%23toc%235228%232002%23999129996%23329058!&_cdi=5228&view=c&_acct=C000049650&_version=1&_urlVersion=0&_userid=972067&md5=f910229c999cc83ba455fb792361ab53

Process integration using aqueous two-phase partition for the recovery of intracellular proteins

MARCO RITO-PALOMARES, ANDREW LYDDIATT

PRO-APP The process integration strategy presented *here* for the recovery of intracellular proteins demonstrates that simultaneous disruption and aqueous two-phase extraction can achieve the primary recovery of intracellular proteins from yeast.

RES-EER *In particular* , operating conditions have been established that facilitate the in situ, primary recovery of G3PDH directly and rapidly from disrupted yeast with a significant degree of purification in respect of the reduction of bulk protein and elimination of cell debris in a single operation.

CON-ALP Further studies are required to intensify the biomass loading and increase the yield beyond the current 73% without compromise to the quality of fractionation and product purity.

COM-ACP *However* , the preliminary data presented *here* demonstrate the potential of the integration of ATPS with cell disruption for the direct recovery of specific intracellular protein targets.

Apêndice 8: Manual de anotação das estruturas esquemáticas e estratégias retóricas da seção "Introdução"

As orientações abaixo descrevem o esquema de anotação manual das estruturas esquemáticas e estratégias retóricas para a seção "Introdução" de córpus de artigos científicos em inglês.

IX. Artigo científico: estrutura

Uma característica comum a praticamente todos os textos científicos, que descrevem pesquisa experimental, é o tipo de organização que sua estrutura esquemática deve seguir. Essa estrutura pode ser apresentada como Introdução, Desenvolvimento e Conclusão, sendo que o Desenvolvimento pode ser subdividido em Materiais e Métodos e Resultados, ou ainda Materiais e Métodos, Resultados e Discussão. O objetivo desse tipo de estruturação é guiar o leitor e fazer com que ele siga, na leitura ou

escrita do texto, o movimento do fluxo da informação a ser transmitida que parte do geral-para-específico na Introdução e chega ao específico-para-geral, na Conclusão, conforme pode ser observado na figura abaixo.

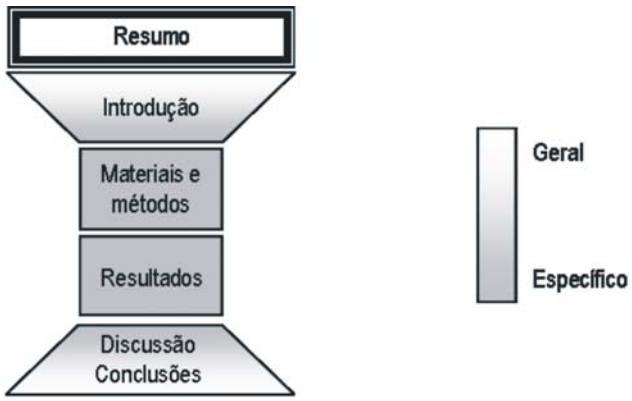


Figura 1: Movimento Geral-Específico-Geral presente na estrutura global do texto científico (Weissberg & Buker, 1990)

X. O que é a seção "Introdução"?

A Introdução serve como uma orientação aos leitores de um artigo científico, mostrando uma perspectiva mais detalhada do que tem que ser conhecido para a compreensão das outras seções do artigo. A sua função é apresentar o assunto do trabalho e, sendo assim, ela deve ser apresentada de maneira clara, simples e sintética, colocando o tema dentro de um quadro de referência teórica atualizado. Embora as seções relativas aos métodos, resultados e discussões possam ser compreendidas somente por especialistas, a introdução deve ser escrita numa linguagem direta e não técnica, de modo que sua apresentação possa ser entendida por todas as pessoas para as quais o trabalho será distribuído.

As introduções de quase todas as áreas de estudo são escritas de uma maneira muito similar. Os tipos de informação incluídos e a ordem em que aparecem são muito convencionais, de modo que podem ser enunciados como modelos de introduções. Tais modelos objetivam guiar o escritor no sentido do tipo de informação que deve ser incluída em uma boa seção "Introdução" e da ordem que tais informações devem aparecer. Assim, O quadro 1, contém as informações típicas que, segundo Weissberg & Buker, (1990:186), podem ser encontradas, em geral, em conclusões de quaisquer áreas do conhecimento.

Contexto	Familiarizar termos, objetos e processos
Revisão da Literatura	Fornece informações para auxiliar o leitor no entendimento do estudo
Lacuna	Indicação de tópico importante ainda não pesquisado
Propósito	Apresentar o propósito principal
Metodologia	Descrição dos materiais e métodos utilizados no experimento
Resultado	Menção dos resultados como gancho para a conclusão que fechará o texto.
Justificativa/Valor	Informações que indiquem o valor do trabalho
Estrutura	Indica as seções de um artigo

Quadro 1: Informações retóricas ou estruturas esquemáticas típicas de Introduções.

III Processo de anotação – o que é?

Consiste, nesse caso, na identificação da função retórica de cada sentença do texto utilizando-se para isso siglas representativas desse papel retórico, as quais serão colocadas no início de cada sentença.

IV Antes do processo de anotação - orientações

- Importante ler o texto antes da anotação, uma vez que a interpretação de determinadas sentenças só se torna possível após uma visão geral do texto.
- Não oriente sua leitura para o entendimento da pesquisa relatada, mas sim para o entendimento da estrutura de argumentação construída pelo autor.
- Não anotar o título ou os subtítulos do texto. Utilize-os apenas como dica/ponteiro do conteúdo que se encontra abaixo dos mesmos.

V Durante o processo de anotação - orientações

O processo de anotação (ou classificação) deve ser feito para cada sentença do corpus, que receberá uma classificação (ou categoria). Entretanto, pode haver sentenças que apresentam características de mais de uma categoria, ou seja, sentenças nas quais os papéis argumentativos se sobrepõem, como por exemplo, sentenças que relatam ao mesmo tempo o procedimento e os equipamentos utilizados. Nesses casos, deve-se identificar a sentença, por meio de uma barra entre as categorias possíveis.

Exemplo:

PRO/EQU Cells were subjected to centrifugation in a Ficoll Hypaque density gradient (Amersham Pharmacia) to further purify PMNs. PRO/PAD The PMNs then were removed by filtration, and the supernatants were analyzed by HPLC analysis.

Exemplos retirados do texto Met_02 do córpus Met comporto por seções "Metodologia" de aritigos científicos da área de Ciências Farmacêuticas.

Note ainda que sentenças consecutivas do texto podem receber a mesma classificação. É comum anotar sentenças consecutivas com a mesma categoria, desde que juntas preencham os critérios de uma dada categoria. Por exemplo: é possível marcar mais de uma sentença como MAT (Materiais) se juntas, elas compõem a lista de materiais utilizados na pesquisa, por exemplo.

MAT Whole antibodies 31154 (human IgG) and 31127 (horse IgG) were obtained from PharMingen. MAT Bovine catalase was obtained from Sigma. MAT All assays were carried out in PBS (10 mM phosphate/160 mM sodium chloride, pH 7.4). MAT Commercial protein solution samples were dialyzed into PBS as necessary. MAT Indigo carmine, isatin sulfonic acid, HOCl, H2O2, vinylbenzoic acid, and 4-carboxybenzaldehyde were obtained from Aldrich.

Exemplo de frases consecutivas com mesma categoria (extraídas do córpus Met):

Se não for possível atribuir nenhuma categoria do esquema utilizado a uma dada sentença, anote-a com um identificador qualquer (por exemplo "?") e descreva, em uma folha a parte, a dificuldade sentida em classificá-la e a função que ela apresenta. Anote também as possíveis dificuldades na classificação de trechos, bem como com a própria categoria que está sendo utilizada. Ex:

? Thiols, at an initial concentration of 1 mM, together with additives as indicated, were equilibrated to 37°C in a shaking water bath.

Importante:

*Qualquer tipo de dúvida é muito importante e deve ser anotada e levada para discussão com o grupo, pois visa a uma melhor caracterização/adequação de uma categoria problemática.

*Não se esqueça de anotar todas as sentenças do texto e suas eventuais dificuldades, que serão discutidas numa reunião com o grupo de anotadores.

*Anote também quais foram os critérios que utilizou para identificar as funções retóricas e, posteriormente, anotar o texto.

VI. Categorias para a anotação retórico-manual do texto

As categorias escolhidas para realizar a anotação retórica dos textos foram inspiradas no modelo de estruturação retórica proposto por Weissberg & Buker, (1990). A seguir, no Quadro 2, temos uma lista de siglas que correspondem as possíveis seções que podem ser encontradas na seção "Conclusão" de um artigo experimental. Para cada uma dessas categorias foi elaborada uma sigla que a representasse para facilitar o processo de anotação do córpus.

Siglas das categorias	Descrição das siglas
CON-DPT	Declarar proeminência do tópico/área
CON-TOP	Familiarizar termos, objetos ou processos
REV-RHA	Revisão histórica da área
REV-TAA	Tendências atuais na área
REV-PA	Progresso na área
REV-RPA	Requisitos para o progresso na área
REV-EA	Estado da arte
REV-CGC	Citações e *gaps* cíclicos
REV-CAA	Citações agrupadas por abordagens
REV-TAA	Citações de trabalhos anteriores do autor
REV-RRR	Revisão de resultados relevantes
LAC-PNR	Existência de conflitos ou problemas não resolvidos
LAC-RTA	Restrições em trabalhos anteriores
LAC-QNC	Questões ainda não consideradas
PRO-RCA	Resolver um conflito entre autores
PRO-AMT	Apresentar uma nova abordagem, metodologia ou técnica
PRO-AML	Apresentar melhorias/avanços em um tópico da literatura
PRO-AEA	Apresentar uma extensão de um trabalho anterior do autor
PRO-AAA	Apresentar uma abordagem alternativa
PRO-EP	Especificar o propósito
PRO-IMP	Introduzir mais propósitos
PRO-AP	Apresentar o propósito
PRO-APR	Apresentar o propósito com resultados
MET-ICC	Indicar critérios ou condições
MET-DMM	Descrever materiais e métodos
RES-AR	Apresentação dos resultados
RES-LR	Listagem dos resultados
RES-CR	Comentários sobre os resultados
JUV-VP	Valor da pesquisa
EST-ISA	Indicar as seções do artigo

Figura 2: Note, na figura acima, que a sigla da categoria é composta sempre por letras que compõem uma dada categoria, de forma a facilitar a memorização e fácil identificação do significado da categoria que deverá ser empregada nas sentenças do corpus.

3.1. O modelo de Weissberg & Buker (1990)

A análise dos Movimentos, desenvolvida por Weissberg & Buker (1990), visa a representar artigos científicos em termos de sua organização textual hierárquica construída por seções distintas.

Cada seção pode ser dividida em estruturas esquemáticas, que por sua vez, pode ser subdividida em Estratégias Retóricas.

A estrutura esquemática consiste em uma unidade textual funcional, utilizada com algum propósito retórico identificável. Essas estruturas realizam-se lingüisticamente em um texto de diferentes formas, ou seja, por diferentes estratégias retóricas, conforme a estrutura esquemática na qual se encontram contidas.

Vale lembrar que essas estruturas visam a descrever de maneira geral as funções retóricas que podem ser encontradas em um córpus. Mas é bem possível que outras funções possam ser encontradas no córpus em análise e acrescentadas a esse modelo. Aliás, esse modelo é um ponto de partida para que seja aprimorado segundo as características que o córpus em estudo/análise apresentar.

3.2 Detalhamento das categorias

Nessa seção, são apresentados as categorias e a forma como essas podem aparecer no texto. A definição das estruturas esquemáticas abaixo apresentadas tiveram suas definições baseadas no trabalho de Feltrim (2004).

Estratégia Retórica 1 - Declarar proeminência do tópico/área - Sigla CON-DPT

Given *these* potential physiological and pathological roles, it is important to understand the regulation of IDO in macrophages.

The compound is a potent oxidant that has been receiving increasing attention as a potential pathogenic mediator in human diseases and as a cellular toxin in host defense mechanisms against invading microorganisms (3-6).

Estratégia Retórica 2 - Familiarizar termos, objetos ou processos - Sigla CON-TOP

As is almost invariably the case with autoxidation reactions, thiols do not react directly with oxygen and the reaction is catalyzed by transition metals.

Oxidation of thiols is a very complex process, largely due to different liganding of transition metals both to the thiol itself and to other complexing agents that may be present.

Estratégia Retórica 3 - Revisão histórica da área - Sigla REV-RHA

It has been proposed that peroxynitrous acid (ONOOH) promotes one-electron oxidations following a rate-limiting unimolecular activation to a species whose chemical identity, an activated form of peroxynitrous acid (ONOOH*) or the hydroxyl radical, remained under debate for a long time (3, 14-18).

It was only recently that clear experimental evidence was obtained demonstrating that a significant portion of the oxidative activity of peroxynitrous acid is *because* of the hydroxyl radical.

Estratégia Retórica 4 - Tendências atuais na área - Sigla REV-TAA

These data suggest that physiologic cells expressing IDO inhibit the generation of T cell responses in vivo.

At present , a significant part of the biological reactivity of peroxynitrite is ascribed to the adduct produced by its reaction with carbon dioxide (7-13).

Estratégia Retórica 5 - Progresso na área - Sigla REV-PA

There is now considerable evidence to suggest that early inflammation plays an important role in the development of chronic lung disease (10-14).

Neutrophils are attracted to the lungs, and ongoing neutrophil infiltration is associated with poor respiratory outcome (13, 15-17).

Estratégia Retórica 6 - Requisitos para o progresso na área - Sigla REV-RPA

To provide a firm basis for any intervention or management strategies, more evidence that reactive oxidants have a pathologic role in the disease is needed.

Equally important is to identify the oxidant source.

Estratégia Retórica 7 - Estado da arte - Sigla REV-EA

Currently , it is proposed that in resting state, p47PHOX is folded in a masked conformation involving intramolecular interactions between the two SH3 domains.

Upon activation, phosphorylation of p47PHOX disrupts the SH3-mediated intramolecular interaction and p47PHOX adopts a conformation that allows it to interact with the p22PHOX (13), bringing p67PHOX in proximity with cytochrome b558.

Estratégia Retórica 8 - Citações e *gaps* cíclicos - Sigla REV-CGC

The mechanism of HOCl-mediated protein aggregation, and the properties that make some proteins more susceptible than others, have not been established.
Aggregation has been observed as high-molecular-mass bands on SDS/PAGE or by size exclusion chromatography, even under reducing conditions, and is generally assumed to represent intermolecular covalent cross-linking of the protein.
Estratégia Retórica 9 - Citações agrupadas por abordagens - Sigla REV-CAA
There are two principal methods to produce the heterodimeric form of HIV-1 RT, namely, expressing both subunits individually either in the same cell or different cells [11, 12 and 13], or relying on an endogenous Escherichia coli protease or HIV-1 protease to convert the p66 protein to the heterodimer [11, 12, 13 and 14].
These strategies have been successfully employed to produce pure proteins, *however* , the yield and homogeneity of such preparations have often been limiting [12, 13, 14 and 15].
Estratégia Retórica 10 - Citações de trabalhos anteriores do autor - Sigla REV-TAA
Recently we discovered that all antibodies can catalyze the formation of H2O2 from 1O and H2O (8) via the postulated intermediacy of dihydrogen trioxide (H2O3) (9).
An oxidative component of the cascade of reactive intermediates generated during this process possesses the chemical signature of ozone (10).
Estratégia Retórica 11 - Revisão de resultados relevantes - Sigla REV-RRR
There is evidence that this toxicity results from thiol oxidation with concomitant generation of "active oxygen" species (4, 5 and 6).
A proportion of the peroxide reacts with more cysteine (11, 12, 13 and 14)
Estratégia Retórica 12 - Existência de conflitos ou problemas não resolvidos - Sigla LAC-PNR
However , thiols themselves can be harmful, *for example* by causing necrosis, apoptosis, chromosome aberrations, DNA damage, and mutagenesis (3, 4, 5, and 6).
This is not an uncommon picture for metal-catalyzed reactions.
Estratégia Retórica 13 - Restrições em trabalhos anteriores - Sigla LAC-RTA
Other aspects of the reaction are more controversial, and there are many anomalies in the literature.
Studies on the effects of catalase (20, 21 and 22) and chelating agents (12, 13 and 24) have *also* been inconsistent and open to various interpretations.
Estratégia Retórica 14 - Questões ainda não consideradas - Sigla LAC-QNC
The function of Rac remains unclear, *although* it is absolutely required for NADPH oxidase activation (8).
The function of p40PHOX *also* is not well defined.
Estratégia Retórica 15 - Resolver um conflito entre autores - Sigla PRO-RCA
To resolve the apparent differences, the reaction of peroxynitrite with recombinant purified rat TH in vitro was re-examined, and no evidence of cysteine oxidation was found.
In the present study , we have undertaken a thorough kinetic analysis of the sensitivity of caspase-3 in cell lysates and its recombinant form to oxidation by H2O2.
Estratégia Retórica 16 - Apresentar uma nova abordagem, metodologia ou técnica - Sigla PRO-AMT
Therefore , the current study focuses *specifically* on the interaction of the neutrophil and the pulmonary microvascular endothelium as two early components of the host innate immune response to bacterial infection in the lung.
In this report , we describe an approach to the validation of proven acceptable ranges for critical process parameters such as pH and temperature for the production of a humanised monoclonal IgG1 antibody by a murine myeloma cell line in protein-free fed-batch cell culture.
Estratégia Retórica 17 - Apresentar melhorias/avanços em um tópico da literatura - Sigla PRO-AML
In this study we examined whether antioxidants regulate the induction of IDO in IFN-gamma-activated hMDM.
To investigate the structural features underlying the chemical reactivity of IDO in more detail, we have measured the resonance Raman spectra of recombinant human indoleamine 2,3-dioxygenase (hIDO) and its cyanide, carbon monoxide, and L-Trp complexes.
Estratégia Retórica 18 - Apresentar uma extensão de um trabalho anterior do autor - Sigla PRO-AEA

To complement pharmacological studies and to *further* address relationships between IDO activity and inhibition of T cell responses, we used two molecular genetic strategies to enhance IDO activity in transfected cell lines and in new strains of transgenic mice.
To *further* delineate the capabilities of different TLRs to discriminate microbial products, we hypothesized that GBS would be recognized by both known and novel Toll receptors.
Estratégia Retórica 19 - Apresentar uma abordagem alternativa - Sigla PRO-AAA
We have tried to trap the carbonate radical in systems containing peroxynitrite and bicarbonate under different experimental conditions but did not succeed.
Consequently , we considered it worth trying to detect the carbonate radical directly by continuous fast flow EPR of peroxynitrite and bicarbonate solutions.
Estratégia Retórica 20 - Especificar o propósito - Sigla PRO-EP
In the current study, we directly evaluated the hypothesis that enhanced IDO activity in cells or tissues inhibits T cell responses.
We *specifically* examined the role of the TLR adapter molecule MyD88 and the receptors TLR1, 2, 4, and 6 using the companion approaches of cellular transfection with cDNA constructs (gain of function) and the examination of macrophages from genetically deficient animals (loss of function).
Estratégia Retórica 21 - Introduzir mais propósitos - Sigla PRO-IMP
Furthermore we add further weight to our original observation that a powerful oxidant with the chemical signature of ozone is generated by human PMNs with the use of a second ozone probe, vinylbenzoic acid.
Additionally , we show that massive deposition of complement and hemorrhagic necrosis occurs at the maternal-fetal interface when mice carrying an allogeneic fetus are exposed to 1-methyl-tryptophan and that this inflammation is driven by T cell recognition of fetal antigens.
Estratégia Retórica 22 - Apresentar o propósito - Sigla PRO-AP
In this report , we describe the exceptional sensitivity of copper-catalyzed cysteine oxidation to the presence of iron.
We now report the effect of modifying both surface antibody concentration and the presence of catalase on the production of ozone by PMNs.
Estratégia Retórica 23 - Apresentar o propósito com resultados - Sigla PRO-APR
In the present study , we show a direct interaction between p67PHOX and cytochrome b558 and find that this interaction increases when the proteins are incubated in the presence of Rac1-GTP/GDP.
We demonstrate *here* that the degree of genetically determined tissue incompatibility between parental strains directly determines the rate of pregnancy failure in mice exposed to 1-methyl-tryptophan.
Estratégia Retórica 24 - Indicar critérios ou condições - Sigla MET-ICC
The major factors that limit E. coli culture densities are a lack of oxygen and the accumulation of metabolic byproducts [1 and 2].
Thus , important considerations for any bacterial fermenter are the aeration and mixing systems.
Estratégia Retórica 25 - Descrever materiais e métodos - Sigla MET-DMM
The expression of several endothelial adhesion molecules (ICAM-1, ICAM-2, VCAM-1, and E-selectin) in response to these pathogens was characterized using intact cell ELISA and immunofluorescence microscopy (IFM), and chemokine expression by the endothelium or the neutrophils was quantitated by ELISA of the supernatant for IL-8, monocyte chemoattractant protein-1 (MCP-1), IL-6, RANTES, TNF-alpha, and growth-related oncogene-alpha (GRO-alpha).
We have studied human haemoglobin and horse heart myoglobin, as well characterized model proteins that readily undergo aggregation, and used the haem-depleted apo forms to avoid complications due to the reaction of HOCl with the haem groups.
Estratégia Retórica 26 - Apresentação dos resultados - Sigla RES-AR
The data show that the distal and proximal heme environments of hIDO are distinctly different from that of conventional Mb and that L-Trp binds closely *but* not directly to the distal side of the heme iron.
Oxidation of one cysteine residue per molecule of TH was observed only at high peroxynitrite concentrations, and three cysteine residues were oxidized in partially unfolded protein.
Estratégia Retórica 27 - Listagem dos resultados - Sigla RES-LR
We report the effects of pulse feeding on cell growth, broth viscosity, recombinant enzyme productivity, and oxygen mass transfer.

Furthermore , we describe an accurate and reliable spectrophotometric assay that should allow detailed future kinetic analysis.

Estratégia Retórica 28 - Comentários sobre os resultados - Sigla RES-CR

It needs to be taken into account when interpreting experimental studies and may explain some of the anomalies in the literature.

Although interaction between iron and copper in cysteine oxidation has been noted previously (11), the high sensitivity to iron has not been described and implications of the effect have not been featured in subsequent studies.

Estratégia Retórica 29 - Valor da pesquisa - Sigla JUV-VP

The nature and stability of the association complexes formed by PLA-PEG copolymers in aqueous dispersions are of fundamental importance as regards their potential drug carrying capacity.

Estratégia Retórica 30 - Indicar as seções do artigo - Sigla EST-ISA

In the Materials and Methods section we describe the materials and experimental methods utilized in this investigation.

Next, we present a review of the protein partitioning theories based on excluded-volume interactions (Nikas et al., 1992) and electrostatic interactions (Kamei et at., 2002a) to predict protein partition coefficients in two-phase aqueous mixed (nonionic/ionic) micellar systems.

O texto a seguir é uma seção "Introdução" que possui sua estrutura esquemática e respectivas estratégia retóricas anotadas. A primeira identificada pela primeira parte da sigla (lado esquerdo do hífen da sigla) e a segunda pelo lado direito da sigla. Trata-se de um resultado de um artigo científico da área de Ciências Farmacêuticas, retirado da Base de Casos do ambiente SciPo-Farmácia (http://www.nilc.icmc.usp.br/scipo-farmacia/).

Introdução - Caso intro_19

Link:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12062443&dopt=Abstract

Hampton, M.B., Stamenkovic, I. and Winterbourn, C.C. Interaction with substrate sensitises caspase-3 to inactivation by hydrogen peroxide. FEBS Lett. 517:229-232, 2002

Interaction with substrate sensitises caspase-3 to inactivation by hydrogen peroxide

CON-TOP Cysteine residues play a fundamental role in protein structure and catalytic activity.

CON-TOP Oxidation and reduction of these residues can regulate a variety of signal transduction pathways in cells [1].

CON-TOP Proteins whose function can be altered by cysteine oxidation include protein kinases and phosphatases, transcription factors, membrane receptors and channel proteins.

CON-TOP H2O2 is one of the oxidants speculated to regulate cell function via its interaction with thiol proteins.

CON-TOP While cysteine itself has a low reactivity with H2O2 at neutral pH, the local protein environment can enhance the reactivity of selected cysteine residues.

CON-TOP One common *example* is the presence of positively charged residues that stabilise the thiolate anion.

CON-TOP *This* enables some form of selectivity during redox signaling.

CON-TOP The caspases are a family of cysteine proteases that play an essential role in the execution of apoptosis [2].

CON-TOP They are expressed as inactive zymogens, and become proteolytically active during apoptosis.

CON-TOP Caspases have an active site cysteine that mediates nucleophilic attack on its target substrate.

CON-TOP The thiol has to be reduced for the enzyme to function, and dithiothreitol [DTT] is regularly included in activity assays.

REV-RRR In a study with apoptotic cells it was observed that H2O2, depending on the time of its addition, could delay the onset of caspase activation or impair the activity of those effector caspases present immediately prior to harvest [4].

REV-RRR *This* suggested that oxidation could directly inhibit caspase activity, and observations of oxidative stress during apoptosis raised the possibility of a physiological and pathological role for caspase oxidation [4].
REV-RRR Consistent with these observations, inactivation of recombinant caspases at low concentrations of H2O2 has been reported [5], as has the effect of other thiol oxidants and reductants [6, 7 and 8].
REV-RRR Reactive nitrogen species have also been shown to inhibit caspase-3 activity in vitro via thiol modification [9, 10, 11 and 12], and one report provided evidence that the pro-form of caspase-3 is S-nitrosylated in resting cells [13].
LAC-PNR *However* , there are contradictory reports with cell extracts and purified caspases showing minimal inactivation by H2O2 [14 and 15].
PRO-RCA *In the present study* , we have undertaken a thorough kinetic analysis of the sensitivity of caspase-3 in cell lysates and its recombinant form to oxidation by H2O2.
RES-AR We have discovered an unusual phenomenon whereby caspase-3 becomes more sensitive to oxidative inactivation in the presence of its substrate.
RES-CR *This* explains contradictory results in the literature, and it identifies a novel mechanism for sensitising a thiol enzyme to oxidative inactivation.

Apêndice 9: Instruções de utilização do WordSmith Tools na extração de palavras-chaves

O *WordSmith Tools* é um *software*, desenvolvido por Mike Scott e publicado pela Oxford University Press desde 2001, somente obtido pela Internet, nos seguintes endereços: www.liv.ac.uk/~ms2928/;www.lexically.net/; www.oup.com/elt/global/isbn/6890/ (Berber-Sardinha, 2004; 1999). Nesses endereços, o usuário baixa a versão demo e se desejar a versão completa, precisa pagar uma licença para receber um código que o habilitará para converter a versão demo para uma completa. É de fácil manuseio e, por isso, seu uso se estende em diferentes áreas da comunidade lingüística. A *Oxford University Press*, por exemplo, a utiliza em trabalhos de lexicografia, que envolvem a preparação de dicionários; professores de língua, estudantes e pesquisadores na análise de padrões de uma dada língua podem por sua vez, utilizá-la na investigação de concordâncias, por exemplo.

Essa ferramenta disponibiliza ao seu usuário diferentes recursos que, se bem aproveitados, propiciam uma análise bastante consistente acerca de vários aspectos da

linguagem, como, por exemplo, sobre a composição lexical (freqüência, tamanho e balanceamento) do córpus coletado. Esses recursos são descritos na próxima seção.

O *WordSmith Tools* é composto por três ferramentas (*WordList, KeyWords e Concord*) e quatro utilitários (*Renamer, Text Converter, Splitter e Viewer*). Dentre esses diferentes recursos, as duas primeiras ferramentas citadas são utilizadas em nosso estudo, portanto serão descritas com maiores detalhes a seguir.

***a)* WordList**

Esse recurso produz, a cada vez que utilizado, listas de palavras (individuais ou multipalavras[2]) ordenadas de três formas distintas: 1) por ordem alfabética crescente (identificada pela letra A entre parênteses), 2) com base em medidas estatísticas (identificada pela letra S) ou 3) por ordem crescente de freqüência (identificada pela letra F, com as palavras mais freqüentes encabeçando o topo dessa lista). Para se obter uma lista de palavras de um córpus, basta seguir os seguintes comandos:

(1) Na tela inicial (*Controller*), clique na opção *Tools* e em *Word List*.

(2) Na janela do *Word List*, clique em *File* e *Start*.

(3) Na janela *Getting Started*, clique em *Choose Texts Now*.

(4) Na janela *Choose Texts*, clique no diretório (pasta) que contém os textos, clique nos textos desejados e, finalmente, em Ok.

(5) Na janela *Getting Started*, clique em *Make a WordList Now*.

***b)* KeyWords**

Esse recurso, também disponibilizado pela ferramenta *WordSmith Tools*, contrasta uma lista de palavras (ou mais de uma) de um córpus de estudo com uma lista de palavras de um córpus de referência, seja ele de outra área de especialidade ou de língua geral, produzindo uma terceira contendo somente as palavras-chaves do córpus em estudo.

O que se entende por palavras-chaves obtidas por essa ferramenta não tem relação com a lista das palavras mais importantes do córpus, uma vez que a característica de uma dada palavra ser considerada palavra-chave é definida por sua freqüência no córpus. Assim, uma palavra pode ser chave se sua freqüência for muito alta (positiva) ou muito baixa (negativa) em relação ao de referência. Poder-se-ia questionar a utilização de palavras-chaves eleitas pelos autores dos textos compilados. No entanto, as mesmas quando escolhidas

[2] Multiplavras: termo corrente na área de fraseologia, mas também conhecido como multi-word units, polywords. Na área de PLN é comparável ao termo n-grama.

aleatoriamente, isto é, sem o auxílio de uma análise estatística do córpus para levantamento das palavras-chaves, as mesmas podem não fazer parte do corpo do texto. Portanto, optou-se pela seleção das palavras-chaves geradas pelo recurso *KeyWords*. Para se obter uma lista de palavras-chaves de um córpus, basta seguir os seguintes comandos:

(1) Na tela inicial, clique em *Tools* e depois em *KeyWord*.

(2) Na janela do *KeyWord*, clique em *File* e depois em *Start* ou no botão iniciar (bolinha verde)

(3) Na janela *Getting Started*, clique em *Find the key words in a text*, o que resulta no aparecimento da janela *Choose Word Lists.*

(4) Na janela da esquerda, clique sobre a *WordList* do córpus de estudo e na janela da direita, clique sobre a *wordlist* do córpus de referência, o BNC (British National Corpus).

(5) Clique em OK.

(6) O processamento é então iniciado. Para interrompê-lo, clique em *Suspend*, na janela de andamento e, a seguir, em *Stop Now*.

(7) A lista será então mostrada na tela.

4.4.2.2 Exemplo de palavras-chaves extraídas pelo Word Smith Tools

Nessa subseção são apresentadas as palavras-chaves geradas a partir do córpus Met, com o auxílio das ferramentas computacionais *KeyWord* e *WordList*, a pouco apresentadas.

Ao final dos procedimentos descritos para o uso da ferramenta *WordList* com o Córpus Met, obteve-se a lista mostrada na Figura 4.6.

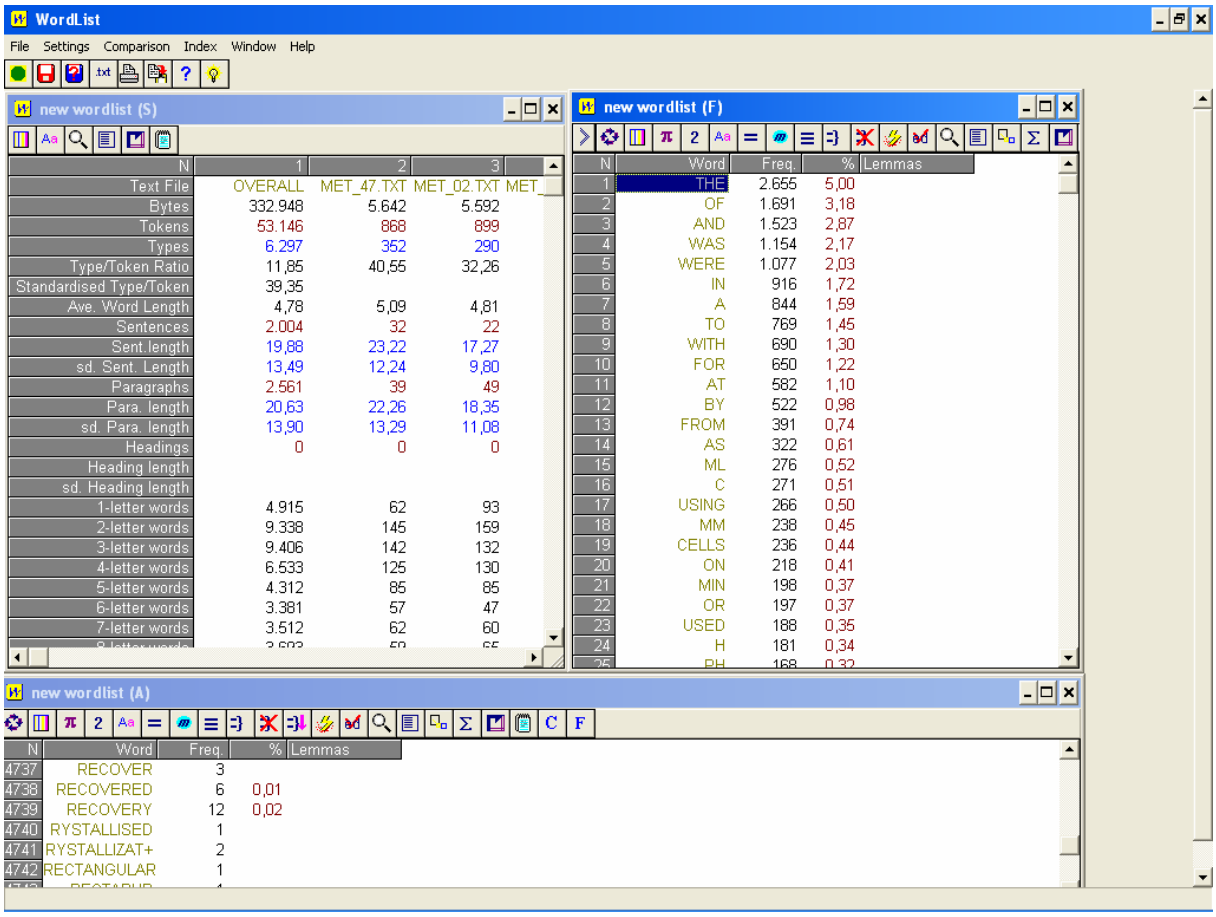


Figura 4.6: *WordList* gerada com o Córpus Met.

As listas alfabética (A) e freqüencial (F) possuem as seguintes informações:

Coluna Word: os itens lexicais (palavras) contidos nos textos.

Coluna Freq.: quantas vezes cada item apareceu no córpus.

Coluna %: a porcentagem do total de itens do texto a que corresponde cada item.

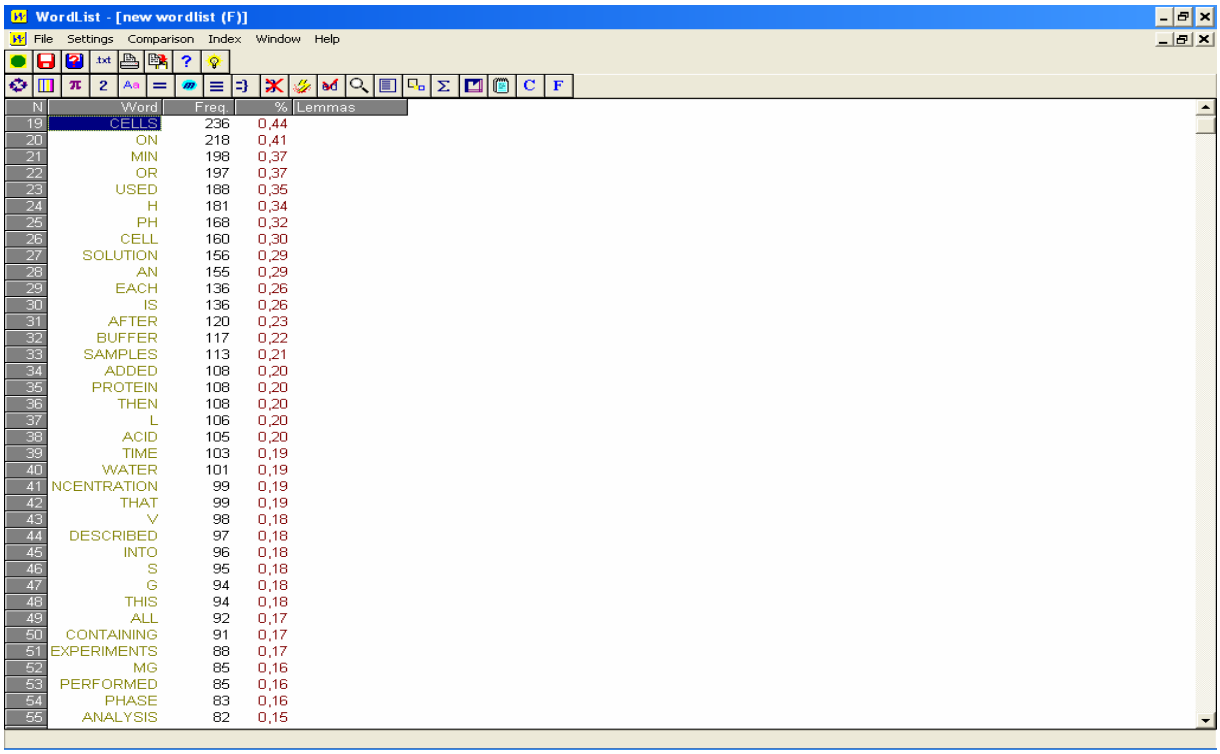
Coluna Lemas: outros itens cujas freqüências foram adicionadas ao item corrente, por exemplo, o lema "amar" compreende as formas "amo", "ame", "amando", etc. Nota-se que essa coluna não apresenta dados porque a opção de lematizar os itens não foi desejada em nosso trabalho.

Como vimos, a lista de freqüência de palavras fornecida pelo programa *WordSmith Tools* mostra dados sobre o número de ocorrências das palavras do córpus, o que possibilitou a identificação das palavras de baixa e de alta freqüência.

Ao observarmos as palavras que possuem uma freqüência alta, pode-se dizer que as mesmas são importantes, pois poderiam ser candidatas a termos recorrentes de uma dada área de especialidade. Assim como as palavras de baixa freqüência no córpus também são importantes de serem consideradas, pois elas podem vir a representar termos muito

específicos utilizados em uma determinada subárea da área de especialidade. Uma outra utilidade dessa lista de freqüência é o fato dela poder servir como ponto de partida para a elaboração de glossários de termos técnicos.

Ao analisarmos nosso córpus, verificamos que foi contabilizado um total de 6.297 palavras, com freqüências que variam de 1 a 2.655 (o artigo "the"). As primeiras 18 palavras do córpus são artigos e preposições que, em geral, tem grande aparição em textos. Mas se fizermos um corte e considerarmos as freqüências das palavras a partir do ponto 19, podemos ver as de maior freqüência de uso e, conseqüentemente, identificar os termos recorrentes da área de Farmácia (Figura 4.7).



N	Word	Freq.	%	Lemmas
19	CELLS	236	0,44	
20	ON	218	0,41	
21	MIN	198	0,37	
22	OR	197	0,37	
23	USED	188	0,35	
24	H	181	0,34	
25	PH	168	0,32	
26	CELL	160	0,30	
27	SOLUTION	156	0,29	
28	AN	155	0,29	
29	EACH	136	0,26	
30	IS	136	0,26	
31	AFTER	120	0,23	
32	BUFFER	117	0,22	
33	SAMPLES	113	0,21	
34	ADDED	108	0,20	
35	PROTEIN	108	0,20	
36	THEN	108	0,20	
37	L	106	0,20	
38	ACID	105	0,20	
39	TIME	103	0,19	
40	WATER	101	0,19	
41	NCENTRATION	99	0,19	
42	THAT	99	0,19	
43	V	98	0,18	
44	DESCRIBED	97	0,18	
45	INTO	96	0,18	
46	S	95	0,18	
47	G	94	0,18	
48	THIS	94	0,18	
49	ALL	92	0,17	
50	CONTAINING	91	0,17	
51	EXPERIMENTS	88	0,17	
52	MG	85	0,16	
53	PERFORMED	85	0,16	
54	PHASE	83	0,16	
55	ANALYSIS	82	0,15	

Figura 4.7: *WordList* de freqüências gerada com o Córpus Met.

Em contrapartida, se olharmos para as palavras de baixa freqüência, poderemos ver os termos bem específicos contidos em cada subárea da Farmácia (Figura 4.8).

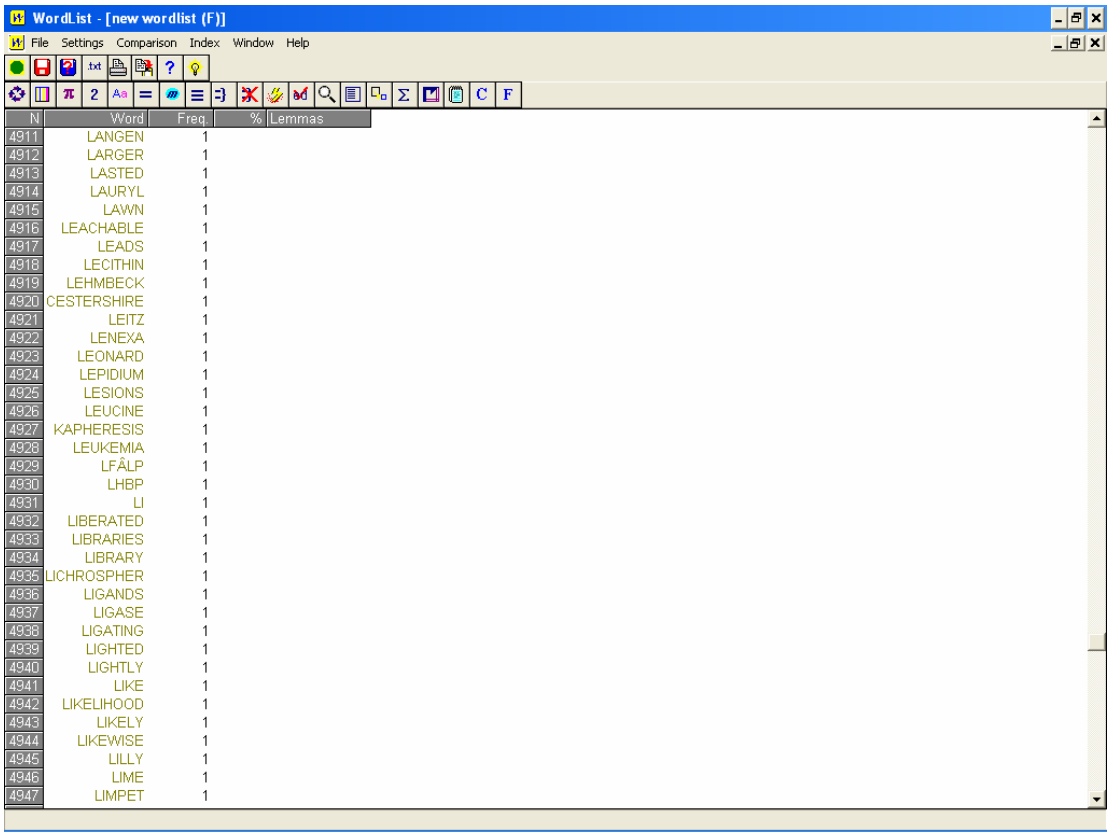


Figura 4.8: *WordList* de freqüências gerada com o Córpus Met.

Ao considerarmos a lista estatística (S), veremos que ela possui:

Coluna 1,2,3,...: número de cada arquivo do córpus.

Text File: o nome de cada texto analisado. Overall é a coluna que reúne todos os textos analisados, o texto Met_47, o Met_02, e assim por diante.

Tokens: é o número de ocorrências. Nesse item está indicado o número total de palavras do córpus ou do subcórpus escolhido para ser analisado. A importância desse dado está na possibilidade de se verificar o tamanho do córpus ou subcórpus por meio do número de palavras que possuem.

Types: o número de vocábulos. Esse dado nos mostra quantas palavras diferentes, excetuando as suas ocorrências, o córpus ou subcórpus contém. É um dado importante para verificarmos se um córpus possui ou não um material variado. Por exemplo, se um córpus for compilado com textos de apenas um tema, o número de tipos de palavras será baixo, mesmo se a quantidade de textos for grande, pois os termos se repetirão ao longo do córpus.

TypeToken Ratio: a razão vocábulo-ocorrência de palavras. Esse número nos fornece a razão dos vocábulos pelo número de ocorrências, cujo resultado indicará a riqueza lexical do texto: quanto maior for essa razão, mais diversificado será o córpus ou subcórpus analisado. Em contrapartida, um valor baixo indicará um número alto de repetições, o que pode indicar um

texto menos rico do ponto de vista de seu vocabulário, sendo necessário, portanto, variar mais as fontes e os textos coletados.

Em nosso trabalho, o propósito de utilizarmos palavras-chaves extraídas de cada texto de nosso córpus Met consiste em detectar, por meio delas, a terminologia específica da linguagem de especialidade contida em cada um deles e, conseqüentemente, podermos alocá-los sob as quatro subáreas da árvore de domínios das Ciências Farmacêuticas gerada para nosso estudo. Assim, por meio desse pequeno conjunto de palavras que fornecem a indicação do conteúdo tratado em cada texto (*aboutness*), tem-se a possibilidade de verificar se o nível de balanceamento do córpus compilado está adequado. Conforme é sugerido por Atkins *et al* (1992), a construção de um dado córpus passa por etapas cíclicas, que se repetem de acordo com as metas e critérios inicialmente estabelecidos. Dessa maneira, é nessa primeira versão de um córpus, que se avalia seu balanceamento, isto é, se as áreas contidas na árvore receberam textos suficientes e em similar quantidade. Caso contrário, uma nova coleta (ou novas coletas) será(o) necessária(s) até que se obtenha um córpus balanceado, equilibrado.

Assim, para produzimos listas de palavras-chaves para cada um dos textos coletados, utilizamos uma lista de palavras de cada texto do córpus Met (córpus de referência da área de especialidade), gerada anteriormente pelo recurso *WordList* e uma lista de palavras do córpus BNC (córpus do inglês britânico escrito e falado, córpus de referência de língua geral), que pode ser obtida no próprio site que disponibiliza a ferramenta, www.liv.ac.uk/~ms2928 ou pelo site http://lael.pucsp.br/direct. O único requisito que se tem nessa tarefa de extração de palavras chaves é o tamanho recomendado para um córpus de referência. Segundo estudo de Berber-Sardinha (2005) o tamanho recomendado para um córpus de referência é de ser 5 X (vezes) maior que o tamanho do córpus de estudo. O estudo ainda sugere haver diferenças de resultados se o córpus de referência é de amostras ou de textos integrais, pois em textos curtos não há tanta repetição de palavras, o que influencia a freqüência.

É importante salientar que nos textos do córpus Met foi realizada uma limpeza da linguagem computacional, na qual foram excluídas determinadas palavras que não têm relação com o domínio das Ciências Farmacêuticas, mas que por constarem no cabeçalho e, às vezes também, pelo corpo do texto, pudessem aparecer listadas como palavras-chaves: http, figura, etc.

Um dos procedimentos mais delicados envolvidos em uma análise de córpus via auxílio do *KeyWords* é a seleção de um subconjunto de palavras-chaves para serem investigadas em detalhe. Essa seleção faz-se necessária, uma vez que o tamanho do léxico considerado chave de um córpus de estudo, em geral, é grande. Uma alternativa que é proposta por Berber-

Sardinha (2005) é a aplicação de um ponto de corte generalizado que indicaria a região da lista de palavras-chaves na qual há maior probabilidade de ocorrência de léxico chave exclusivo. O léxico chave exclusivo é composto por palavras-chaves que ocorrem somente no córpus de estudo em questão. Nesse contexto, os parâmetros utilizados para a extração das palavras-chaves foram os seguintes:

(1) Teste estatístico (ou prova estatística) utilizado na comparação das freqüências das palavras: Log-likelihood, segundo indicação de estudo de Berber-Sardinha (1999) que ao discutir o uso do $\chi 2$ com o Log-likelihood, indica que a melhor escolha para se trabalhar com o *KeyWords* é pelo segundo.

(2) Nível de significância (p) utilizado na comparação: um índice em porcentagem que indica a parcela em palavras-chaves que se deveria manter para se ter a probabilidade de inclusão das palavras-chaves exclusivas de um córpus de estudo. Outro estudo de Sardinha (2005) sobre o ponto de corte generalizado mais eficiente na ferramenta *Keywords* indica o valor de p = 0.0000001 para se reduzir a lista de palavras-chaves que implica num recorte escolhendo as 53 primeiras palavras-chaves da lista ordenada pelas palavras-chaves exclusivas de seu córpus. E já que as palavras-chaves exclusivas são um tipo de léxico categorizador, elas provavelmente serão as mais caracterizadoras de seu córpus de estudo.

A seguir, foram contrastadas trinta listas de palavras do córpus Met com a lista de palavras do BNC para obtermos hipóteses de palavras-chaves específicas da subárea de especialidade a qual pertencem. A partir desses resultados, o próximo passo será disponibilizar essas listas de palavras-chaves para um especialista da área de Farmácia para que ele faça concordâncias com as mesmas, verifique por meio de agrupamentos lexicais se essas palavras são termos ou partes de um termo e, por fim, possa determinar se as palavras escolhidas correspondem realmente a uma das subáreas de especialidade contidas em nossa árvore de domínios. O auxílio desse especialista se tornou necessário uma vez que a compilação desse córpus é feita por uma pessoa não pertencente à área de Ciências Farmacêuticas, portanto, não habilitada/possibilitada a realizar a tarefa de julgamento dos termos, que deve ser feita não só com base em dados da língua, mas também com o auxílio do conhecimento do vocabulário especializado contido em qualquer área do conhecimento.

Portanto, para que se complete a etapa de balanceamento do Córpus_Met, as listas de palavras-chaves poderão ser avaliadas futuramente por um especialista da área de Farmácia, que poderá alocar os textos em cada uma das subáreas da árvore gerada, para podermos avaliar o balanceamento final obtido, isto é, a quantidade de textos existentes em cada uma das subáreas da árvore de domínios gerada para esse córpus.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo