2015

Padrões léxico-gramaticais na especificação de propósito e resultado em abstracts de artigos científicos: aplicações no ensino de EAP e na construção de ferramentas de suporte à escrita científica

PESQUISAS E PERSPECTIVAS EM LINGUÍSTICA DE CORPUS

Coleção Espaços da Linguística de Corpus

Conselho Editorial

ANA MARIA T. IBAÑOS
LÍVIA PRETTO MOTTIN
SIMONE SARMENTO
TONY BERBER SARDINHA
(ORGANIZADORES)

PESQUISAS E PERSPECTIVAS EM LINGUÍSTICA DE CORPUS

MERCADO® LETRAS

AGRADECIMENTOS

Primeiramente, gostaríamos de agradecer à CAPES, Fapergs, Macmillan, STB, SBS, Cambridge University Press, DISAL, Consulado dos Estados Unidos (São Paulo), British Council e John Benjamins, pelo apoio e auxílio à realização do evento que deu origem a este livro. Somos também gratos às equipes da Faculdade de Letras e ao Programa de Pós-Graduação em Letras da Pontifícia Universidade Católica do Rio Grande do Sul por não terem medido esforços para sediar e garanti o sucesso do evento.

Agradecemos a todos que de alguma forma participaram do evento: organizadores, comissão científica, monitores, conferencistas e pesquisadores que contribuíram com a apresentação de seus trabalhos.

De forma especial, agradecemos a todos os autores que nos enviaram seus trabalhos para publicação neste livro. Aos pareceristas anônimos pelas leituras cuidadosas, correções e sugestões, as quais contribuíram para a qualidade do material publicado.

Finalmente, agradecemos à Sylviane Granger e à John Benjamins Publishing Company por permitirem a publicação da versão traduzida do artigo *The contribution of learner corpora*

to second language acquisition and foreign language teaching: A critical evaluation, originalmente publicado na obra *Corpora and Language Teaching* (Aijmer 2008), publicada pela mesma editora.

Simone, Tony, Lívia e Ana

SUMÁRIO

9. PADRÕES LÉXICO-GRAMATICAIS NA ESPECIFICAÇÃO DE *PROPÓSITO* E *RESULTADO* EM *ABSTRACTS* DE ARTIGOS CIENTÍFICOS: APLICAÇÕES NO ENSINO DE EAP E NA CONSTRUÇÃO DE FERRAMENTAS DE SUPORTE À ESCRITA CIENTÍFICA[1]

Carmen Dayrell
Arnaldo Candido Jr.
Mariana Curi
Stella Tagnin
Sandra Aluísio

1. Agradecimentos: Aos Profs. Drs. Adalberto Pessoa Jr. (FCF), Sandra Aluísio (ICMC), Valtencir Zucolotto e Oswaldo Oliveira (IFSC) da Universidade de São Paulo (USP) e Solange Aranha (IBILCE) da Universidade Estadual Paulista (UNESP) de São José do Rio Preto, coordenadores dos cursos de escrita acadêmica em inglês nos seus respectivos departamentos, e Ethel Schuster do *Northern Essex Community College* (EUA) instrutora do curso no ICMC, os nossos sinceros agradecimentos pela cooperação e apoio na coleta dos dados analisados neste estudo. Agradecemos também a todos os alunos participantes desses cursos por permitirem a utilização de seus textos nesta pesquisa e à FAPESP pelo total apoio a este projeto (Processos 2007/52405-3 e 2008/08963-4).

Introdução

Por ser a língua franca da comunidade científica internacional, o inglês para fins acadêmicos (*English for Academic Purposes* – EAP) tornou-se um requisito essencial para pesquisadores de todo o mundo (Hyland 2009, pp. 3-5; Swales e Feak 2009, pp. ix-xi). No entanto, para falantes não-nativos do inglês, principalmente se pesquisadores iniciantes, adquirir competência comunicativa em suas respectivas áreas de estudo é um grande desafio. Além de dominar as estruturas lexicais e sintáticas da língua inglesa, devem também reconhecer as principais convenções e características do discurso acadêmico, que pode apresentar diferenças relevantes em relação àquelas de seu idioma materno (Milton e Hyland 1999; Vold 2006; Davoodifard 2008).

Nesse contexto, cursos de escrita acadêmica em inglês e ferramentas de auxílio à escrita científica assumem um papel fundamental para estimular pesquisadores iniciantes a divulgar seus trabalhos de pesquisa em nível internacional. Citamos aqui os cursos de escrita acadêmica em inglês que têm sido oferecidos a pós-graduandos (mestrado e doutorado) por algumas instituições brasileiras, sendo elas: a Faculdade de Ciências Farmacêuticas (FCF), o Instituto de Física de São Carlos (IFSC) e o Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP), o Instituto de Biociências, Letras e Ciências Exatas (IBILCE) da Universidade Estadual Paulista de São José do Rio Preto (UNESP) e o Departamento de Genética e Evolução da Universidade Federal de São Carlos (UFSCar). Seja como disciplina optativa na grade curricular da pós-graduação ou cursos extracurriculares de curta duração, o principal objetivo desses cursos é orientar e auxiliar os alunos na escrita de artigos científicos em inglês. Como veremos na Seção 3, o *corpus* de estudo aqui analisado é composto por textos coletados nesses cursos.

Em relação a sistemas computacionais, diversas ferramentas de auxílio à escrita científica têm sido desenvolvidas para oferecer a escritores iniciantes suporte nos níveis lexical, sentencial e retórico. Enquanto o auxílio em nível lexical e sentencial visa ajudar na superação de problemas com convenções da língua, o nível retórico foca nas dificuldades relativas às convenções estruturais do gênero científico. Exemplos de sistemas desenvolvidos para o gênero científico são o AMADEUS (Fontana *et al.* 1993; Aluísio 1995; Aluísio e Oliveira 1995; Aluísio e Oliveira Jr. 1996; Aluísio e Gantenbein 1997; Aluísio *et al.* 2001), o Academic Writer (Broady e Shurville 2000), o BEAR (Narita 2000a, 2000b, *et al*. 2003), o Mover (Anthony e Lashkia 2003), o SciPo-Farmácia (Aluísio *et al*. 2005) e o CARE (Wu *et al.* 2006).

Destacamos aqui o SciPo-Farmácia (http://www.nilc.icmc.usp.br/scipo-farmacia), um conjunto de ferramentas desenvolvido no NILC (Núcleo Interinstitucional de Linguística Computacional) para auxiliar o usuário a compor a estrutura do texto. O sistema oferece exemplos de possíveis componentes da estrutura esquemática de um artigo científico (Fig. 1), a saber: resumo (*abstract*), introdução, metodologia, resultados, discussão e conclusão. O SciPo-Farmácia vem sendo usado com sucesso, desde sua criação em 2005, nos cursos de escrita acadêmica oferecidos na FCF da USP-SP e no IFSC da USP-São Carlos.

No presente estudo, nosso foco de interesse são resumos (*abstracts*) de artigos científicos, por desempenharem um papel importante em diversas atividades acadêmicas. Por exemplo, no processo de escolha dos artigos a serem lidos na íntegra, a decisão do leitor é geralmente tomada com base nas informações apresentadas no resumo (Swales e Feak 2009, p. 2; Wanner 2009, p. 157). Para Wanner (2009, p. 157), tal característica tem se tornado cada vez mais evidente, visto que a grande maioria dos periódicos, independente se eletrônicos ou não, disponibilizam os resu-

mos *on-line* para que os não-assinantes possam decidir comprar ou não o acesso a um determinado artigo. No caso de trabalhos submetidos para publicação, os resumos oferecem ao corpo editorial uma visão geral do estudo a ser avaliado. Embora o aceite do artigo dependa do mérito do trabalho como um todo, um resumo bem elaborado certamente servirá para enaltecer a impressão dos pareceristas sobre o trabalho apresentado (Swales e Feak 2009, p. 2). No caso de congressos, os resumos são um fator importante na decisão do comitê organizador em aceitar ou rejeitar um trabalho para apresentação (Swales e Feak 2009, p. 43).

Fig. 1: Scipo-Farmácia: tela principal
do componente resumo (abstract)

Os resumos são considerados textos independentes, com características lexicais, sintáticas, estilísticas e retóricas próprias. Tendem a focar nas informações mais essenciais e nos principais argumentos do trabalho, apresentando seu conteúdo e estrutura de forma sucinta. Com isso, são textos altamente elaborados, densos e compactos (Gledhill 2005, p. 41; Wanner 2009, p. 156), e têm a função de atrair a atenção dos leitores e convencê-los a ler o artigo completo (Hyland e Tse 2005; Hyland 2009, p. 70). Tais características fazem com que a tarefa de escrever um resu-

mo claro e eficiente não seja simples, mesmo para pesquisadores experientes e com um grande número de publicações (Swales e Feak 2009, p. xiii).

O presente trabalho concentra-se na investigação de resumos de artigos científicos escritos em inglês, das áreas de Ciências da Vida e da Saúde. Apoiando-se na Linguística de *Corpus* como base metodológica, o nosso principal objetivo é investigar as diferenças linguísticas entre resumos escritos por pós-graduandos (mestrandos e doutorandos) brasileiros e aqueles extraídos de artigos publicados por periódicos de excelência nas mesmas áreas de estudos. Focamos aqui nas partes dos resumos referentes à especificação do *propósito* e *resultado*. Dois aspectos são abordados, a saber:

número médio de segmentos por resumo, ou seja, orações ou sentenças usadas para expressar os componentes de *propósito* e *resultado*;

padronização léxico-gramatical, que se refere a combinações recorrentes de palavras, sem variação significativa de seus elementos. Nesse caso, além de repetição lexical (*the aim of this study is to*), consideramos também diferentes formas de um mesmo lema (*the aim of this study* ***is/was*** *to*), outros itens da mesma classe gramatical que se encaixem no contexto (*the* ***purpose*** *of this* ***paper*** *is to*), ou mesmo, a inserção ou ausência de determinado item (*the aim of this* ***in vitro*** *study was to*).

Trabalhos relacionados

O presente estudo enquadra-se em duas vertentes que, embora relacionadas, são totalmente distintas. Por um lado, buscamos explorar os aspectos linguísticos de resumos científicos escritos em inglês com o intuito de fornecer subsídios para os

cursos acima citados. De outro, esperamos também contribuir para o aprimoramento de ferramentas computacionais de suporte ao sistema SciPo-Farmácia como, por exemplo, um detector automático da estrutura do discurso. Nesta seção discutimos essas duas correntes, apresentando uma visão geral dos diversos trabalhos desenvolvidos até o momento e assim oferecendo ao leitor o contexto no qual este estudo se insere.

Estudos linguísticos

O uso de *corpora* para análise das características lexicais e sintáticas do discurso acadêmico tem crescido rapidamente. Alguns trabalhos focam exclusivamente na linguagem produzida por falantes nativos e/ou experientes[2] e examinam a prosa acadêmica em comparação a outros gêneros textuais (e.g. Biber *et al.* 1999, pp. 988-1036; Biber, Conrad e Cortes 2004). Outros investigam as principais semelhanças e diferenças entre gêneros científicos em uma mesma área de estudo e entre áreas (Gledhill 2005; Groom 2005; Peacock 2006, dentre outros). Existem ainda aqueles que se concentram em uma determinada seção de um artigo científico (e.g. Brett 1994; Gledhill 2000; Lim 2010).

Mais fortemente relacionados ao presente trabalho são aqueles que contrastam textos acadêmicos em inglês produzidos por pesquisadores iniciantes, geralmente não-nativos de inglês, com artigos científicos que tenham sido publicados por periódicos de excelência na área em análise. Por exemplo, Hyland

2. Neste estudo, optamos por empregar o termo falante nativo e/ou experiente (em vez de apenas falante nativo) pelo fato de que, como veremos a seguir, diversos estudos com base em *corpora* de textos acadêmicos utilizam artigos científicos publicados como padrão de comparação. Em outras palavras, toma-se como referência a linguagem aceita e utilizada por periódicos de excelência nas áreas analisadas, independente se produzida por falantes nativos ou não-nativos.

(2008) examina o uso de pacotes lexicais (sequências ininterruptas de palavras) em dissertações e teses escritas em inglês por mestrandos e doutorandos chineses. Já Hewings e Hewings (2002) analisam o uso do *it* antecipatório (sequências como *it has been shown that* ou *it is important to*, em que o pronome assume a posição de sujeito e a oração complementar funciona como o sujeito lógico da oração principal) em dissertações de mestrado escritas por estudantes não-nativos do inglês da Universidade de Birmingham. Aktas e Cortes (2008) examinam artigos científicos escritos por pós-graduandos não-nativos do inglês, cursando uma universidade americana, focando em substantivos usados como recursos de coesão (tais como *effect* e *result*), que englobam ou antecipam o significado do discurso ao seu redor.

No que diz respeito a resumos científicos em inglês, diferentes perspectivas têm sido adotadas na investigação de suas características linguísticas. Alguns trabalhos focam na organização textual, dando ênfase aos movimentos retóricos encontrados no texto (por exemplo, Hyland 2004, pp. 64-81; Lorés 2004). Por movimento retórico, entende-se uma unidade discursiva que reflete a função comunicativa de um determinado segmento do discurso (Swales 2004, p. 228). Outros estudos concentram-se na relação entre movimentos retóricos e características linguísticas (Salager-Meyer 1992; Pho 2008).

No entanto, pouco tem sido investido em análises de resumos escritos por pesquisadores iniciantes, cuja primeira língua não é o inglês. Uma exceção é o estudo de Hyland e Tse (2005) sobre o uso do pronome avaliativo *that* em resumos de dissertações e teses escritas em inglês por mestrandos e doutorandos chineses. Os dados foram coletados de um *corpus* de 465 resumos de seis áreas distintas: linguística aplicada, biologia, administração de empresas, computação, engenharia elétrica e administração pública. Como padrão de comparação, os autores usam um *corpus* com o mesmo número de resumos e das mesmas áreas

de estudos, extraídos de artigos científicos publicados por periódicos de relevância nas áreas analisadas. O pronome avaliativo *that* é amplamente usado nos dois *corpora*, visto que ambos fazem uso de estruturas impessoais, envolvendo, por exemplo, sujeitos abstratos (e.g. *the results suggest that*) e o *it* antecipatório (*it is very clear that*). No entanto, a frequência total do pronome é bem mais baixa no *corpus* dos alunos, pois esses parecem evitar o uso do pronome de primeira pessoa (*we believe that*).

Vale também mencionar nossos trabalhos anteriores (Dayrell e Aluísio 2008; Dayrell 2009a, 2009b, 2010, 2011a, 2011b) que serviram de base para o presente estudo. Tais trabalhos comparam características lexicais e sintáticas de resumos escritos em inglês por pós-graduandos brasileiros de diversas áreas e aqueles extraídos de artigos publicados por periódicos de excelência nas mesmas áreas de estudos. Foram identificadas diferenças relevantes entre os dois *corpora* tanto em relação ao sobreuso ou subuso de determinados itens lexicais quanto a padrões léxico-gramaticais empregados. O presente estudo visa, portanto, incrementar essa comparação e vincular tais estruturas à função retórica do segmento analisado, concentrando-se na especificação do *propósito* e *resultado* da pesquisa apresentada no resumo.

Sistemas de detecção automática da estrutura do discurso em textos científicos

Ferramentas de auxílio à escrita de artigos científicos, como é o caso do SciPo-Farmácia, partem do pressuposto que o texto acadêmico segue um padrão de estruturação que pode ser reproduzido automaticamente, o que favorece o desenvolvimento de sistemas que apoiem tanto a composição quanto a revisão do texto produzido. Neste contexto, a detecção automática da estrutura do discurso de artigos científicos é uma área de pesquisa

em franca expansão, contando com vários esforços paralelos na elaboração de diferentes tipos e abordagens de anotação.

Uma divisão importante é feita entre os sistemas que focam em resumos e aqueles para artigos completos. Nesse caso, os resumos ganham ênfase e diversos sistemas, geralmente baseados em aprendizado de máquina, têm sido propostos para a detecção automática dos componentes de sua estrutura esquemática (Anthony e Lashkia 2003; Mcknight e Srinivasan 2003; Shimbo *et al*. 2003; Ito *et al.* 2004; Feltrim *et al*. 2005; Yamamoto e Takagi 2005; Lin *et al.* 2006; Wu *et al.* 2006; *Genovês et al.* 2007; Ruch *et al.* 2007; Hirohata *et al.* 2008). A diferença entre eles se dá pelo esquema/modelo adotado e a abordagem de aprendizado de máquina escolhida.

Outro ponto é se os classificadores são treinados com *corpus* de uma área específica (por exemplo, Teufel 1999; Anthony e Lashkia 2003; Genovês *et al.* 2007) ou se trabalham com textos de áreas diferentes (e.g. Pendar e Cotos 2008). Finalmente, é também importante mencionar a unidade de anotação. No caso específico de resumos, a maioria dos classificadores (e.g. Anthony e Lashkia 2003; Genovês *et al.* 2007) opta por efetuar a anotação por sentença, ou seja, cada sentença é associada a um único movimento retórico. Já a anotação multirrótulo, que permite mais de uma etiqueta por sentença, ainda é um desafio (Dayrell *et al.* 2012).

Como veremos a seguir, neste trabalho, utilizamos o AZEA (*Argumentative Zoning for English Abstracts*, Genovês et al. 2007), cuja principal função é identificar e etiquetar os movimentos retóricos de resumos de artigos científicos em inglês. As seguintes etiquetas são consideradas: (i) <background> (contexto); (ii) <gap> (lacuna); (iii) <purpose> (propósito); (iv) <method> (metodologia); (v) <result> (resultado); (vi) <conclusion> (conclusão). Tal anotação automática foi posteriormente revisada manualmente (para mais detalhes sobre esses procedi-

mentos, veja a Seção *Metodologia*). A versão beta do AZEA está disponível on-line (http://www.nilc.icmc.usp.br/azea-web/).

Corpora usados neste estudo

Neste trabalho, utilizamos dois *corpora* de resumos de artigos científicos escritos em inglês das áreas de Ciências Biológicas e da Saúde. Um *corpus* consiste em resumos de artigos científicos escritos por pós-graduandos brasileiros (mestrandos e doutorandos) e o outro é composto por resumos extraídos de artigos científicos publicados em periódicos internacionais de referência nas mesmas áreas de estudo. Os critérios de seleção e categorização desses textos são apresentados em Dayrell (2011a) e, por conveniência, repetidos aqui de forma sucinta.

O *corpus* de resumos escritos por estudantes brasileiros (doravante EST) totaliza 138 resumos (26.516 palavras). Tais textos foram coletados nos cursos de escrita acadêmica em inglês oferecidos pela FCF/USP-SP em 2004, 2005 e 2009, pelo IBILCE/UNESP de São José do Rio Preto entre 2004 e 2006, e pelo Departamento de Genética e Evolução da UFSCar em 2005. Os resumos foram agrupados conforme o departamento que rege o programa de pós-graduação ao qual o aluno está filiado. Assim, o EST contempla as seguintes áreas: Odontologia, Ciências Farmacêuticas (Farmácia, Bioquímica e Engenharia de Alimentos), Biologia/Genética, Biofísica, Bioengenharia e Ciências Biomédicas (confira Tabela 1). A maior parte dos textos é das subáreas de Odontologia e Ciências Farmacêuticas, por contarem com o maior número de alunos nos cursos mencionados acima.

Os resumos que compõem o EST são aqueles entregues no início de cada curso, antes da incorporação de comentários e sugestões propostos pelos instrutores, colegas e orientadores.

Outro ponto a ser mencionado é que o nível de conhecimento de inglês dos alunos cuja produção é aqui investigada apresenta ampla variação, desde um domínio intermediário até o estágio avançado. Tal característica deve-se ao fato de que, para ingresso nesses cursos, exige-se apenas que o aluno tenha um conhecimento razoável de inglês e seja integrante de um programa de mestrado ou doutorado na universidade onde o curso é oferecido.

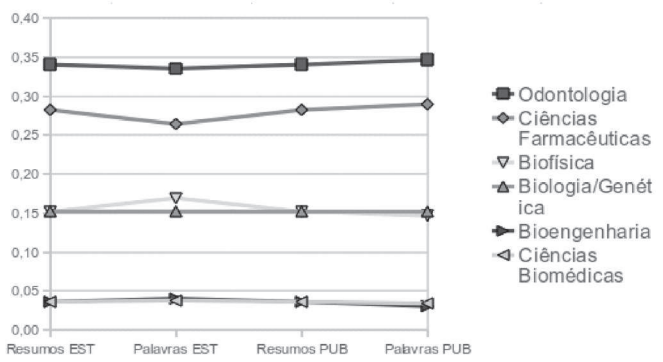
Como padrão de comparação ao EST, utilizamos um *corpus* de 690 resumos (150.247 palavras) extraídos de artigos em inglês, das mesmas áreas de estudo, publicados em periódicos internacionais com conceito 'A' no programa Qualis da CAPES. Para o *corpus* de resumos publicados (doravante PUB), o número de resumos de cada subárea foi estabelecido de forma a conter cinco vezes o número de resumos daquela mesma subárea no EST (confira Tabela 1). Tal decisão permite aos pesquisadores tirar conclusões mais sólidas sobre as preferências linguísticas de autores publicados.

Tabela 1: Composição do corpus de resumos escritos por estudantes brasileiros (EST) e do corpus de resumos publicados (PUB)

	EST		PUB	
Subáreas	Número de resumos	Número de palavras	Número de resumos	Número de palavras
Bioengenharia	5	1.072	25	4.612
Ciências Biomédicas	5	1.003	25	5.159
Odontologia	47	8.899	235	52.061
Ciências Farmacêuticas	39	7.011	195	43.533
Biofísica	21	4.491	105	22.041
Biologia/ Genética	21	4.040	105	22.841
Total	138	26.516	690	150.247

O interessante a mostrar aqui é que o EST e o PUB possuem a mesma distribuição de subáreas, tanto em termos de número de resumos quanto em número de palavras. A Fig. 2 abaixo mostra o percentual de textos e de palavras por área de estudo para cada um dos *corpora*. Por exemplo, se considerarmos a subárea odontologia, observamos que, no EST, corresponde a 35% do número total de resumos e 34% do número total de palavras. De forma semelhante, no PUB, o *subcorpus* de odontologia representa 34% dos resumos e 35% do número de palavras.

Fig. 2 Distribuição das subáreas do EST e do PUB em termos de número de resumos e número de palavras



Para seleção dos resumos publicados focamos naqueles de artigos de pesquisa (*research papers*), assumindo-se que os alunos estão mais propensos a escrever este tipo de artigo. Além disso, para agilizar a coleta, utilizamos apenas resumos disponíveis *on-line*, de periódicos tanto de circulação eletrônica ou impressa. A área de pesquisa em que o estudo se enquadra foi determinada de acordo com a filiação do primeiro autor ou da maioria dos autores, descartando-se os resumos de autores vinculados a departamentos de duas ou mais áreas aqui analisadas. A filiação dos autores também serviu de base para que fosse dada preferência a resumos de

autores afiliados a universidades de países de língua inglesa. No entanto, vale enfatizar aqui que isso não significa que os resumos incluídos no PUB sejam necessariamente escritos por falantes nativos do inglês. Como ressalta Williams (2006), a localização de uma universidade ou centro de pesquisa não diz nada a respeito da língua materna de seus pesquisadores, visto que a grande maioria trabalha com cientistas do mundo todo. O que se preza aqui é que o resumo seja parte de um artigo que tenha sido publicado por um periódico de excelência na área, ou seja, tenha passado pelo crivo de um comitê científico reconhecido. Tal abordagem já foi adotada por diversos estudos relacionados (Aktas e Cortes 2008; Hewings e Hewings 2002; Hyland 2008; Hyland e Tse 2005).

Por fim, vale mencionar que os resumos incluídos no PUB são em sua grande maioria de autoria múltipla enquanto que os resumos do EST foram, pelo menos em princípio, escritos por um único autor. Cada aluno contribui com um único resumo. Portanto, tentou-se diversificar o PUB o máximo possível em termos de autores selecionados.

Metodologia

A metodologia para extração dos dados deu-se em duas etapas. Em um primeiro momento, identificamos e etiquetamos os segmentos (oração ou sentença) referentes à especificação do *propósito* e *resultado* do estudo apresentado no resumo. A partir daí, procedeu-se à extração e análise dos dados de cada movimento.

Movimentos retóricos

Como mencionado anteriormente, a identificação e etiquetagem dos movimentos retóricos dos resumos foi, em um pri-

meiro momento, realizada automaticamente pelo sistema AZEA (Genovês *et al.* 2007). Seis possíveis etiquetas são consideradas, a saber: (i) <background> (contexto); (ii) <gap> (lacuna); (iii) <purpose> (propósito); (iv) <method> (metodologia); (v) <result> (*resultado*); e (vi) <conclusion> (conclusão).

Já de antemão, previa-se que seria necessário validar manualmente a anotação automática, por diversas razões. Primeiramente, porque o AZEA foi treinado com um *corpus* de 74 resumos de artigos científicos da área de Ciências Farmacêuticas, e estamos trabalhando aqui com *corpora* maiores, que contêm resumos de diversas áreas de estudo. Além disso, o AZEA estabelece uma relação de um-para-um entre sentenças e movimentos, e o presente trabalho pretende realizar a anotação multirrótulo, ou seja, a ocorrência de mais de um movimento retórico por sentença. Diante desses fatores, optamos por adotar um procedimento sistemático de validação manual dos dados, visto que este trabalho reporta um novo tipo de anotação.

O primeiro passo foi verificar se existia concordância entre anotadores humanos na realização da tarefa. Para tal, recorremos a três especialistas na anotação da estrutura esquemática de artigos científicos para etiquetagem de 10 resumos, depois da leitura de um manual com diretrizes de como proceder. Após a anotação, as principais divergências foram discutidas e o manual atualizado. Concluída essa fase de calibração, 34 resumos foram selecionados aleatoriamente do PUB e anotados separadamente pelos mesmos três especialistas.

A estatística Kappa (Carletta 1996) foi então aplicada para verificar o nível de concordância entre os especialistas. O cálculo é feito com base no número de respostas concordantes, ou seja, no número de casos em que os juízes propõem a mesma anotação. Para o nosso estudo, o valor do Kappa foi de 0.69 (N=529,k=3,n=21), sendo N o número de sentenças anotadas, k o número de juízes e n o número de padrões distintos para as sen-

tenças que incluem as várias combinações dos seis componentes retóricos. Embora esses valores sejam mais baixos do que aqueles encontrados em grande parte das tarefas de Processamento de Língua Natural (PLN), indicam boa concordância entre os anotadores, principalmente para a tarefa de anotação retórica. Uma vez discutidas e acordadas as divergências entre os anotadores, o manual de anotação foi revisado a fim de descrever a tarefa de anotação multirrótulo e ilustrar casos mais problemáticos, servindo de base para a anotação dos demais resumos, inclusive aqueles do EST.

Extração dos padrões léxico-gramaticais

Os *corpora* dos alunos e dos publicados foram manipulados separadamente, assim como os movimentos de *propósito* e *resultado* que foram tratados como sub*corpora* independentes. A seguinte nomenclatura é utilizada para nos referirmos aos *subcorpora*: EST-P e PUB-P para os *subcorpora* dos segmentos de *propósito* e EST-R e PUB-R para aqueles de *resultado*. Todos os procedimentos de extração dos dados foram efetuados por meio da ferramenta *WordSmith Tools (WST)* 5.0 (Scott 2007) e, para os cálculos estatísticos, utilizamos o pacote de programas estatísticos R (http://www.r-project.org/).

Inicialmente, efetuamos uma análise quantitativa das ocorrências dos movimentos de *propósito* e *resultado* no EST e no PUB, considerando-se a frequência total e por disciplina em cada *corpus*. Nessa análise, focamos na proporção de resumos com e sem os movimentos em questão e na comparação do número médio de segmentos de *propósito* e *resultado* por resumo.

Passamos então para a extração dos padrões léxico-gramaticais mais frequentemente empregados nos resumos dos alunos e/ou nos resumos publicados para especificação do *propósito* e *resultado* da pesquisa apresentada. Nosso ponto de partida fo-

ram sequências ininterruptas (pacotes lexicais)[3] de três palavras que ocorreram em, pelo menos, 3% dos resumos do *subcorpus* em questão. Por exemplo, o movimento *propósito* ocorre em 113 resumos do EST e, portanto, nosso ponto de partida nesse *subcorpus* foram os pacotes lexicais com no mínimo três ocorrências. Essa frequência mínima é arbitrária, tendo sido estabelecida apenas para delimitar o escopo do estudo e manter o volume de dados em um tamanho manejável.

Os padrões léxico-gramaticais aqui analisados foram extraídos a partir da análise das linhas de concordância de cada uma das sequências selecionadas, com o intuito de eliminar redundâncias e estabelecer os limites dos padrões. Por exemplo, no EST-P, as cinco sequências mais frequentes são: *the aim of*, *aim of this*, *of this study, study was to* and *this study was*. No entanto, 64% das ocorrências de *the aim of* referem-se à sequência *the aim of this study was to.*

Uma vez identificado um padrão, o próximo passo foi analisar as demais linhas de concordância a fim de identificar possíveis variações do mesmo. Nesse caso, os seguintes aspectos foram considerados:

(a) diferentes formas de um mesmo lema, como as ocorrências do verbo *to be* abaixo:

3. Para recuperar pacotes lexicais mais relevantes, além do valor mínimo de frequência, estipulamos também que o processamento deveria ser feito dentro dos limites da sentença. Ademais, o WST também recorre a uma série de cálculos estatísticos, sendo eles: coeficiente de dice, informação mútua, t-score, z-score and log likelihood, cujos pontos de corte são estabelecidos pelo usuário. Para efeitos deste estudo, adotamos os valores mínimos *default* do WST: -999,000 para o coeficiente dice; 3,000 para a informação mútua; 1,000 para o z-score; 3,000 para o log likelihood e 2,000 para o t-score.

01 The aim of this study **was** to assess the knowledge of dental students in the last year of colle

02 The aim of this study **is** to analyze the population structure of pink shrimp F. paulensis at th

03 The aim of this study **is** to relate a case-report of an 8-year-old child, with right primary se

(b) outros itens da mesma classe gramatical, ocorrendo no mesmo contexto. Por exemplo, no EST-P, ao permitirmos que a segunda lacuna do padrão acima possa variar (*the * of this study was/is/has been to*), recuperamos, além de *aim*, também *aims*, *purpose*, e *objective*:

04 The **aims** of this study were to assess age and gender differences in the growth of the nasophar

05 The **objective** of this study is to optimize cancer treatments and to explore uses and limitatio

06 The **purpose** of this study was to evaluate the tooth size discrepancies according to the Bolton

(c) pequenas variações lexicais devido à inserção ou ausência de determinado item, e também erros ortográficos ou gramaticais. Para ilustrar, citamos a inserção de *in vitro* na linha 07 abaixo e de *present* nas linhas 08 e 09. Na linha 09, observamos além da ausência do marcador de infinitivo *to*, onde incluímos o símbolo P para facilitar a visualização, erros gramaticais como o uso do *were* e *identified* em vez de *was* e *identify*.

07 The aim of this **in vitro** study was to compare the cutting effectiveness of an of an ultrasound

08 The objective of the **present** study was to evaluate the interference of GHF dentinal desensitiz

09 The objective of the **present** work were **P** identified the poly(hydroxyalkanoate) synthase (PHA

Esse procedimento de extração dos padrões foi executado para todas as sequências selecionadas nos quatro sub*corpora*, EST-P, PUB-P, EST-R e PUB-R, até que todas as linhas de concordância do movimento foram analisadas.

Como padrão léxico-gramatical, consideramos as combinações com uma frequência mínima de uma ocorrência a cada 100 segmentos do movimento em análise. Por exemplo, *we VERB that* aparece menos de uma vez (0,8) a cada 100 segmentos no EST-R, mas é aqui considerado como padrão por ocorrer 3,5 vezes a cada 100 segmentos no PUB-R. Esse ponto de corte é também arbitrário e foi adotado apenas para garantir que o padrão fosse recorrente em, pelo menos, um dos *subcorpora*.

Para representação dos padrões, utilizamos as seguintes convenções:

- itens opcionais aparecem entre parênteses. Por exemplo, o item *in* em *(in) the [presence] of*. Para efeito de simplificação, representamos apenas aqueles que ocorrem em, no mínimo, 50% das ocorrências do padrão em um ou ambos os *corpora*.
- os lemas são representados em versalete, como o verbo *to be* em *the aim of this study* BE *to*.
- agrupamento de itens do mesmo campo semântico aparecem entre colchetes. Por exemplo, *[presence]* em *(in) the [presence] of*, refere-se a *presence*, *absence*, *existence* e *lack*.
- as classes gramaticais aparecem em letras maiúsculas (e.g. *results VERB that*).

Comparação das frequências dos padrões léxico-gramaticais

Na comparação das frequências dos padrões léxico-gramaticais de cada movimento nos dois *corpora* optamos por aplicar

um teste de significância estatística,[4] com o intuito de aumentar a confiabilidade da interpretação dos dados. Em outras palavras, o analista pode inferir que as distribuições de frequências diferem entre o EST e o PUB. O que se pretende é verificar, por exemplo, se a diferença entre 47 ocorrências de *the [purpose] of [this] [study] be (to VERB)* no EST-P, que contém 123 segmentos, e 127 ocorrências desse mesmo padrão em um sub*corpus* com 626 segmentos (PUB-P) é estatisticamente significativa.

Para esse cálculo, recorremos ao qui-quadrado de Pearson e ao teste exato de Fisher quando o teste qui-quadrado pode não ser válido, i.e., em situações em que o número esperado de observações é menor do que 5 (Baroni e Evert 2009). A hipótese nula de cada teste realizado é a igualdade das proporções de segmentos com o referido padrão entre os resumos em EST e PUB. Em um teste estatístico de significância, o valor resultante (p) é interpretado em relação a um nível de significância pré-estabelecido. Nas ciências sociais, é comum adotar-se 0,05 como o ponto acima do qual a probabilidade de se aceitar um resultado como significativo quando este tenha ocorrido ao acaso é de no máximo 5 em 100 chances (5%) (Hinton 1995, p. 38; Oakes 1998, p. 9). Nesse caso, um valor de $p \geq 0{,}05$ indica que a diferença entre as frequências analisadas *não* é estatisticamente significativa. Já para um $p < 0{,}05$, o analista pode afirmar com confiança que as frequências observadas são diferentes, pois a diferença entre as frequências observadas nos dois *corpora* é estatisticamente significativa.

Vale mencionar que o cálculo estatístico é aplicado às frequências absolutas de cada padrão. No entanto, a fim de facilitar

4. Testes estatísticos de significância são aplicados para testar a hipótese nula de que não existe diferença significativa entre dois grupos de dados e as diferenças entre eles ocorrem puramente por acaso (Oakes 1998, p. 9). Tais testes indicam o grau de confiança com que um pesquisador pode aceitar ou rejeitar dada hipótese (Hinton 1995, p. 38).

a visualização dos dados, os gráficos apresentados na próxima seção mostram as frequências normalizadas, considerando-se o número de ocorrências a cada 100 segmentos do movimento analisado. Os padrões são apresentados conforme a frequência no EST, do mais para o menos frequente. Por fim, ressaltamos ainda que as frequências apresentadas a seguir incluem todas as variações encontradas para um dado padrão.

Análise de dados

Para cada um dos movimentos, *propósito* e *resultado*, apresentamos primeiro uma análise quantitativa do número total de segmentos no EST e no PUB e depois por subárea. Em um segundo momento, discutimos as semelhanças e diferenças entre os dois *corpora* em relação aos padrões léxico-gramaticais empregados. Todos os exemplos mostrados abaixo foram extraídos dos resumos dos alunos, a não ser quando explicitado.

Movimento de *Propósito*

Para o movimento de *propósito*, não foram encontradas diferenças significativas entre os dois *corpora* no que diz respeito a aspectos quantitativos, ou seja, nem em relação ao percentual de resumos em que o movimento ocorre nem em termos do número médio de segmentos por resumo.

A Tabela 2 abaixo mostra o número de resumos em cada *corpus* contendo um ou mais segmentos de *propósito* e também o percentual de ocorrência em relação ao número total de resumos. Observamos que a grande maioria dos resumos (84% no EST e 87% no PUB) contém, pelo menos, um segmento para especificar o *propósito*. Essa tendência ocorre independente da subárea, já que o percentual mais baixo de ocorrência do movimento é 74% nos resumos da subárea de Ciências Farmacêuticas do EST.

Tabela 2: Número e percentual de resumos com um ou mais segmentos de propósito nos corpora de resumos de alunos (EST) e daqueles publicados (PUB)

	EST		PUB	
Subáreas	Número de resumos com propósito	% de resumos com segmentos de propósito	Número de resumos com propósito	% de resumos com segmentos de propósito
Odontologia	41	87%	227	97%
Ciências Farmacêuticas	29	74%	171	88%
Biofísica	16	76%	82	78%
Biologia/ Genética	20	95%	77	73%
Bioengenharia	5	100%	20	80%
Ciências Biomédicas	5	100%	20	80%
Total	116	84%	597	87%

Ao considerarmos apenas os resumos com o movimento de *propósito*, observamos que ambos os *corpora* tendem a empregar, em média, um segmento (oração ou sentença) por resumo na especificação do *propósito* do estudo apresentado. Essa média é encontrada nos resumos de todas as subáreas, como mostrado na Tabela 3 abaixo.

Por outro lado, diferenças significativas foram encontradas na comparação dos padrões léxico-gramaticais empregados pelos alunos e pelos autores publicados para especificar o *propósito* da pesquisa apresentada no resumo.

Cinco padrões foram selecionados para análise. Nos padrões (i) a (iv), [*study*] engloba os seguintes itens lexicais: *study*, *work*, *investigation*, *project*, *clinical trial*, *research, paper*, *review*, *article, chapter* and *manuscript*.

Tabela 3: Número médio de segmentos de propósito nos resumos de alunos (EST) e naqueles publicados (PUB)

	EST-P		PUB-P	
Subáreas	Número total de segmentos	Número médio de segmentos por resumo	Número total de segmentos	Número médio de segmentos por resumo
Odontologia	43	1.0	239	1.1
Ciências Farmacêuticas	33	1.1	183	1.1
Biofísica	20	1.3	84	1.0
Biologia/ Genética	20	1.0	85	1.1
Bioengenharia	6	1.2	25	1.3
Ciências Biomédicas	5	1.0	23	1.2
Total	127	1.1	639	1.1

(i) **the [purpose] of [this] [study] be (to VERB)**

[*Purpose*] inclui *purpose(s)*, *aim(s)*, *objective(s)*, *goal* e *intent*; [*this*] refere-se a *this*, *the present, the current, the* e *our*.

01 **The purpose of this study was to evaluate** complete caries removal time (CRT) and patient accep

02 **The aim of the present study was to follow** the nucleolar cycle in the spermatogenesis of Rattu

03 **The objective of this work is to investigate** the in vitro toxicity of gold and silver nanopart

O verbo *to be* é sempre seguido por um verbo no infinitivo, com exceção de sete ocorrências no EST-P (15% do total de ocorrências do padrão), como nos exemplos abaixo:

04 **The aim of this paper is** about the treatment CL I squeletictal - CL III dentoalveolar (with ov

05 **The aim of this work is** valueted the gene expression of the RECK and MMPs in three cell lines

06 With this, **the objective of this study is** the microsatellite prospecting of Rimpaenaeus constr

(ii) [this] **[study] ([aim] to/was [conducted] to) VERB**

[this] refere-se a *this*, *the present, the current* ou apenas *the*. [*study*] aparece como sujeito da oração que especifica o objetivo da pesquisa apresentada. Três estruturas são consideradas:

(a) [this] **[study] VERB**

07 **This work describes** the development and validation of a method for the determination of aceton

08 **This paper presents** the isolation and characterization of the camptosemin , a protein of legum

09 **The present study evaluated** the sensory acceptability of a frozen dessert that associates a ce

(b) [this] **[study] [aim] to VERB**

Um verbo indicando 'intenção' aparece entre [*study*] e o verbo que especifica o *propósito*. Nesse caso, [*aim*] engloba os seguintes verbos *aim*, *intend* e *seek*.

10 Thus, **this work aims to offer** organic milk as a potential raw material for the manufacture of

11 **This project aims to construct** a nonlinear optical microscope to acquire images of fluorescenc

12 **This study intends to examine** the changes in bones and teeth caused by use of Herbst appliance

(c) [this] **[study] was [conducted] to VERB**

Como no tipo (b), um outro verbo, nesse caso na voz passiva, é inserido entre [*study*] e o verbo que indica o *propósito*. [*Conducted*] refere-se a *conducted, designed, performed, organized, made* e *considered*. Essa estrutura ocorre apenas no PUB, de onde os exemplos abaixo foram extraídos:

13 **This study was conducted to determine** the effects of the inflammatory cytokine interferon-(IFN

14 **This study was performed to test** some common food ingredients and additives for their effect o

15 **This study was designed to evaluate** the suitability of an aqueous solution of voriconazole sol

(iii) **(here/in [this] [study]) we ([seek] to) VERB (the)**

No padrão (iii), o pronome de primeira pessoa do plural (*we*) é usado como sujeito da oração que especifica o *propósito* da pesquisa sendo apresentada. [this] refere-se a *this* e *the present*. Dois tipos de estrutura são aqui considerados:

(a) **(here/in [this] [study]) we VERB (the)**

16 **Here we report the** major structural differences between the subtypes B and C. The protease was

17 **In this work we show the** fluorescence anisotropy of the binding of two different constructions

18 **We describe** here **the** treatment protocol applied to a male patient with desquamative lesions in

Incluímos aqui os casos em que os autores fazem referencia à hipótese sendo testada, o que indiretamente representa o objetivo do estudo. Tais ocorrências foram encontradas apenas no PUB (8% das ocorrências) e incluem a sequência *we hypothesize that*:

19 **We hypothesize that** differences in DNA methylation of specific CpG dinucleotides between forme

20 **We hypothesize that** ATP and NPY can regulate regulate catecholamine synthesis in chromaffin ce

21 **we hypothesized that** salivary mRNA stability is mediated by ARE-binding proteins in human sali

(b) **(here/in [this] [study]) we [seek] to VERB (the)**

[seek] engloba os verbos *seek* e *set out*. Essa estrutura aparece no PUB apenas (5% das ocorrências), de onde os exemplos foram coletados:

22 **We sought to examine** key points of intersection between cannabinoid receptor 1 (CB1) signallin

23 **We seek to determine** whether cell membranes contain sensors that trigger a downstream response

24 **Here we set out to define the** mechanism by which UPEC enters host cells by investigating four

(iv) **in [this] [study] (we)**

[this] refere-se a *this*, *the present, the current* ou *our*, como em:

25 **In this study, we** have determinated the dissociation constants (Kd) of the RXR-different HREs

26 **In the present study**, the genotoxic and antigenotoxic potential of Ab extracts were investigat

27 have been studied **in this project**. APRT catalyzes the conversion of adenine into adenosine mon

(v) **the (ADJECTIVE) effect(s) of**

Esse padrão inclui a sequência *the effect(s) of* assim como os casos em que um ou mais adjetivos modificam o substantivo. O padrão é pouco frequente no EST e, portanto, os exemplos abaixo foram extraídos do PUB:

28 In this study, we examined **the effects of** syringolin A (SylA) and glidobactin A (GlbA) as well

29 sthetized rats to determine **the** cardiovascular **effects of** acute repeated intravenous administr

30 ne the role of a5 nAChRs in **the** initial pharmacological **effects of** nicotine, nicotine reward,

Na grande maioria dos casos, além do complemento nominal iniciado pela preposição *of*, encontramos um segundo complemento iniciado pela preposição *on*, que especifica o que é afetado:

31 gated **the effects of** pH **on** nail permeability and the transport of ions such as sodium (Na) and

32 gated **the effects of** specific K+ channel inhibitors on basal and oestrogen-stimulated prolifer

33 luate **the effects of** TGA and TGA-MABP pretreatment **on** an orthopedic allograft involving menisc

Nesse padrão, incluímos também alguns poucos casos em que os dois complementos aparecem na ordem inversa, ou seja, o complemento iniciado por *on* vem antes do complemento iniciado por *of*:

34 study were to determine **the effect on** the flexural properties of FRC and metal post materials

35 ted an investigation of **the effect on** enamel and dentin **of** the interaction among bleaching, er

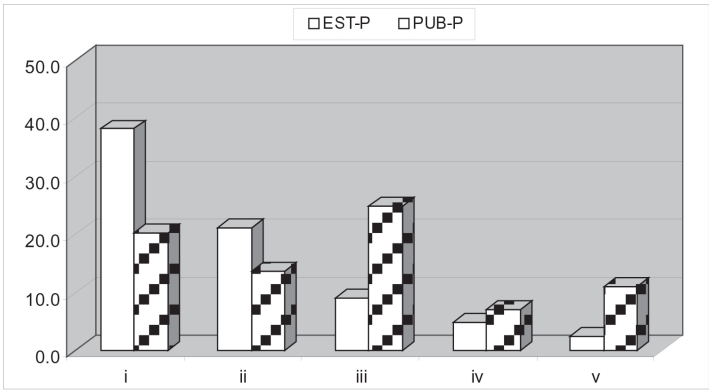
36 f practice assumptions, **the** simulated **effect on** profitability **of** treating patients covered by

A Figura 5 mostra as frequências normalizadas (a cada 100 segmentos) dos cinco padrões léxico-gramaticais supracitados, no EST-P e no PUB-P.

Como mostrado na Figura 3, o padrão (i) (e.g *the aim of this study is to VERB*) é mais frequente no *corpus* dos alunos. Embora também empregado pelos autores publicados, não ocorre de forma tão recorrente no PUB-P e a diferença entre as frequências totais do padrão nos dois *corpora* é estatisticamente

significativa ($p < 0,05$). De forma semelhante, o padrão (ii) (e.g. *the present study evaluated*) é também mais frequente no EST-P do que no PUB-P, o que é comprovado estatisticamente ($p < 0,05$). Já os resumos publicados demonstram uma preferência marcante pelo padrão (iii), e.g. *here we report the*, que é bem menos frequente no EST-P, sendo a diferença entre as frequências nos dois *corpora* estatisticamente significativa ($p < 0,05$). Tais *resultado*s corroboram a sugestão de outros estudos (confira Hyland e Tse 2005; Hyland 2008; Dayrell 2010, 2011a) de que aprendizes tendem a optar pela impessoalidade, como no caso dos padrões (i) e (ii) acima. Ao mesmo tempo, parecem evitar dispositivos linguísticos que explicitem a posição do pesquisador como agente, como é o caso dos pronomes de primeira pessoa (padrão (iii)).

Figura 3: Frequências normalizadas (a cada 100 segmentos) dos cinco padrões léxico-gramaticais recorrentes na especificação do propósito da pesquisa



O padrão (iv), e.g. *in this study we*, é o único para o qual a diferença entre as frequências nos dois *corpora* não é estatisticamente significativa. O interessante a notar aqui é que o pronome da primeira pessoa do plural (*we*) representa 55% das

ocorrências no PUB-P e apenas 33% das ocorrências no EST-P. Em outras palavras, ao assumirem a posição de agentes com o uso do pronome *we*, os autores publicados tendem a chamar a atenção para o trabalho sendo apresentado com a sequência *in [this] [study]*. Já para os resumos dos alunos, não podemos afirmar o mesmo, visto que encontramos pouca regularidade lexical no entorno de *in [this] [study]* no EST-P (veja, por exemplo, as linhas 25-27 acima).

Por fim, o padrão (v) – e.g. *the effect of* –, ocorre de forma recorrente no PUB-P, mas parece ser negligenciado pelos alunos, sendo a diferença entre as frequências nos dois *corpora* estatisticamente significativa ($p < 0,05$). Além disso, observamos que, na grande maioria das ocorrências no PUB-P, o padrão (v) é precedido pelos padrões (i) (26% dos casos), (ii) (25%) ou (iii) (24%), como ilustram as linhas 37-39 abaixo. Essa tendência não é observada nas ocorrências do padrão (v) no EST-P, o que sugere pouca familiaridade dos alunos com a co-ocorrência desses padrões.

37 The purpose of this study was to determine the effect of miniscrew implant orientation on the
38 This study examines the effects of distraction on the condyle in a large animal model by expli
39 We have investigated the effect of deletions of a post-synaptic density, disc large and zo-1 pr

Movimento de *Resultado*

No que diz respeito ao movimento de *resultado*, diferenças significativas são observadas entre os dois *corpora*, tanto em relação ao percentual de resumos em que o movimento ocorre quanto em termos do número médio de segmentos por resumo.

O primeiro ponto a ser mencionado é que a proporção de resumos que incluem a descrição dos *resultado*s é bem mais alta no PUB (97%) do que no EST (65%) (Tabela 4). Essa tendência é também observada quando a análise é feita por subárea. Uma possível explicação para essa diferença talvez esteja relacionada ao fato de que as pesquisas desenvolvidas por grande parte dos alunos ainda estivesse em andamento no momento da produção do resumo e, portanto, eles ainda não tinham os resultados da pesquisa. Outra possível justificativa seria a pouca familiaridade dos alunos com o gênero alvo que, como mostram os resumos publicados, parece privilegiar a descrição dos resultados da pesquisa apresentada.

Tabela 4: Número e percentual de resumos com um ou mais segmentos de resultado nos resumos de alunos (EST-R) e daqueles publicados (PUB-R)

	EST		PUB	
Subáreas	Número total de resumos com resultados	% de resumos com segmentos de resultado	Número total de resumos com resultados	% de resumos com segmentos de resultado
Odontologia	30	64%	232	99%
Ciências Farmacêuticas	28	72%	192	98%
Biofísica	10	48%	100	95%
Biologia/ Genética	17	81%	104	99%
Bioengenharia	2	40%	19	76%
Ciências Biomédicas	3	60%	24	96%
Total	90	65%	671	97%

Outra diferença relevante entre os dois *corpora* é que o número de segmentos é muito mais alto no PUB-R do que no EST-R (Tabela 5). Embora tal diferença fosse esperada, visto que o PUB é cinco vezes maior que o EST, ela excede qualquer estimativa, pois o número de segmentos é 10 vezes maior no

PUB-R. Além disso, ao considerarmos apenas os resumos que apresentam os *resultado*s da pesquisa, observamos que o número médio de segmentos no EST-R tende a ser a metade do seu correspondente no PUB-R. Tais dados podem ser um indicativo de que os alunos não apresentam os *resultado*s de suas pesquisas de forma tão detalhada quanto os autores publicados o fazem. Também podem estar relacionados com a sugestão acima de que os alunos talvez ainda não tenham os resultados da pesquisa no momento da redação do resumo.

Tabela 5: Número médio de segmentos de resultado nos resumos de alunos (EST) e naqueles publicados (PUB)

	EST-R		PUB-R	
Subáreas	Número total de segmentos de resultado	Número médio de segmentos por resumo	Número total de segmentos de resultado	Número médio de segmentos por resumo
Odontologia	84	2.8	919	4.0
Ciências Farmacêuticas	73	2.6	866	4.5
Biofísica	18	1.8	344	3.4
Biologia/ Genética	47	2.8	422	4.1
Bioengenharia	3	1.5	45	2.4
Ciências Biomédicas	9	3.0	100	4.2
Total	234	2.6	2,696	4.0

Diferenças relevantes também são observadas em relação aos padrões léxico-gramaticais identificados nos dois *corpora*, sendo eles:

(i) **(the) [results] VERB that**

[results] refere-se tanto a *results* quanto a *findings*, sendo que esse último ocorre apenas no PUB, correspondendo a 7% das ocorrências:

01 second to the fourth year. **The results showed that** 92 % of the pupils in the second year, 95,1

02 d MgCl[2] in this binding. Our **result shows that** TR DL can bind DNA better than TR DBD for pal

03 nce of the storage period. **The results indicated that** the addition of a whey protein-derived f

(ii) **BE not PARTICIPLE**

O padrão (ii) inclui ocorrências da voz passiva na forma negativa, como em:

07 actose. The main fatty acids **were not influenced** by the type of milk and higher amounts of con

08 seminiferous tubules light, **were not revealed** by Feulgen reaction, indicating indicating that

09 C ($p > 0.05$). This difference **was** also **not observed** when the two subgroups were compared regar

(iii) **there BE (no/not) (statistically) (significant) difference(s) (between)**

Esse padrão engloba todas as ocorrências do verbo *there BE*, tanto na forma afirmativa quanto negativa, seguidas do item lexical *difference(s)*. Tanto o advérbio *statistically* quanto o adjetivo *significant* aparecem de forma recorrente no padrão:

10 and cells entrapment. Statistically **there was no difference between** control group and citric a

11 and Class II (92,06). **There were no significant differences between** genders. When all the grou

12 groups. In FG, **there was statically [sic] significant difference between** QL1 and QL2. We did n

Encontramos também alguns casos no PUB-R que incluem o determinante *a*, ou o adjetivo *few* ou um advérbio, como em:

13 pared to moist enamel and there was not **a** significant difference in the SFL of OBP on dry and

14 at 1 and 30 d of storage. There were **also** no significant differences in modes of failure betwe

15 sed preparation strategy. There were **few** differences regarding geographic regions. More than 7

(iv) **(in) the [presence] of**

[presence] refere-se a: *presence*, *absence*, *existence* e *lack*. Nesse padrão, incluímos também ocorrências da sequência *(in) the presence and absence of*:

04 nd tissue preparations. **In the presence of** 1 μM N/OFQ a high-affinity GDP binding site was als

05 pEC50 for phenylephrine **in the absence of** desmethylimipramine was greater than control after b

06 ssion of these enzymes and **the lack of** significant induction in the pediatric samples.

(v) **BE PARTICIPLE to (be)**

O padrão (v) inclui todos os casos em que a voz passiva é seguida de um verbo no infinitivo. PARTICIPLE refere-se a dois tipos de verbos, comumente encontrados no inglês acadêmico (Carter e McCarthy 2006, p. 790): (i) verbos que reportam os resultados da pesquisa, tais como *found*, *shown*, *demonstrated*, *proved* e *reported*; (ii) verbos cognitivos, como *estimated*, *suspected, perceived, observed* e *predicted*. Os exemplos a seguir foram extraídos do PUB, por ser esse padrão pouco frequente no EST:

16 age canal wall thickness **was found to be** thinner (33%) on the distal, suggesting a "danger zon

17 revalence of dental pain **was estimated to be** 30.4%, and care-seeking in those reporting pain w

18 etry, the magnetic field **is shown to restrict** the radial spread of secondary electrons to a sm

(vi) **we [find] that**

[find] refere-se a *find*, *show*, *demonstrate*, *report, observe, confirm, verify* e *determine*. Os exemplos abaixo foram extraídos do PUB-R:

19 n microscopy, **we found that** Sterne strain 7702 spores were able to adhere to and subsequently

20 In this paper **we show that** Map triggers filopodia formation by activating Cdc42; expression of

21 e first time, **we demonstrate that** both 19S and 20S components of the 26S proteasome complex ar

Nesse caso, incluímos também ocorrências com um advérbio ou uma locução adverbial, como em:

22 s unexplored. **We now demonstrate that** proteasome stringently controls transcription of inflamm

23 helial cells. **We showed for the first time that** internalized spores were able to survive and t

24 degradation. **We furthermore observed that** LeTx promoted the formation of Processing bodies, r

Foram descartados todos os casos em que o verbo principal pede um objeto direto, assim, a oração iniciada pelo pronome *that* deixa de assumir o papel de predicado e passa a complementar o objeto. Os exemplos abaixo foram extraídos do PUB:

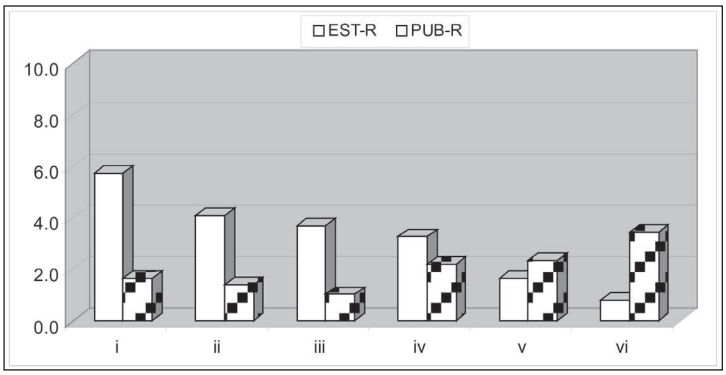
25 Asians. We have identified **21 SNPs** that act as markers of the inverted, i.e., H2, haplotype. T

26 ganics. We also describe **approaches** that have been taken to describe chemical mixtures as a wh

27 people. We rediscovered **genes** that were known to play a role in the ER-stress response and unc

Apresentamos na Figura 4 as frequências normalizadas (a cada 100 segmentos) dos padrões léxico-gramaticais identificados no EST-R e no PUB-R.

Figura 4: Frequências normalizadas (a cada 100 segmentos) dos seis padrões léxico-gramaticais recorrentes na especificação de resultados



Em relação aos padrões léxico-gramaticais empregados no componente *resultado*, observamos que ambos *corpora* recorrem frequentemente a predicados verbais iniciados pelo pronome *that*. No entanto, enquanto os alunos preferem usar o item lexical [*results*] (resultados) como sujeito da oração (padrão (i), e.g. *the results showed that*), os autores publicados optam pelo pronome de primeira pessoa do plural (padrão (vi), e.g. *we find that*) (Figura 4). Em ambos os casos, a diferença entre as frequências nos dois *corpora* é estatisticamente significativa ($p < 0,05$). Portanto, assim como na especificação do *propósito*, observamos que também para os *resultado*s, os alunos parecem preferir sujeitos abstratos tais como os "*resultado*s" (padrão (i)) a se posicionarem frente ao argumento apresentado (padrão (vi)).

Já no caso da voz passiva, é interessante notar que o padrão (ii) (*BE not PARTICIPLE, e.g. were not revealed*) é mais recorrente no EST-R do que no PUB-R ($p < 0,05$), ou seja, os alunos empregam a forma negativa da voz passiva mais frequentemente que os autores publicados. Tal dado talvez seja um indício de que esses últimos tendam a dar mais ênfase a resultados positivos. Por outro lado, o padrão (v) (*BE PARTICIPLE to (be)*, e.g. *were shown to be*) ocorre em uma frequência semelhante nos dois *corpora*, pois a diferença entre sua ocorrência nos dois *corpora* não é significativa ($p \geq 0,05$). No entanto, o padrão (v) é o segundo mais frequente no PUB-R (Figura 4) e, em relação aos outros padrões identificados no EST-R, é um dos menos frequentes. Uma possível explicação para isso pode ser o fato de o padrão não ter correspondentes diretos no português, daí o desconforto dos alunos em utilizá-lo.

No caso do padrão (iii), e.g. *there were **no** significant differences between*, a grande maioria das ocorrências nos dois *corpora* refere-se à forma negativa do padrão: 78% no EST-R e 79% no PUB-R. O padrão é bem mais frequente no EST-R do que no PUB-R, o que é confirmado pelo teste estatístico ($p < 0,05$). Portanto, novamente, os alunos parecem a dar mais atenção à não confirmação de determinado *resultado* do que os publicados o fazem. Outra possível razão para a mais alta incidência no *EST-R* talvez seja que os autores publicados, em vez de repetirem esse padrão, recorram a outras estruturas para expressar a mesma ideia.

Por fim, o padrão (iv) – (in) the [presence] of – parece ser usado de forma semelhante nos dois *corpora*. Em termos de frequência, a diferença entre os dois *corpora* não é significativa do ponto de vista estatístico ($p \geq 0,05$). Também em relação aos itens lexicais que compõem o padrão, o item *presence* é o mais frequente nos dois *corpora*: 75% no EST-R e 62% no PUB-R.

Considerações Finais

Neste trabalho, analisamos o número de segmentos (orações ou sentenças) e a padronização léxico-gramatical usados na especificação do *propósito* e *resultado* em resumos de artigos científicos escritos em inglês, das áreas de Ciências da Vida e da Saúde. As seguintes subáreas foram consideradas: Odontologia, Ciências Farmacêuticas (Farmácia, Bioquímica e Engenharia de Alimentos), Biologia/Genética, Biofísica, Bioengenharia e Ciências Biomédicas. O nosso principal objetivo era contrastar resumos escritos por pós-graduandos brasileiros com aqueles extraídos de artigos publicados por periódicos internacionais de referência nas mesmas áreas de estudos.

Em termos quantitativos, no que se refere à especificação do *propósito* da pesquisa apresentada no resumo, não foram encontradas diferenças significativas entre os dois *corpora*. Já em relação aos *resultado*s da pesquisa, enquanto os resumos publicados parecem considerá-los altamente relevantes e mencioná-los em praticamente todos os resumos, os alunos muitas vezes não os incluem. Alem disso, os autores publicados tendem a apresentá-los de forma muito mais detalhada do que os alunos. O estudo também aponta uma série de diferenças relevantes entre os dois *corpora* no que se refere à frequência dos padrões léxico-gramaticais empregados, tanto em relação à especificação do *propósito* quanto dos *resultado*s.

O presente trabalho representa, portanto, uma contribuição importante para o ensino de EAP, visto que os resultados podem ser aplicados diretamente em sala de aula e também em ferramentas de auxílio à escrita científica, como é o caso do Scipo-Farmácia. Ao chamarmos a atenção dos aprendizes para os principais aspectos que distinguem sua linguagem daquela de falantes nativos e/ou experientes, contribuímos para que percebam

o que sinaliza a não naturalidade da linguagem por eles produzida. Ao tomarem consciência das convenções e características mais tipicamente usadas por suas comunidades acadêmicas, poderão certamente aprender a escrever de forma mais eficaz.

Além disso, vale mencionar que este estudo abre espaço para o aprimoramento do sistema AZEA. Ao validarmos manualmente a anotação dos componentes retóricos de 690 resumos, possibilitamos que a ferramenta seja novamente treinada com base em um *corpus* maior do que feito anteriormente e dividido por subárea de estudo. Além de aumentar a precisão da ferramenta, podemos assim adequá-la às características dos textos reais, cujas sentenças podem refletir mais de um movimento retórico. Resultados preliminares de um estudo inédito sobre o tópico são apresentados em Dayrell *et al.* (2012). Ademais, intencionamos ainda abordar as peculiaridades de cada área de estudo, criando vários classificadores, um para cada grande área de pesquisa. Além dessas aplicações, não podemos esquecer que um classificador preciso é uma ferramenta de grande valia para linguistas de *corpus* que desejam basear seus estudos em grandes *corpora* textuais.

Referências bibliográficas

AKTAS, R. N. e CORTES, V. (2008). "Shell nouns as cohesive devices in published and ESL student writing." *Journal of English for Academic Purposes*, vol. 7, n.º 1, Elsevier Ltd., pp. 3-14.

ALUÍSIO, S. M. (1995). *Ferramentas para auxiliar a escrita de artigos científicos em inglês como língua estrangeira*. Tese de Doutorado. São Carlos: Instituto de Física de São Carlos, Universidade de São Paulo.

ALUÍSIO, S. M.; BARCELOS, I.; SAMPAIO, J. e OLIVEIRA JR, O. N. (2001). "How to learn the many unwritten 'rules of the game' of the academic discourse: a hybrid approach based on critiques and cases to support scientific writing." *IEEE International Conference on Advanced Learning Technologies*, vol. 1, Madison/EUA, pp. 257-260.

ALUÍSIO, Sandra Maria e GANTENBEIN, R. E. (1997) "Towards the application of systemic functional linguistics in writing tools." *Proceedings of the International Conference on Computers and their Applications*, vol.1, Tempe/EUA, pp. 181-185.

ALUÍSIO, Sandra Maria e OLIVEIRA JR, O. N. (1995). "A case -based approach for developing writing tools aimed at non-native English users." *Proceedings of the 1st International Conference – ICCBR-95*. *Lecture Notes in Artificial Intelligence*, vol. 1010. Berlin: Springer-Verlag, pp. 121-132.

_____. (1996) "Detailed schematic structure of research papers introductions: an application in support-writing tools." *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, vol. 1, Espanha, pp. 141-147.

ALUÍSIO, Sandra Maria; SCHUSTER, E.; FELTRIM, V. D.; PESSOA JR, A. e OLIVEIRA JR, O. N. (2005) "Evaluating scientific abstracts with a genre-specific rubric." *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005)*, vol.1. Amsterdã/Holanda, 18 a 22 de julho, pp. 738-740.

ANTHONY, L. e LASHKIA, G. V. (2003). "Mover: A machine learning tool to assist in the reading and writing of technical papers." *IEEE Transactions on Professional Communication*, vol. 46, nº. 1, pp. 185-193.

BARONI, M. e EVERT, S. (2009). "Statistical methods for corpus exploitation", *in*: LÜDELING, A. e M. KYTÖ (eds.), *Corpus linguistics: an international handbook*, vol. 2. Berlin: Mouton de Gruyter, pp. 777-802.

BIBER, D.; CONRAD, S. e CORTES, V. (2004). "'If you look at…': lexical bundles in university teaching and textbooks." *Applied Linguistics*, vol. 25, n.º 3. Oxford: Oxford University Press, pp. 371-405.

BIBER D.; JOHANSSON, S.; LEECH, G.; CONRAD, G. e FINEGAN, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson Education.

BRETT, P. (1994). "A genre analysis of the results section of sociology articles." *English for Specific Purposes*, vol. 13, n.º 1, Elsevier Ltd., pp. 47-59.

BROADY, E. e SHURVILLE, S. (2000). "Developing academic writer: designing a writing environment for novice academic writers", *in*: BROADY, E. (ed.) *Second Language Writing in a Computer Environment*. Londres: CILT, pp. 131-151.

CARLETTA, J. (1996). "Assessing agreement on classification tasks: the kappa statistic." *Computational Linguistics*, vol. 22, n.º 2. Michigan: The MIT Press, pp. 249-254.

CARTER, R. e MCCARTHY, M. (2006). *Cambridge grammar of English: a comprehensive guide*. Cambridge: Cambridge University Press.

DAVOODIFARD, M. (2008). "Functions and hedges in English and Persian academic discourse: effects of culture and the scientific discipline." *ESP Across Cultures*, n.º 5. Bari: Edizioni B.A. Graphics, pp. 23-48.

DAYRELL, C. (2009a). "Sense-related verbs in English scientific abstracts: a corpus-based study of students' writing." *ESP Across Cultures*, n.º 6. Bari: Edizioni B.A. Graphics, pp. 61-78.

_____. (2009b). "Lexical bundles in English abstracts: a corpus -based study of published and non-native graduate writing." *Proceedings of the CL2009 (5th Corpus Linguistics Conference)*, Liverpool/Grã-Bretanha, 21-23 julho. Disponível em: http://ucrel.lancs.ac.uk/publications/cl2009.

_____. (2010). "Sense-related verbs in published and student writing: a corpus-based study of English and Portuguese abstracts", *in*: XIAO, R. (ed.) *Using corpora in contrastive and translation studies*. Newcastle: Cambridge Scholars Publishing, pp. 486-507.

_____. (2011a). "*Corpora* no ensino do inglês acadêmico: padrões léxico-gramaticais em *abstracts* de pós-graduandos brasileiros",

in: VIANA, V. e TAGNIN, S. (eds.) *Corpora no Ensino de Língua Estrangeira*. São Paulo: HUB Editorial, pp. 137-172.

_____. (2011b). "Anticipatory '*it'* in English abstracts: a corpus-based study of non-native student and published writing", *in*: GOŹDŹ-ROSZKOWSKI, S. (ed.) *Explorations across Languages and Corpora*. Łódź Studies in Language. Frankfurt am Main: Peter Lang, pp. 581-598.

DAYRELL, C. e ALUÍSIO, S. (2008). "Using a comparable corpus to investigate lexical patterning in English abstracts written by non-native speakers." *Proceedings of LREC 2008 (6th Language Resources and Evaluation Conference), Workshop Building and Using Comparable Corpora*. Marraquexe/ Marrocos, 26 de maio a 1 de junho. Disponível em: http://www.lrec-conf.org/lrec2008/.

DAYRELL, C.; A. CANDIDO Jr.; LIMA, G.; MACHADO Jr., D.; COPESTAKE, A.; FELTRIM, V. D.; TAGNIN, S. e ALUÍSIO, S. M. (2012). "Rhetorical Move Detection in English Abstracts: Multi-label Sentence Classifiers and their Annotated Corpora." *Proceedings of LREC 2012 (8th International Conference on Language Resources and Evaluation)*. Istambul/Turquia, 21 a 27 de maio..

FELTRIM, V. D.; TEUFEL, S.; NUNES, M. G. V. e ALUÍSIO, S. M. (2005). "Argumentative zoning applied to critiquing novices' scientific abstracts." *Computing Attitude and Affect in Text: Theory and Applications*, vol. 1. 1ª ed. Dordrecht/The Netherlands: Springer, pp. 159-170.

FONTANA, N.; ALUÍSIO, S. M.; DE OLIVEIRA, M. C. F. e OLIVEIRA JR., O. N. (1993). "Computer assisted writing: applications to English as a foreign language." *CALL (Computer Assisted Language Learning Journal)*, n.º 6. Londres/Nova York: Routledge, pp. 145-161.

GENOVÊS JR, L. C.; FELTRIM, V. D., DAYRELL, C. e ALUÍSIO, S. M. (2007). "Automatically detecting schematic structure components of English abstracts: building a high accuracy classifier for the task." *Proceedings of the International Conference RANLP'2007*, *Workshop on Natural Language Processing for Educational Resources*, Borovetz/Bulgária, 26 de setembro, pp. 23-29.

GLEDHILL, C. (2000). "The discourse function of collocation in research article introductions." *English for Specific Purposes*, vol. 19, n.º 2, Elsevier Ltd., pp. 115-135.

_____. (2005). *Collocations in science writing*. Tübingen: Gunter Narr Verlag.

GROOM, N. (2005). "Pattern and meaning across genres and disciplines: an exploratory study." *English for Academic Purposes*, vol. 4, Elsevier Ltd., pp. 257-277.

HEWINGS, M. e HEWINGS, A. (2002). "'It is interesting to note that…': a comparative study of anticipatory 'it' in student and published writing." *English for Specific Purposes*, vol. 21, n.º 4, Elsevier Ltd., pp. 367-383.

HINTON, P. R. (1995) *Statistics explained: a guide for social science students*. Londres/Nova York: Routledge.

HIROHATA, K.; OKAZAKI, N.; ANANIADOU, S. e ISHIZUKA, M. (2008). "Identifying sections in scientific abstracts using conditional random fields." *Proceedings of the IJCNLP 2008 (3rd International Joint Conference on Natural Language Processing)*, Hyderabad/Índia, 7 a 12 de janeiro, pp. 381-388.

HYLAND, K. (2004). *Disciplinary Discourses: social interactions in academic writing*. Michigan: Michigan University Press.

_____. (2008). "Academic clusters: text patterning in published and postgraduate writing." *International Journal of Applied Linguistics*, vol. 18, n.º1. Chichester: John Wiley and Sons, pp. 41-61.

_____. (2009). *Academic discourse*. Londres/Nova York: Continuum.

HYLAND, K. e TSE, P. (2005). "Hooking the reader: a corpus study of evaluative *that* in abstracts." *English for Specific Purposes*, vol. 24, n.º 2, Elsevier Ltd., pp. 123-139.

ITO, T.; SIMBO, M.; YAMASAKI, T. e MATSUMOTO, Y. (2004). "Semi-supervised sentence classification for medline documents." *IPSJ SIG Technical Report*, vol. 2004-ICS-138, pp. 141-146.

LIM, J. Miin-Hwa (2010). "Commenting on research results in applied linguistics and education: a comparative genre-ba-

sed investigation." *English for Specific Purposes*, vol. 9, n.º 4, Elsevier Ltd., pp. 280-294.

LIN, J.; KARAKOS, D.; DEMNER-FUSHMAN, D. e KHUDANPUR, S. (2006). "Generative content models for structural analysis of medical abstracts." *Proceedings of the HLT/NAACL 2006, Workshop on Biomedical Natural Language Processing (BioNLP'06)*, Nova York/EUA, 5 a 9 de junho, pp. 65-72.

LORÉS, R. (2004). "On RA abstracts: from rhetorical structure to thematic organization." *English for Specific Purposes*, vol. 23, n.º 3, Elsevier Ltd., pp. 280-302.

MCKNIGHT, L. e ARINIVASAN, P. (2003). "Categorization of sentence types in medical abstracts." *AMIA 2003 Symposium Proceedings*, pp. 440-444.

MILTON, J.; HYLAND, K. (1999). "Assertions in students' academic essays: a comparison of English NS and NNS student writers", *in*: BERRY, R.; ASKER, B. e HYLAND, K. (eds.) *Language analysis, description and pedagogy*. Hong Kong: Language Centre HKUST, pp. 147-161.

NARITA, M. (2000a). "Constructing a tagged e-j parallel corpus for assisting Japanese software engineers in writing English abstracts." *Proceedings of the LREC 2000 (2nd Language Resources and Evaluation Conference*), Atenas/Grécia, 30 de maio a 02 de junho, pp. 1187-1191.

NARITA, M. (2000b). "Corpus-based English language assistant to Japanese software engineers." *Proceedings of the MT-2000 (Machine Translation and Multilingual Applications in the New Millennium*), 20 a 22 de novembro, Exeter/Inglaterra, pp. 24-1 – 24-8.

NARITA, M.; KUROKAWA, K. e UTSURO, T. (2003). "Case Study on the development of a computer-based support tool for assisting Japanese software engineers with their English writing needs." *IEEE Transactions on Professional Communication*, vol. 48, n.º 3, IEEE Professional Communication Society, pp.194-209.

PEACOCK, M. (2006). "A cross-disciplinary comparison of boosting in research articles." *Corpora*, vol. 1, n.º 1. Edinburgo: Edinburgh University Press, pp. 61-84.

PENDAR, N. e COTOS, E. (2008). "Automatic identification of discourse moves in scientific article introductions." *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus/EUA, 19 de junho, pp. 62-70.

PHO, P. D. (2008). "Research article abstracts in applied linguistics and educational technology: a study of linguistic realizations of rhetorical structure and authorial stance." *Discourse Studies*, vol. 10, n.º 2. Los Angeles/Londres/Nova Deli/Singapura: SAGE Publications, pp. 231–250.

OAKES, M. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

RUCH, P.; BOYER, C.; CHICHESTER, C.; TBAHRITI, I.; GEISS-BUHLER, A.; FABRY, P.; GOBEILL, J.; PILLET, V.; REBHOLZ-SCHUHMANN, D.; LOVIS, C. e VEUTHEY, A. L. (2007). "Using argumentation to extract key sentences from biomedical abstracts." *International Journal of Medical Informatics*, vol. 76, Elsevier Ltd., pp. 195- 200.

SALAGER-MEYER, F. (1992) "A text-type and move analysis of verb tense and modality distribution in medical English abstracts." *English for Specific Purposes*, vol. 11, n.º 2, Elsevier Ltd., pp.93-115.

SCOTT, M. (2007). *WordSmith Tools*. Versão 5. Oxford: Oxford University Press.

SHIMBO, M.; YAMASAKI, T. e MATSUMOTO, Y. (2003). "Using sectioning information for text retrieval: a case study with the medline abstracts." *Proceedings of the 2nd International Workshop on Active Mining (AM'03)*, Maebashi/Japan, 28 de outubro, pp. 32-41.

SWALES, J. M. (2004). *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.

SWALES, J. M. e FEAK, C. B. (2009). *Abstracts and the writing of abstracts*. Michigan: University of Michigan Press.

TEUFEL, S. (1999). *Argumentative Zoning: Information Extraction from Scientific Text*. Tese de Doutorado. Edinburgo: School of Cognitive Science, University of Edinburgh.

VOLD, E. T. (2006). "Epistemic modality markers in research articles: a cross-linguistic and cross-disciplinary study." *International Journal of Applied Linguistics*, vol. 16, n.º 1. Chichester: John Wiley and Sons, pp. 61-87.

WILLIAMS, G. (2006). "Challenging the native-speaker norm: a corpus-driven analysis of scientific usage", *in*: BARNBROOK, G.; DANIELSSON, P. e MAHLBERG, M. (eds.) *Meaning texts: the extraction of semantic information from monolingual and multilingual corpora*. Londres/Nova York: Continuum, pp. 115-127.

WU, J.; CHANG, Y.; LIOU, H. e CHANG, J. S. (2006) "Computational analysis of move structures in academic abstracts." *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, Sydney, 17 a 21 de julho, pp. 41-44.

WANNER, A. (2009). *Deconstructing the English passive*. Berlin: Mouton de Gruyter.

YAMAMOTO, Y. e TAKAGI, T. (2005). "A sentence classification system for multi-document summarization in the biomedical domain." *Proceedings of the International Workshop on Biomedical Data Engineering (BMDE2005)*, Tóquio/Japão, 3 e 4 de abril, pp. 90-95.