DOI: 10.14393/DL29-v11n2a2017-6

Português para Fins Acadêmicos sob o aporte da Linguística de *Corpus* e do Processamento de Linguagem Natural

Portuguese for Academic Purposes with the contribution of Corpus Linguistics and Natural Language Processing

Cristina Becker Lopes Perna*
Lucelene Lopes**
Lucas Zambrano Rollsing***

RESUMO: O presente artigo intenta apresentar um projeto de interface entre Linguística de *Corpus* e Processamento de Linguagem Natural (PLN) em andamento na Pontifícia Universidade Católica do RS. Tal projeto consiste na exploração, via *corpus* escrito, de teses e dissertações em Linguística oriundas do Programa de Pós-Graduação em Letras da referida instituição de ensino superior, por meio de um software chamado ExATO (LOPES, 2012). A partir dessa ferramenta produzida em PLN, podemos ter em mãos uma série de recursos linguísticos que permitem a prossecução da nossa análise, tais como, a detecção de Hierarquia de Conceitos, Listas de Termos Extraídos, Concordanciador de Termos e Nuvens de Conceitos a fim de embasarmos uma proposta de ensino de Português como Língua Adicional, com vistas à proficiência dentro do gênero acadêmico por alunos não-falantes de Língua Portuguesa. Trazemos uma série de resultados já disponibilizados pelo ExATO, os quais fomentam a discussão acerca dos temas visualizados a partir dos dados obtidos, bem como fomentam a interdisciplinaridade necessária entre Linguística e Ciência da Computação, no que concerne à descrição e explicação do viés pragmático da LP em grande escala dentro dos limites da esfera discursiva universitária.

ABSTRACT: The article hereby intends to present an interface project between Corpus Linguistics and Natural Language Processing (NLP) in progress at the Pontifical Catholic University of RS. This project consists in the exploration, through a written corpus, of theses and dissertations in Linguistics from the Postgraduate Program in Letters of said institution of Higher Education, through software called ExATO (LOPES, 2012). From this tool produced in NLP, we can have in hand a series of linguistic resources that allow the continuation of our analysis, such as the detection of Concept Hierarchy, Extracted Term Lists, Concordance of Terms and Concepts Clouds in order to base a proposal of teaching of Portuguese as an Additional Language, with a view to the proficiency within the academic genre by non-Portuguese speakers. We bring a series of results already available by ExATO, which foment the discussion about the subjects visualized from the obtained data, as well as foment the necessary interdisciplinarity between Linguistics and Computer Science, as far as the description and explanation of the pragmatic bias of the LP on a large scale, within the limits of the university discursive register.

*Doutora em Linguística, Escola de Humanidades – Programa de Pós-Graduação em Letras, PUCRS.
**Doutora em Ciência da Computação, PUCRS.
***Mestrando em Linguística, Escola de Humanidades – Programa de Pós-Graduação em Letras, PUCRS.

PALAVRAS-CHAVE: Linguística de *Corpus*. Processamento de Linguagem Natural. Português para Fins Acadêmicos.

KEYWORDS: Corpus Linguistics. Natural Language Processing. Portuguese for Academic Purposes.

1. Considerações iniciais

O presente artigo visa apresentar os resultados iniciais de um estudo sendo desenvolvido na interface entre dois programas de Pós-Graduação da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), a saber, o Programa de Pós-Graduação em Letras (PPGL), na sua área de concentração em Linguística, através do Grupo de Pesquisa sobre o Uso e Processamento de Língua Adicional (UPLA)[1], e o Programa de Pós-Graduação em Ciência da Computação (PPGCC), através do Grupo de Processamento em Linguagem Natural[2]. Cabe mencionar que o grupo UPLA opera sob a perspectiva pragmática para descrição e análise de suas pesquisas, utilizando o aporte da Linguística de *Corpus* (LC). Essa união epistemológica assumida pelo grupo proporciona condições de compilar e descrever *corpora* de Língua Portuguesa, seja em registro acadêmico, como veremos ao longo do presente trabalho, seja em qualquer esfera discursiva.

Essa profícua relação é bem apontada por Finatto, Lopes e Ciulla (2015, p. 43), ao atestarem que o diferencial da LC: "[...] é o estudo da língua em uso, verificado o uso sempre em grande escala, com apoio informatizado e tratamento estatístico", e acrescentam que: "Seu principal objetivo é a descrição de usos". Logo, através da metodologia de análise de *corpora* proposta pela LC em conjunto com a Pragmática, obtemos dados reais sobre o registro acadêmico, em condições de extrairmos deles considerações que possam nos levar a um entendimento mais profundo sobre os gêneros do discurso.

Nosso objeto de estudo, portanto, são textos acadêmicos produzidos no âmbito do PPGL, dentro da área de concentração em Linguística, conforme descreveremos na Metodologia. As pesquisas sobre o Português para Fins Acadêmicos (PFA) vêm crescendo nos últimos anos e, diante dessa nova demanda, é necessário que o ensino e a aprendizagem da língua portuguesa se tornem mais especializados, de acordo com os propósitos específicos para os quais a língua será utilizada.

[1] http://dgp.cnpq.br/dgp/espelhogrupo/0705653235733182
[2] http://dgp.cnpq.br/dgp/espelhogrupo/9398077368852737

Ressaltamos que este artigo está sob a égide das demais pesquisas já iniciadas no PPGL-PUCRS na área de PFA, vinculadas ao Grupo UPLA, o qual almeja ampliar os estudos sobre Português para Fins Acadêmicos (PFA), ao desenvolver *corpora* de estudo e metodologias de análise, para futuramente serem empregados como uma pedagogia de ensino de Português como Língua Adicional (PLA) para alunos estrangeiros em mobilidade acadêmica que necessitem de proficiência em PFA.

Com isso em mente, Molsing e Perna (2015, p.3), extraem e postulam as diretrizes para o trabalho em PFA, quando atestam que o Grupo UPLA: "[...] têm adotado, por enquanto, construtos teóricos similares àqueles que guiam a pesquisa na área de ESP (*English for Specific Purposes*), fazendo as adaptações específicas para a língua portuguesa, para a cultura brasileira e para a realidade da educação superior brasileira[3]".

Tendo delineado nossa motivação, organizamos nosso artigo a partir da seguinte estrutura: na seção 2 apresentaremos os construtos da Linguística de *Corpus* e do Processamento de Linguagem Natural, que dão suporte à análise dos textos da produção acadêmica do PPGL-PUCRS; na seção 3 apresentaremos a metodologia empregada na compilação do *corpus* de estudo, denominado CorpAcad, e como se deu o processamento pelo software ExATO; na seção 4 faremos a discussão dos dados obtidos através do software em questão e, por fim, na seção 5 apresentaremos as considerações finais.

2. A Linguística de *Corpus* e o Processamento de Linguagem Natural

As definições das áreas de Linguística de *Corpus* (LC) e de Processamento de Linguagem Natural (PLN) são sujeitas a certas discrepâncias segundo a fonte consultada. No entanto, é um consenso entre as principais fontes da área (e.g., BIBER et al., 1998; MANNING; SCHTZE, 1999; DALE et al., 2000; KENNEDY, 1998; MITKOV, 2003; TEUBERT; CERMÁKOVÁ, 2007) que se trata de duas áreas de pesquisa de origens distintas, mas que colaboram estreitamente em diversas iniciativas práticas e teóricas. Logo, nos parágrafos que se seguem a este, conceituaremos ambas as áreas e as caracterizaremos, a fim de contextualizarmos a presente pesquisa.

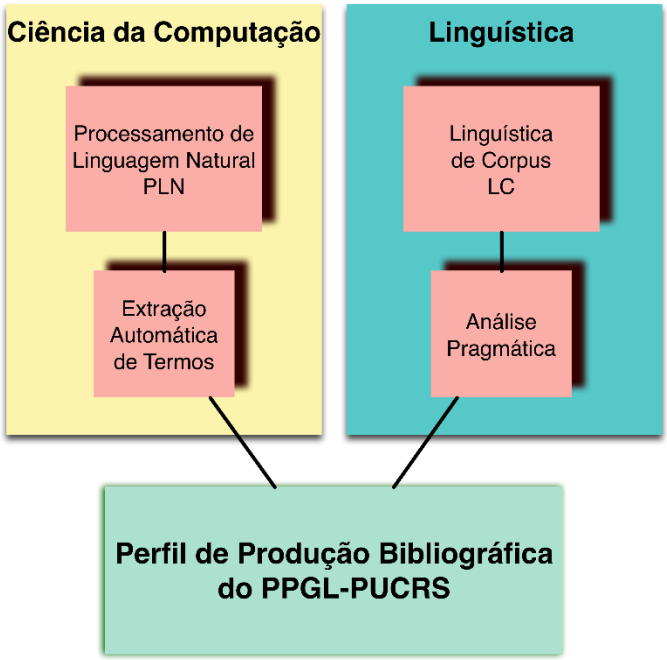
[3] Texto original: *"[...] we have adopted, for the time being, similar theoretical constructs as those that guide the research in the area of ESP, making specific adaptations for the Portuguese language, Brazilian culture and the reality of Brazilian higher education."*

O PLN é "uma área de Ciência da Computação, mais especificamente, da área de Inteligência Artificial, que estuda o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas, ou linguagens naturais" (LOPES; VIEIRA, 2010, p. 184). A partir disso, podemos afirmar que o PLN tem por propósito buscar o desenvolvimento de técnicas, algoritmos e, consequentemente, programas de computador que possam tratar informações complexas expressas por linguagens que possuem todas as características que são encontradas na expressão humana, seja ela escrita ou falada. O PLN se destaca de outras áreas usuais da Ciência da Computação que tratam linguagens ditas formais, onde os conceitos não se prestam à interpretação subjetiva ou à existência de ambiguidade, como é o caso das linguagens de programação ou mesmo das expressões matemáticas ou de lógica de primeira ordem.

Já a LC é uma área da Linguística, mais especificamente, da área de Linguística Aplicada, que não pretende desenvolver técnicas, algoritmos ou programas capazes de analisar a linguagem natural dos textos acadêmicos por nós reunidos, e sim valer-se destas ferramentas teóricas e práticas para promover um estudo sistemático sobre grandes volumes de expressões humanas. Conceitualmente, essas expressões também podem ser faladas, porém é mais frequente limitar-se a textos escritos. Esses conjuntos de textos são denominados *corpora* (no singular: *corpus*) e representam amostras de expressões relativas a entidades específicas. Dessa forma, um *corpus* é uma maneira prática bastante eficaz e usualmente livre de preconceitos de descrever, através de exemplos de utilização, uma determinada entidade, seja ela uma língua, um domínio, um grupo ou uma pessoa. Dessa forma, o que usualmente define um *corpus* é o critério com o qual ele é construído. Caso construa-se um *corpus* com os textos de um determinado autor, teremos uma amostra do tipo de discurso utilizado por este autor. Consequentemente, a análise deste *corpus* permitirá chegar a conclusões sobre este autor de forma sistemática e sem o viés de uma análise tradicional.

Apresentadas as duas áreas, a Figura 1 apresenta esquematicamente a relação entre as áreas envolvidas (LC e PLN). Especificamente, detalha-se que dentro de PLN utiliza-se como suporte a Extração Automática de Termos, enquanto que dentro de LC utiliza-se a análise pragmática de *corpora*.

Figura 1 – Organização das áreas de pesquisa que embasam este estudo.



Fonte: elaborada pelos autores.

De acordo com o esquema da Figura 1, podemos ter um resumo sobre a organização teórica do trabalho em questão. Essa estruturação nos conduzirá a análise e descrição do perfil da produção intelectual do Programa de Pós-Graduação em Letras da PUCRS, a fim de dar continuidade aos estudos que se debruçam sobre o português acadêmico e, por consequência disto, sobre o Português como Língua Adicional.

Na seção seguinte, explicitaremos como se procedeu com a extração automática de termos relevantes do *corpus* do PPGL, a partir da ferramenta de software ExATO.

3. Metodologia

Tomamos a Linguística de *Corpus* (LC) como a abordagem necessária para compilarmos o CorpAcad, bem como para a sua descrição e análise pragmática; conjuntamente a isso, utilizamos o software elaborado em PLN, a saber, o Extrator Automático de Ontologias[4] (ExATO), para o processamento dos dados. Escolhemos o software ExATO, desenvolvido por Lopes desde sua tese de doutorado (2012), no PPGCC da Pontifícia Universidade Católica do

[4]A ferramenta ExATO na sua versão atual é uma evolução que inclui diversas otimizações do ponto de vista de informática, bem como a capacidade de estender o seu escopo de aplicação a corpora com textos em Inglês, enquanto que a versão inicial, denominada EχATO$_{LP}$ (Extrator Automático de Ontologias em Língua Portuguesa), só era capaz de extrair termos relevantes de corpora com textos em Português (LOPES et al., 2009).

Rio Grande do Sul, pois se trata de um software robusto e moderno, capaz de processar grandes quantidades de dados em linguagem natural. O ExATO é um extrator automático de termos relevantes (sintagmas nominais/verbais) que permite traçar um panorama do *corpus* de estudo, através da detecção de Hierarquia de Conceitos e outros recursos linguísticos, tais como Listas de Termos Extraídos, Concordanciador de Termos e Nuvens de Conceitos. Ele processa *corpora* em Português e Inglês, utilizando métodos linguísticos e estatísticos da área de PLN (LOPES; FERNANDES; VIEIRA, 2016).

De acordo com Lopes (2012, p. 33), conceito "é uma generalização associada a uma ideia, podendo ter várias manifestações textuais". Devido a um conjunto de regras heurísticas, o ExATO nos fornece com precisão sintagmas nominais representativos para o domínio em questão, sob a forma de conceitos. Esses conceitos são gerados em listas ordenadas por grau de relevância, e podem ser visualizados de diferentes formas, como por exemplo, nuvens de conceitos ou árvores hiperbólicas. No contexto deste trabalho serão utilizadas as listas textuais e as nuvens de conceitos, cada qual com seu propósito, uma vez que a lista de conceitos abarca um número muito mais expressivo de sintagmas nominais, enquanto que a nuvem nos fornece somente os 150 sintagmas mais relevantes, mas os disponibiliza graficamente.

De um ponto de vista metodológico, o processo de extração realizado pelo ExATO pode ser detalhado em cinco etapas bem distintas:

- Anotação sintática e estrutural do texto, através de um *PoS-tagger* que anota, dentre outras informações, a classe gramatical de cada palavra do texto (substantivo, preposição, pronome, etc.) e um *parser* que anota as funções gramaticais desempenhadas por conjuntos de palavras (sujeito, verbo, objeto, etc.);
- Detecção e refinamento de sintagmas nominais através de heurísticas de base linguística que permitem fornecer termos de alto valor terminológico;
- Computação de um índice de relevância para cada sintagma nominal detectado através de um método de forte base estatística;
- Aplicação de um ponto de corte às listas de termos extraídos e organizados segundo seu índice de relevância de forma a escolher dentre todos os termos aqueles que são suficientemente relevantes para serem considerados conceitos do *corpus* alvo;

- Geração de recursos linguísticos sofisticados que vão desde uma simples lista de conceitos até sofisticadas visualizações como é o caso, neste artigo, das nuvens de conceitos.

A anotação sintática e estrutural dos textos foi feita através do *parser* PALAVRAS[5] (BICK, 2000), que executa tanto a tarefa de *PoS-tagging* quanto à estruturação gramatical das frases. A detecção de sintagmas nominais, por sua vez, segue um processo proposto por Lopes e Vieira (2012) e engloba várias heurísticas, que vai desde a eliminação de símbolos até a detecção de sintagmas implícitos. O cálculo do índice de relevância está baseado na utilização de *corpora* contrastantes com o uso do índice *tf-dcf* (*term-frequency, disjoint corpora frequency*), uma métrica com resultados mais efetivos que as atuais alternativas existentes (LOPES; FERNANDES; VIEIRA, 2016). Uma vez calculados os índices de relevância dos termos, estes podem ser ordenados pelo índice e então se aplica uma política de ponto de corte que escolhe quais termos possuem os valores de *tf-dcf* altos o suficiente para serem considerados relevantes para o domínio, ou seja, conceitos do *corpus* alvo (LOPES; VIEIRA, 2015). Finalmente, a geração de recursos linguísticos inclui um concordanciador, além das listas, nuvens e árvores citadas anteriormente. Considerando o material utilizado nesse artigo, a seguir trazemos exemplos das formas de saída do ExATO, a fim de ilustrarmos os resultados iniciais já obtidos.

Para a construção do CorpAcad, selecionamos o período de produção bibliográfica de 2006 a 2016 do PPGL da PUCRS. Já contamos com mais de 200 textos, dentre dissertações e teses que versam sobre Linguística, finalizando assim a etapa de compilação. Podemos conferir na Tabela 1 abaixo, a quantia de teses e dissertações, o número de sentenças, *tokens*[6] e termos (sintagmas nominais) de acordo com cada ano, a quantia de conceitos segundo o ponto de corte estabelecido pelo ExATO e a representação percentual desses conceitos no CorpAcad como um todo. Cabe salientar que o período selecionado contempla o início do arquivamento digital das teses e produções dos programas de pós-graduação da PUCRS. A partir do ano de 2006, podemos acessar as produções intelectuais da universidade, ainda que nesse mesmo ano, nem todas as produções tivessem sido armazenadas virtualmente. No entanto, conseguimos coletar

[5]Mais detalhes em: http://visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf.
[6]"Uma única unidade linguística, na maioria das vezes uma palavra, embora dependente do sistema de codificação utilizado, uma única palavra pode ser dividida em mais de um token, por exemplo, em inglês "*he's*" (*he* + *'s*)". Traduzido e adaptado de Baker, Hardie e McEnery (2006, p. 159).

o que estava disponível para cada ano, sendo crescente o número de trabalhos disponíveis até os dias de hoje.

Tabela 1 – Dados Gerais do CorpAcad.

Ano	Teses	Dissert.	Sentenças	*Tokens*	Termos/SN	Conceitos Ponto de Corte	Porcent.
2006	5	3	13483	347225	18803	1640	9%
2007	5	13	30758	700224	35971	3275	9%
2008	9	13	44695	1017660	49479	4796	10%
2009	10	13	45936	992365	48288	4636	10%
2010	6	12	33619	738295	37509	3680	10%
2011	9	13	44181	954984	48884	4534	9%
2012	5	13	35246	828666	46699	4096	9%
2013	8	12	32390	817770	43070	3777	9%
2014	11	13	43825	959624	50284	4460	9%
2015	14	16	75614	1722420	84942	5439	6%
2016	0	3	4106	96355	5985	309	5%
Geral	82	124	403853	9175588	333429	8056	2%

Fonte: elaborada pelos autores.

O CorpAcad já processado através do ExATO pôde nos propiciar, para cada ano de produção bibliográfica do PPGL, as nuvens e listas de conceitos relevantes para a área da Linguística, assim como uma nuvem e uma lista geral para o período total observado. Trazemos, como exemplo, as nuvens de conceitos (Figura 2) e a lista dos 20 termos mais relevantes (Quadro 1) referentes respectivamente a cada um dos anos de 2006 a 2016 e ao período de onze anos de produção.

Figura 2 – Nuvens de conceitos por ano.

Fonte: extraído da análise no programa ExATO.

Quadro 1 – Lista dos 20 termos mais relevantes por ano.

2006	2007	2008	2009
Discurso	Verbo	Criança	Criança
Enunciados	Discurso	falante	Oração
oração	Crianças	Consciência Fonológica	Verbos
Verbos	Aprendizes	Verbo	Adjetivos
Alocutário	sinonímia	Expressões Idiomáticas	Consciência fonológica
enunciação	falante	Implicatura	Sintagma
Encadeamentos	Processamento Auditivo	Orações	Discurso
Pronomes	mulher	aprendiz	Síndrome de Down
oração relativa	enunciado	Substantivos	falante
interlocutor	oclusiva dental	Discurso	Vogais
criança	Pragmática	palavra gramaticais	Advérbios
Enunciador	item lexical	enunciado	enunciado
Ducrot	palatalização de oclusiva dental	Pragmática	Brasil
Espanhol	Implicatura	Editoriais	João
r forte	UG	Informante	dialeto
Onset	Grupo Experimental	Grupo Experimental	Guarda chuva
r fraco	orelha	Epêntese	Encadeamentos
sentido de enunciado	GJTS	léxico	postônica
Retórica	paciente	GC	Flor
encadeamento argumentativo	oração	Português Brasileiro	pronome

2010	2011	2012	2013
Consciência Fonológica	Criança	Discurso	Crianças
Discurso	Discurso	Enunciados	Discurso
enunciado	Verbo	Emoção	Enunciados
criança	Narrativa	Metáforas	HR
encadeamento	Mãe	falante	Tempos Verbais
falante	enunciado	interlocutor	Surdos
Informante	Latim	Encadeamento	pesquisadora
Vogais	nomeação	signo	crianças bilíngues
aprendiz	traço distintivo	corpus	Enunciação
conjunção	DFE	ato	Signo
Negação	menino	Diálogo	falante
enunciação	Vogal	input	Compreensão de textos
paciente	Enunciação	Serviços Sexuais	Acentos
verificação acústica	Adultos	Implicatura	Informante
Sílaba	Sílaba	mulher	Elisão
Ducrot	Menina	enunciação	interlocutor
narrativa	inteligibilidade	ANL	encadeamento
Redução vocálica	gravidade de DFE	Assibilação	Zona rural
Francês	pesquisadora	aluna	ANL
verificação perceptual	alocutário	Reescrita	Brasil

2014	2015	2016	2006-2016
Discurso	Discurso	Discurso	Discurso
Criança	Criança	ANL	Criança
Enunciado	Enunciado	Ducrot	Enunciado
Enunciação	interlocutor	diálogo	Verbo
Adjetivo	implicatura	asno	Falantes
Falantes	Signo	transposição didática	Consciência Fonológica
ANL	falante	língua adicional	Encadeamento
Encadeamento	Bush	produção oral	interlocutor
implicatura	encadeamento	competência interacional	Enunciação
AP	Enunciação	ensino de leitura	Oração
signo	TR	Alunos	Implicatura
Provérbio	diálogo	Alteridade	Signo
Teoria dos Topoi	Narrativas	entidade linguística	Aprendizes
Linguística	Charge	CLG	Brasil
Ducrot	ANL	Semântica Argumentativa	ANL
Sílaba	Rima	Saber Científico	Vogais
Vogais	Poema	Bloco Semântico	Narrativa
Aprendizes	Brasil	encadeamento argumentativo	Ducrot
Oração	Igreja	Ensino Fundamental	Diálogo
Acento	Atos	noção de valor	Ato

Fonte: elaborado pelos os autores.

4. Discussão dos resultados iniciais

Nas nuvens da Figura 2, como dito anteriormente, constam os 150 conceitos mais relevantes para o CorpAcad. Embora não seja um número muito expressivo, a partir das nuvens juntamente com as listas de conceitos, podemos notar que existe um padrão de produção intelectual no PPGL-PUCRS naquele dado período ao observarmos atentamente cada conceito extraído. Nossa análise inicial já pôde distinguir três grandes áreas de trabalho ao longo desses 11 anos, a saber, 1) Fonética/Fonologia, 2) Análise do Discurso e 3) Pragmática. Essas foram as três grandes áreas nas quais houve maior intensidade de produção devido à alta presença de conceitos pertinentes as suas terminologias.

Essa terminologia foi comprovada posteriormente por profissionais das respectivas áreas, pois repassamos as listas de conceitos do Quadro 1 a doutores e doutorandos em Linguística para que os termos ali expostos fossem analisados como pertencentes ou não ao domínio de sua expertise. Os especialistas puderam corroborar alguns conceitos como específicos da terminologia de sua área, assim como puderam acrescentar outros, cuja inserção não estava prevista na nossa análise inicial. Com isso chegamos à seguinte triagem de conceitos por área, demonstrada na Tabela 2, na qual temos os termos classificados em 5 categorias:

- Termos exclusivos da Fonética/Fonologia estão marcados com fundo roxo;
- Termos exclusivos da Análise do Discurso estão marcados com fundo amarelo;
- Termos exclusivos da Pragmática estão marcados com fundo verde;
- Termos comuns à Análise do Discurso e Pragmática estão marcados com fundo azul;
- Termos de classificação da Linguística em geral.

Existe, portanto, uma terminologia básica de cada área, que ocorre naturalmente nos textos do *corpus* de estudo, e acaba por caracterizar o seu aparato teórico-metodológico. A constatação desse perfil de especialidade também nos indica outras áreas não ali contempladas, as quais podem incrementar os estudos linguísticos do PPGL, impulsionando novas pesquisas em diferentes campos até então não explorados, ou não bem estabelecidos dentro da tradição linguística, como o já apresentado PFA. A constatação acima, através dos conceitos extraídos, das áreas da Linguística nos incentiva a continuarmos explorando o PFA a fim de estabelecê-lo como uma pedagogia de ensino de LP, assentada em metodologias oriundas de pesquisa científica que nos forneçam informações confiáveis e valiosas sobre o registro acadêmico.

Somente definir as áreas de trabalho do PPGL durante os últimos 11 anos não é algo que nos traga uma real novidade. No entanto, temos, a partir das nuvens e outros recursos gerados pelo ExATO, a possibilidade de realizar uma análise mais profunda de cada área, apontando para questões eminentemente práticas, como a produção de glossários, dicionários terminológicos, ementas entre outras possíveis aplicações. Não só restrito às áreas em si, podemos refletir sobre os conceitos que as perpassam e tipificam a própria ciência linguística e, por extensão, conceitos que são necessários para uma pedagogia de ensino de PLA para alunos estrangeiros em mobilidade acadêmica que necessitem de proficiência em PFA.

Tomemos, por exemplo, os resultados referentes ao ano de 2015. Ao nos debruçarmos mais atentamente sobre os 150 conceitos encontrados, podemos observar, em ordem decrescente de relevância, os seguintes termos:

Tabela 2 – Alguns conceitos extraídos do ano de 2015.

posição	termo	frequência	posição	termo	frequência
3	enunciado	666.5	83	gêneros discursivos	90
4	interlocutor	644	88	substantivo	88
9	encadeamento	364	92	anáfora	84
12	diálogo	279	93	ensino de leitura	83
20	atos	224	94	estratégias de leitura	82
23	verbos	214	97	aprendizagem de leitura	81
33	negação	174	98	compreensão de texto	80
45	compreensão leitora	128	111	bloco semântico	75
57	pronomes	117	122	sintagma	71
66	oração	107	124	sentido de enunciado	70
70	preposição	101	125	encadeamento argumentativo	70
72	léxico	98			

Fonte: elaborada pelos autores.

À frente de cada sintagma nominal está sua colocação na lista de 1 até 150, e sucedendo cada termo está o seu índice de relevância indicado pelo ExATO, na etapa de computação desse índice através de um método de forte base estatística, como dito anteriormente.

A partir desses sintagmas, podemos intuir a seguinte consideração: tendo em vista que o ensino de PLA, assumido nesse artigo sob a égide da Pragmática, parte, portanto, do pressuposto de que a língua deve ser ensinada e adequada à real necessidade de ser utilizada

em determinado contexto, devemos abarcar, então, esses conceitos em atividades desenhadas que visem à proficiência do aluno no contexto acadêmico.

Tomemos, por exemplo, os conceitos encontrados na Tabela 3:

> "verbos / pronomes / oração / preposição / léxico / substantivo / anáfora / sintagma / sentido de enunciado / encadeamento argumentativo".

Ao refletirmos sobre eles e sobre o ensino de PLA, podemos intuir que tais são essenciais para a escrita, por exemplo, de um texto acadêmico, no qual a tipologia textual requer um domínio maior das habilidades de escrita do aluno, a fim de se obter um texto coeso e coerente. Por isso, é valioso que acrescentemos às atividades de ensino de PLA tópicos como esses, cujo objetivo final é o desenvolvimento da proficiência linguística do aprendente de LP - ou mesmo do falante nativo - cuja proficiência no contexto acadêmico também deve ser alcançada para obter aprovação em suas atividades de estudo.

Da mesma forma, ao longo de uma tese ou dissertação, caso seja um aluno de pós-graduação, é fundamental que se encontre no seu texto "encadeamento argumentativo", pois esses gêneros acadêmicos em questão são baseados em discussões científicas, nas quais existem argumentos que corroboram outros, assim como argumentos conflitantes sobre o mesmo assunto.

5. Considerações finais

O Português para Fins Acadêmicos está se solidificando no PPGL-PUCRS através das pesquisas desenvolvidas dentro do Grupo UPLA, o qual está crescendo e refinando sua abordagem pragmática. Esse refinamento se alcança devido à congregação epistemológica da LC e PLN. Essa interface bem estabelecida se estabelece para que o trabalho**,** tanto em PFA quanto em PLA**,** seja mais eficaz no que concerne a uma descrição e análise mais empírica sobre a LP no registro acadêmico. Ainda que a contribuição do artigo seja um pouco sucinta devido ao fato dos resultados de análise serem iniciais, já temos em mãos dados pertinentes que nos propiciam a descrever concretamente o registro acadêmico em seus gêneros tese e dissertação de forma mais eficiente.

Além disso, é importante salientar que trabalhos futuros aprofundarão a análise dos resultados obtidos, abrindo novos horizontes para a exploração do português dentro da academia. Com foco no ensino e aprendizagem de PLA, de maneira mais prática, já podemos

vislumbrar um glossário de termos acadêmicos, um dicionário terminológico de Linguística, uma ementa que enfoque nas habilidades linguísticas necessárias para o aluno que utilize a LP como meio de estudo, quanto para aquele que pesquisa a LP, entre outras possíveis aplicações de extrema importância para a pesquisa, referência e consulta por alunos nativos e não nativos de LP.

Referências Bibliográficas

BAKER, P.; HARDIE, A.; MCENERY, T. **A Glossary of Corpus Linguistics.** Edinburgh: Edinburgh University Press, 2006. 187p.

BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus Linguistics**: Investigating Language Structure and Use. Cambridge: Cambridge University Press, 1998. https://doi.org/10.1017/CBO9780511804489

BICK, E. **The parsing system PALAVRAS**: automatic grammatical analysis of Portuguese in constraint grammar framework. 2000. Ph.D. (Doctoral Thesis). Arhus University, Arhus, 2000.

DALE, R.; MOISL, H.; SOMERS, H. **Handbook of Natural Language Processing** (first edition). New York: Marcel Dekker, 2000.

FINATTO, M. J. B.; LOPES, L.; CIULLA, A. Processamento de Linguagem Natural, Linguística de *Corpus* e Estudos Linguísticos: uma parceria bem-sucedida. In: **Domínios de Lingu@gem.** v. 9, n. 5 (dez. 2015).

KENNEDY, G. **An introduction to Corpus Linguistics**. London & New York: Longman, 1998.

LOPES, L. **Extração Automática de Conceitos a partir de Textos em Língua Portuguesa.** 2012. Tese (Doutorado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2012.

LOPES, L.; VIEIRA, R. Processamento de Linguagem Natural e o Tratamento Computacional de Linguagens Científicas. In: **Linguagens Especializadas em *Corpora***: modos de dizer e interfaces de pesquisa. PERNA, C.; DELGADO, H.; e FINATTO, M. (orgs.). Porto Alegre: EdiPucrs, 2010. p.183-201.

LOPES, L.; FERNANDES, P.; VIEIRA, R. ExATO - High Quality Term Extraction for Portuguese and English. In: **2016 IEEE/WIC/ACM International Conference on Web Intelligence**, 2016, Omaha. Proceedings of International Conference on Web Intelligence. Omaha - Nebraska - USA, 2016. p. 1-6. https://doi.org/10.1109/WI.2016.0092

LOPES, L.; VIEIRA, R. Evaluation of cutoff policies for term extraction. **Journal of the Brazilian Computer Society**, v. 21(1), p. 1-9, Elsevier, 2015.

LOPES, L.; FERNANDES, P.; VIEIRA, R. Estimating term domain relevance through term frequency, disjoint *corpora* frequency - tf-dcf. **Knowledge-Based Systems**, v. 97: p. 237-249, Elsevier, 2016.

LOPES, L.; VIEIRA, R. Improving Portuguese Term Extraction. In: **International Conference on Computational Processing of the Portuguese Language** - PROPOR, 2012, Coimbra. Lecture Notes in Computer Science - Proceedings of PROPOR 2012. Heidelberg: Springer, 2012. v. 7243. p. 85-92. https://doi.org/10.1007/978-3-642-28885-2_9

LOPES, L.; FERNANDES, P.; VIEIRA, R.; FEDRIZZI, G. ExATOlp - An Automatic Tool for Term Extraction from Portuguese Language *Corpora*. In: **LTC'09 - 4th Language and Technology Conference**, 2009, Poznan, 2009, Poznan. Proceedings of the Fourth Language and Technology Conference. Poznan: Adam Mickiewicz University, 2009. p. 427-431.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge: The MIT Press, 1999.

MITKOV, R. (ed.). **The Oxford Handbook of Computational Linguistics**. Oxford: Oxford University Press, 2003.

MOLSING, K. V.; PERNA, C. B. L.-P. Research and Teaching in Portuguese for Specific Purposes. **BELT-Brazilian English Language Teaching Journal** 5.2 (2015): 1-7. https://doi.org/10.15448/2178-3640.2014.2.19701

TEUBERT, W.; CERMÁKOVÁ, A. **Corpus Linguistics. A short introduction**. London: Continuum, 2007.

Artigo recebido em: 15.12.2016 Artigo aprovado em: 21.03.2017