

unesp  **UNIVERSIDADE ESTADUAL PAULISTA**
“JÚLIO DE MESQUITA FILHO”
Faculdade de Ciências e Letras
Campus de Araraquara – SP
Departamento de Letras Modernas

FERNANDA SCHELUCHUAK DIAS

LINGUÍSTICA DE *CORPUS* E LÍNGUA ESPANHOLA:
testando o Unitex com fins lexicográficos



ARARAQUARA – S.P.
2017

FERNANDA SCHELUCHUAK DIAS

**LINGÜÍSTICA DE CORPUS E LÍNGUA ESPANHOLA:
testando o Unitex com fins lexicográficos**

Trabalho de Conclusão de Curso (TCC)
apresentado ao Conselho de Curso de Letras, da
Faculdade de Ciências e Letras –
Unesp/Araraquara, como requisito para obtenção
do título de Bacharel em Letras.

Orientador: Odair Luiz Nadin da Silva

ARARAQUARA – S.P.
2017

SCHELUCHUAK-DIAS, Fernanda.

Linguística de *Corpus* e Língua Espanhola: testando o
Unitex com fins lexicográficos / Fernanda Scheluchuak
Dias. – Araraquara.
48f.; 30 cm

Trabalho de Conclusão de Curso - TCC (Graduação em
Letras) – Faculdade de Ciências e Letras – Universidade
Estadual Paulista “Júlio de Mesquita Filho”, Araraquara,
2017.

Orientador: Odair Luiz Nadin da Silva

1 Linguística de Corpus. 2. Dicionário Pedagógico. 3. Língua
Espanhola.

FERNANDA SCHELUCHUAK DIAS

**LINGÜÍSTICA DE CORPUS E LÍNGUA ESPANHOLA: testando o Unitex
com fins lexicográficos**

Trabalho de Conclusão de Curso (TCC)
apresentado ao Conselho de Curso de Letras, da
Faculdade de Ciências e Letras –
Unesp/Araraquara, como requisito para obtenção
do título de Bacharel em Letras.

Orientador: Odair Luiz Nadin da Silva

Data da defesa/entrega: ____/____/____

MEMBROS COMPONENTES DA BANCA EXAMINADORA:

Presidente e Orientador: Prof. Dr. Odair Luiz Nadin da Silva

Universidade Estadual Paulista "Júlio de Mesquita Filho" – Faculdade de Ciências e Letras de
Araraquara.

Membro Titular: Prof.^a Ms. Mariana Daré Vargas

Universidade Estadual Paulista "Júlio de Mesquita Filho" – Faculdade de Ciências e Letras de
Araraquara.

Membro Titular: Prof. Ms. Renato Rodrigues Pereira

Universidade Estadual Paulista "Júlio de Mesquita Filho" – Faculdade de Ciências e Letras de
Araraquara.

Local: Universidade Estadual Paulista
Faculdade de Ciências e Letras
UNESP – Campus de Araraquara

Dedico esse trabalho aos meus pais Maria de Fátima Scheluchuak e Eduardo Dias, e ao meu irmão Paulo. Sem o apoio deles eu não seria capaz de alcançar meus objetivos. Aos lexicógrafos e estudiosos das palavras que destinam seus estudos à valorização e elaboração de novos dicionários, entre eles os pedagógicos, em especial aos que articulam as duas línguas de origem latina, objeto de estudo desta monografia: o português e o espanhol.

AGRADECIMENTOS

Nesses quase 4 anos de estadia em outra cidade, muitos me apoiaram desde o momento em que eu escolhi cursar uma faculdade longe da minha família e de toda a realidade que eu conhecia. Felizmente, ao longo dessa trajetória, outras pessoas chegaram e fizeram parte desse ciclo inesquecível de muito aprendizado e sabedoria. Seja perto ou a 280 km, vocês tornaram possível a minha graduação e, sobretudo, esse trabalho. Aos que me disseram: *vá! Sem medo, será ótimo, aprenderá muito*. Vocês estavam certos.

Agradeço a minha família que sempre está pronta para me receber e me ajudar. Vocês são essenciais. Em momentos de extremo caos e estresse vocês me transmitiram calma, fazendo com que a distância não fosse notada pela presença diária de vocês. Não há dúvidas de que qualidades como ser responsável, esforçada e honesta foram adquiridas através da nossa convivência. Meu pai Eduardo é a figura de exemplo de homem de caráter. Minha mãe Maria de Fátima é a maior guerreira de todos os tempos, vencendo todas as batalhas que foram impostas a ela, e meu irmão Paulo, sendo um exímio professor de história, que me faz repensar o papel importante que nós, educadores, sempre teremos, além de também me inspirar para ser uma professora tão boa quanto ele.

Aos meus tios, tias, primos e primas, especialmente as minhas madrinhas, Maria Marina Dias Giovanelli e Lusmari Aparecida Nebesmak, que sempre se preocuparam durante esses anos. Minha tia Maria de Lourdes de Paula também tem um lugar querido nesta conquista, pelo carinho de sempre.

Aos meus amigos que continuam comigo, mesmo que não seja na presença física, mas que me acompanham na mente e no coração: Igor Gabia e Leonardo Sisti. De fato, ninguém morre enquanto permanece vivo no coração de alguém.

Meus agradecimentos infinitos e sinceros para o meu orientador Prof. Dr. Odair Luiz Nadin da Silva, que desde 2015 vem me ensinando tudo o que sei, que sempre mostrou acreditar no meu potencial, a fim de extrair o melhor de mim, confiando na minha capacidade, incentivando e impulsionando minha trajetória acadêmica e aprendizagem da língua espanhola. Agradeço pela paciência, leitura meticulosa e contribuições que vão além do academicismo. Eu, como sua *filha acadêmica*, tive muita sorte de poder aprender com você.

Aos amigos de longa data, que me acompanham desde 2010, quando ainda éramos apenas adolescentes com sonhos de fazer uma faculdade, e que felizmente permanecem até hoje no meu círculo de amizades, como Andrezza Ayuso e Nathalia Rodrigues. Vocês me aceitam

como sou e não me poupam honestidade, e estão sempre de braços abertos para me receber. Obrigada por esse apoio e companheirismo incondicional. Faz bem cultivar amizades que não sentem medo de dizer o necessário. Rafaela Oliveira, vocês também conta porque eu sinto e sei que torce por mim, pelo simples motivo de isso ser recíproco.

As pessoas queridas que fizeram da minha estadia em Araraquara menos dolorosa e frustrante, mesmo depois de tantas decepções. Vocês estão marcadas na minha vida, como um sinal de que existem pessoas maravilhosas nesse mundo, principalmente pessoas que destoam da sua realidade da época. Essa é para vocês, Aléxia Zancan e Tainã Stéphanny. Espero que vocês encontrem pessoas como vocês ao longo da vida.

As minhas amizades que nasceram de uma admiração mútua, que viraram reais, Rafaela Chacon e Vitória Bocutti. Vocês são tão fortes quanto quem admiramos. E continuo a acreditar que *you only get one shot, do not miss the chance to blow (...) success is my only motherfuckin# option, failures not*, e assim o fiz. E vocês me ajudaram a fazer. Obrigada.

A todos meus colegas de faculdade, que me ajudaram em muitas outras coisas, e que acolheram essa estrangeira em suas casas, como Gabriel Fernando dos Santos, meu amigo que nunca me diz não, Ana Maria Zacharias, uma das pessoas que eu mais me irrito mas que eu também amo na mesma proporção, Angélica Antonelli, companheira de provas, trabalhos, comidas, e tudo o que for possível, Karla Letícia de Lima, a que me preenche com seus abraços estranhos, Lysllayne Tavela, parceira de bares, cervejas, trabalhos e de opiniões sobre a vida e Raissa Adorno, essa pessoa inteligente e prestativa, que tive o prazer de dividir um Congresso Científico e um Simpósio Internacional. Além do mais, as pessoas que falei pouco durante os 4 anos mas contribuíram para toda a experiência que é morar longe de casa: Ivan Júnior, o meu parceiro espiritual, Luiza Barbieri, que sempre me ajuda quando eu preciso, e Carolina Assenza, pelo acolhimento durante o primeiro ano, tempo da minha adaptação aqui. São incontáveis os momentos de risadas que tivemos, necessárias em meio desse momento delicado que é a graduação.

Aos que chegaram no último ano, como Fabiana Costa: te vejo como uma mulher que não sabe ainda o tamanho e força que tem, mas o tempo te mostrará e te fará mais forte, te espero para assistirmos um jogo do Coringão juntas e Jonas Marques, foi Oxalá que nos uniu, só agradeço pela fé compartilhada que é sempre linda e emocionante.

Meus amigos de São Paulo, Caroline Antonelli, só agradeço pela nossa reaproximação, você é a maior *#teamFernanda* que eu conheço, Carla Biazioli, que é a minha amizade mais antiga e que continua sendo, obrigada por todas as conversas, de verdade, Loraynne Lopes, você é

uma guerreira e toda a sua força também é um pouco parecida com a que eu desejo ter diante da vida, eu te admiro infinitamente, sou sua fã e não se esqueça disso, Gabriela Andrade, minha fisioterapeuta favorita, você é determinada e inteligente, fico feliz com a nossa amizade não ser pesada e sem cobranças, Aline Valéria, que eu sempre nutrirei um carinho bonito e só seu, e acredite, tudo o que eu desejo para nós é paciência e sabedoria para lidar com tantas coisas que as vezes não entendemos, Kayque Felipe, queridão, sempre me lembra do leve que é ser, Luana Miki Sano, que teve de abandonar esse barco louco que é a faculdade, que é o curso de Letras, que ainda tenta me abandonar durante a vida mas que eu jamais permitiria tal distanciamento, mas em compensação, me deu um sobrinho maravilhoso, Mariana Matos, acho que sabe qual é a importância desde 2013, não sabe? Sei que sua preocupação é e sempre será genuína porque seu coração é bonito, e agradeço infinitamente por isso, Caroline de Moraes, o melhor presente que 2017 podia me dar, que amizade a nossa! Samira Martins, que sempre quando me vê parece que é a última, e isso é tão bom porque me sinto a pessoa mais amada do mundo e William Teixeira, sendo esse último responsável pela expansão de conhecimentos, principalmente culturais, reforçando uma conexão forte de pertencimento com o meu país.

Meus queridos amigos que eu compartilho com meu irmão, como Daniele Ayres, Bruno Kellton, Gabriel Camargo. Espero ser recepcionada com meu bolo que está sendo prometido desde 2015.

Agradeço à FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), que me deu a oportunidade de desenvolver o projeto com bolsa de Iniciação Científica que deu origem a presente monografia, e por demonstrar total confiança no meu trabalho durante o ano de 2016-2017.

E, finalmente, agradeço a UNESP. Tudo o que vivenciei só foi possível pela escolha feita, que mostrou ser a melhor para mim.

“Descoberta a palavra, tenta-se compreendê-la. Ao contrário do que parece, as palavras dependem tão intimamente do contexto que, uma vez isoladas, se tornam imprecisas, peçadas de sentidos ancilares, ou vazias de sentido. É o caso de vocábulos tomados a textos poéticos, em que, muita vez, figuram apenas como constante musical, desejadamente misteriosa. A palavra refugia-se dentro de si mesma, é necessário tocaíá-la. Depois, a palavra compreendida – se compreendida – deve ser domesticada. Urge grafá-la, recorrendo à etimologia ou à simples lógica. Como um fidalgo de velha cepa, o dicionarista, mal trava relações com a palavra, sai a investigar-lhe os avós e bisavós, a palmilhar os caminhos do étimo, inçados de cruzamentos e hiperbibasmos, dissimilações e sinalefas, aféreses e paragoges, em que se espetam os mais cautelosos.” Rodrigo Octavio Filho (1961, s/p)

RESUMO

A presente monografia tem por objetivo descrever o processo de etiquetagem dos textos selecionados para a organização do *corpus* GPEL, da inserção deles no *software* Unitex@ e da testagem de seu funcionamento a partir da seleção e análise de amostras de contextos de uso, auxiliando, por meio da etiquetagem, na elaboração de um protótipo de dicionário pedagógico bilíngue para estudantes de Espanhol do Ensino Médio. O *corpus* é composto por textos utilizados em materiais didáticos de espanhol para aprendizes brasileiros do ensino médio, publicados entre 1999 e 2010, e por variados gêneros textuais de domínio jornalístico, veiculados em jornais de sete países hispânicos (Espanha, Argentina, Paraguai, Uruguai, Chile, México e Venezuela), totalizando aproximadamente dois milhões de ocorrências. Com a organização do *corpus* textual supracitado, elaborou-se uma etiqueta cuja função é a de servir à identificação de cada texto e à posterior recuperação das informações bibliográficas no momento da coleta dos contextos para análise. Após a etiquetagem, foi realizada a inserção dos textos no *software* Unitex@, que auxiliou no processo de coleta de contextos a fim de desenvolver os testes necessários, tanto sobre o funcionamento do programa quanto para a recuperação das informações bibliográficas. O uso do *software* citado, bem como as etiquetas utilizadas, demonstraram-se úteis não só para a coleta dos contextos para esse projeto específico, mas também pela forma como os contextos e suas respectivas informações são recuperadas, mostrando-se eficientes para outras pesquisas referentes ao léxico e à língua espanhola. A seleção de contextos seguiu o padrão de extensão, nível de vocabulário e coerência de contexto, questionando a relevância de tal sentença, pois há a possibilidade de mais de uma acepção para um lema. Com a presente monografia, esperamos motivar novos estudos para a área da Linguística de *Corpus* e Lexicografia Pedagógica Bilíngue, bem como priorizar a característica didática de dicionários voltados ao aprendizado do espanhol.

Palavras-chave: Linguística de Corpus, Lexicografia Pedagógica Bilíngue, Dicionário Pedagógico, Língua Espanhola, Aprendizes Brasileiros.

RESUMEN

La presente monografía tiene por objetivo describir el proceso de etiquetado de los textos seleccionados para la organización del *corpus* GPPEL, de la inserción de ellos en el *software* *Unitex@* y de la verificación a partir de la selección y análisis de las muestras de los contextos de uso, auxiliando, a través del etiquetado, en la elaboración de un prototipo de diccionario pedagógico bilingüe para estudiantes de Español de la enseñanza media. El *corpus* es compuesto por textos utilizados en materiales didácticos de español para aprendices brasileños de la enseñanza media, publicados entre 1999 y 2010, y por variados géneros textuales de dominio periodístico, transmitidos en periódicos de siete países hispanicos (España, Argentina, Paraguay, Uruguay, Chile, México y Venezuela), totalizando aproximadamente dos millones de ocurrencias. Con la organización del *corpus* textual citado, se elaboró una etiqueta cuya función es la de servir a la identificación de cada texto y a la posterior recuperación de las informaciones bibliográficas en el momento de la colecta de los contextos para análisis. Después del etiquetado, fue realizado la inserción de los textos en el *software* *Unitex@*, que auxilió en el proceso de la colecta de los contextos a fin de desarrollar los testes necesarios, tanto sobre el funcionamiento del programa cuanto para la recuperación de las informaciones bibliográficas. El uso del *software* citado, también como las etiquetas utilizadas, demostraron ser útiles, no solo para la colecta de los contextos para ese proyecto específico, sino por la forma como los contextos y sus respectivas informaciones son recuperadas, mostrándose eficientes para otras pesquisas referentes al léxico y a la lengua española. La selección de contextos siguió un patrón de extensión, nivel de vocabulario y coherencia de contexto, cuestionando la relevancia de tal oración, pues hay la posibilidad de más de una acepción para un lema. Con la presente monografía, esperamos motivar nuevos estudios para el área de la Lingüística de *Corpus* y Lexicografía Pedagógica Bilingüe, tal como priorizar la característica didáctica de los diccionarios encaminados al aprendizaje de español.

Palabras-claves: Lingüística de Corpus, Lexicografía Pedagógica Bilingüe, Diccionario Pedagógico, Lengua Española, Aprendices Brasileños.

SUMÁRIO

1. INTRODUÇÃO	1
2. LINGUÍSTICA DE CORPUS	4
3. METODOLOGIA	7
3.1 Sobre o nosso corpus e o processo de etiquetagem	7
3.2 Sobre a inserção no Unitex	12
4. ANÁLISE DOS DADOS	18
5. CONSIDERAÇÕES FINAIS	22
6. REFERÊNCIAS BIBLIOGRÁFICAS	24
7. BIBLIOGRAFIA CONSULTADA	27

1. INTRODUÇÃO

A presente monografia dá continuidade às nossas pesquisas desenvolvidas em maio de 2015 até a presente data. Inicialmente, no ano de 2015, elaboramos um projeto nomeado *Etiquetagem e organização de um corpus de textos escritos em língua espanhola*, que foi desenvolvido na pesquisa de Iniciação Científica, na modalidade Sem Bolsa (ICSB), com vínculo ao Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI-CNPq). Depois do fim dessa pesquisa, no ano seguinte, organizamos outra pesquisa relacionada ao assunto, posteriormente aprovada e financiada pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), no período de 01 de março de 2016 a 28 de fevereiro de 2017, intitulada *Lexicografia e Corpus Textual: da etiquetagem dos textos à seleção de contextos de uso com fins lexicográficos* (FAPESP-Processo 2015/25226-7), vinculada ao projeto de pesquisa docente *Lexicografia Pedagógica Bilingue: Elaboração de um Protótipo de Dicionário Português-Espanhol para a Produção de Textos no Ensino Médio* (FAPESP-Processo 2013/26809-0), coordenada pelo Prof. Dr. Odair Luiz Nadin da Silva. Assim, a partir do desenvolvimento das pesquisas mencionadas que tomam a língua espanhola como objeto de suas pesquisas, propusemos o presente tema para a monografia a fim de reafirmar sua importância do ponto de vista cultural, comercial e por número de falantes na atualidade.

Para compreender a relevância da língua espanhola, descrevemos brevemente o panorama contemporâneo que ela se insere. No ano de 2010, o espanhol contava com 450 milhões de falantes, segundo relatório do Instituto Cervantes, tornando-se a segunda língua mais falada no mundo, perdendo apenas para o mandarim. Em 2015, o Instituto Cervantes relatou que cerca de 21 milhões de pessoas estudavam o idioma, e que 6,7% da população mundial a tem como língua materna. O espanhol é a língua oficial de mais de 20 países sendo, portanto, inegável sua importância cultural e política. Entendemos que, frente a tais dados, o valor inferido à língua espanhola tem também implicações políticas, principalmente no Brasil, por fazer fronteira com países hispânicos e pela participação no Mercado Comum do Sul (MERCOSUL).

Apesar do número de falantes, o espanhol não possui a mesma importância da língua inglesa. Com grande influência econômica, as potências anglófonas figuram como as principais empresas dentro do Brasil, movimentando o capital e tornando o idioma requisito básico para se inserir no mercado de trabalho. Além disso, aprender inglês torna-se vantajoso ao simbolizar a inclusão numa cultura globalizada, já que outros países incorporam o ensino

da língua desde a idade mais tenra, fazendo com que a sociedade se habitue a se comunicar com qualquer estrangeiro proficiente nesse idioma. No entanto, mesmo a maior nação política e econômica, Estados Unidos da América, reconhece o prestígio de aprender o espanhol. O intenso fluxo de imigrantes latinos desenvolveu uma demanda considerável, partindo de áreas comerciais e voltadas ao turismo.

Por esta razão, a relevância dessa língua à nível global teve implicações para o sistema de educação brasileiro. Em 2005, o então Presidente Luís Inácio Lula da Silva sancionou a lei 11.161 que determinava a inclusão dessa língua como disciplina de oferta obrigatória no Ensino Médio e facultativo na Educação Básica.¹ Consideramos que o período de 10 anos seja um curto espaço de tempo para a implementação e adaptação do ensino do Espanhol como Língua Estrangeira (E/LE), principalmente nas escolas públicas, as quais sofrem por falta de estímulo, estrutura e outros motivos que não nos cabe dissertar aqui. No entanto, com a obrigatoriedade da oferta do ensino do espanhol, foi preciso pensar em materiais didáticos adequados para essa faixa etária de estudantes do Ensino Médio: em 2010 o Ministério da Educação (MEC), por meio do Programa Nacional do Livro Didático (PNLD), escolhia os materiais didáticos e os renovava a cada ano para a Língua Espanhola.² Além dos livros específicos que auxiliam durante as aulas nas questões gramaticais da língua, oferecendo muitas vezes um ensino dinâmico, por meio de imagens e áudio, incluímos também os dicionários pedagógicos, que podem ser usados nas aulas e como consulta fora do ambiente escolar, com seu caráter independente.

Para a elaboração de um dicionário pedagógico bilíngue, neste caso, português-espanhol, Nadin (2009) diz que é preciso pensar para qual usuário será direcionado, e qual será a função do dicionário para ele. A faixa etária do nosso público alvo engloba jovens de 15 a 18 anos, estudantes do Ensino Médio e do espanhol. Também cabe a definição de dicionário ativo e passivo estabelecida por Nadin (2009): o dicionário ativo será produzido para auxiliar na realização de textos em língua estrangeira, enquanto que o dicionário passivo se encarrega de ajudar na assimilação desses textos. Tratando-se da nossa pesquisa, escolhemos o dicionário ativo. Sua estrutura será pensada e trabalhada de modo a contribuir

¹ Em 2016, essa lei foi revogada pela Medida Provisória nº 746, do atual governo federal.

² Nos anos consultados foram escolhidas as seguintes obras: no ano de 2011 dois materiais didáticos, que são o **Español ¡Entérate!** (por Fatima Cabral Bruno, Margareth M.B. Toni e Sílvia Aparecida Ferrari Arruda) e **Saludos Curso de Lengua Española** (por Ivan Rodrigues Martin); no ano de 2012, três, sendo eles **El arte de leer español** (por Terumi Koto Bonnet Villalba e Deise Cristina de Lima Picanço), **Enlace – Español para jóvenes brasileños** (por Soraia Adel Osman, Neide Elias, Sonia Izquierdo Merinero, Priscila Maria Reis e Jenny Valverde) e **Síntesis – Curso de Lengua Española** (por Ivan Rodrigues Martin).

para a melhoria no desenvolvimento da apreensão léxica do usuário, e consequentemente, na aprimoração da sua competência comunicativa.

Desse modo, ao desenvolver uma pesquisa em Lexicografia Pedagógica a partir de um *corpus* textual, o pesquisador deve preocupar-se com a seleção dos textos que comporão o objeto de estudo, já que a intenção desse *corpus* é o de representar a língua escolhida, ainda que sua completude não seja possível de ser trabalhada. Neste caso, o *corpus* é um recorte de um momento da língua, que jamais abrangerá a totalidade dela. Embora um *corpus* não represente a totalidade da língua, seu uso é um mecanismo eficaz na seleção de contextos reais e adequados para figurarem no dicionário pedagógico bilíngue, tendo em vista auxiliar o público alvo do material didático no processo de aprendizagem da língua espanhola.

Pensando na delimitação do nosso *corpus*, escolhemos e organizamos o *corpus* descrito, que é composto por gêneros textuais disponíveis em materiais didáticos de espanhol específicos para aprendizes brasileiros do ensino médio, publicados entre os anos de 1999 e 2010, e gêneros textuais da esfera jornalística veiculados em jornais de sete países hispânicos (Espanha, Argentina, Paraguai, Uruguai, Chile, México e Venezuela³), totalizando aproximadamente dois milhões de palavras. A escolha desse *corpus* se justifica, por um lado, por serem esses alguns dos materiais didáticos usados em escolas públicas e privadas no Brasil. Por outro, os jornais, intermédio de notícias, e outras seções como opinião, veiculam e divulgam um vocabulário atualizado, ou seja, que representam o estado atual da língua. Também figura como instrumento inclusivo, pois além de jornalistas transmitirem informações, os leitores comunicam-se através de determinadas seções, em um tom menos formal.

Após a manipulação do corpus supracitado, utilizamos de etiquetas para a recuperação de informações bibliográficas. Posteriormente, com as etiquetas padronizadas, os corporas foram inseridos no *software* *Unitex@*. Na sequência, testamos sua funcionalidade e orientamos os demais membros do Grupo GPEL⁴ no uso básico do programa, qual seja, geração de concordâncias para a seleção dos contextos de uso, e posterior análise, para servirem, após aprovação, como exemplos lexicográficos do Dicionário Pedagógico Bilíngue.

³ Os textos dos jornais foram coletados em projetos de Iniciação Científica (PIBIC/CNPq) no período de 2011 a 2014, por Caroline da Costa Lima.

⁴ O GPEL (Grupo de Pesquisa de Estudos do Léxico) é um grupo de estudos criado para debate, pesquisa e entendimento das Ciências do Léxico, como a Linguística de *Corpus*, Lexicologia e Terminologia. O grupo é coordenado pelo Prof. Dr. Odair Luiz Nadin da Silva, na Faculdade de Ciências e Letras da UNESP Araraquara.

2. LINGÜÍSTICA DE CORPUS

Pertencente ao estudo das linguagens, a Linguística de *Corpus* (doravante LC) é uma área que consegue dialogar com outros ramos da Linguística, como a Terminologia, a Lexicografia Pedagógica Bilíngue, a Tradução e as pesquisas no campo do ensino de línguas estrangeiras. Novodvorski e Finatto (2014) afirmam que, logo após o lançamento do livro *Linguística de Corpus* de Berber Sardinha, em 2004, a área de Linguística de Corpus cresceu, e hoje possui pesquisadores em todos os locais do Brasil, ora na Tradução, ora na Terminologia. Berber Sardinha e Almeida (2008) alegam que as pesquisas em Linguística de *Corpus* não se concentram apenas no português, em uma ótica monolíngue, centrando-se apenas nas suas características linguísticas: é possível traçar um diálogo da língua portuguesa com outros idiomas.

A área de Linguística de *Corpus* está ligada intrinsecamente com as inovações computacionais, já que o *corpus* deve estar em arquivo digital, possibilitando a criação de um *corpus* extenso, e, conseqüentemente, mais completo. Dessa forma, começa a surgir uma demanda para a criação de ferramentas que sejam capazes de ler e decodificar o *corpus* textual em toda sua extensão. O advento tecnológico poupa um processo que, se fosse realizado e armazenado manualmente, retardaria os estudos linguísticos. Segundo Berber Sardinha (2004, p. 3),

A Linguística de Corpus ocupa-se da coleta e da exploração de corpora, ou conjuntos de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador.

Sinclair (apud ALUÍSIO & ALMEIDA, 2006, p. 157) atribui uma definição de *corpus* segundo a ótica da Linguística de Corpus. Ele diz que “Um corpus é uma coleção de fragmentos de textos linguísticos em forma eletrônica, selecionados mediante critérios externos para representar, tanto quanto seja possível, uma linguagem ou variação linguística como fonte de dados para pesquisa linguística.”⁵ (Tradução Nossa)⁶. Em contrapartida, o *corpus*, segundo a definição do Centro Virtual Cervantes, é “uma recopilação extensa de textos (escritos, orais ou de ambos os tipos) recolhidos com o objetivo de servir como amostra representativa de uma língua, como conjunto de dados linguísticos reais que refletem o uso da

⁵ “A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.”

⁶ Esta e as demais traduções presentes neste relatório são de nossa autoria e estão identificadas pela sigla TN.

língua (...) do qual querem ser representativos.”⁷ (TN) Ele é o objeto de pesquisa do pesquisador. A análise é realizada tendo o *corpus* como base. Em relação à sua extensão, ainda que o *corpus* seja armazenado eletronicamente, e que isso possa registrar milhões de palavras, o *corpus* é uma fração do léxico da língua estudada. E por não ser possível abranger a totalidade da língua estudada, é necessário pensar sobre quais e quantos serão os tipos textuais que comporão esse *corpus*, pois Piñol (2012, p. 23) afirma que o *corpus* “deve ser representativo da realidade da língua que se vai estudar”⁸ (TN). Não resta dúvida sob o caráter prioritário da escolha do *corpus*.

Sobre a formação do *corpus*, Biber (2012) explica que ao planejá-lo, ele deve ser constituído por variedades de textos na língua e por uma frequência linguística que ressalte a realidade da língua. Santos (2008) diz que a escolha dos textos que formarão o conjunto de um *corpus* precisa ser objetiva: se o *corpus* é coletado sem a reflexão de seu uso, nos depararemos com um conjunto de textos, e não mais um *corpus*. Parodi (2010, p. 23) elenca oito características que um *corpus* adequado deve possuir: “extensão, formato, representatividade, diversificação, marcação ou etiquetagem, procedência, tamanho das amostras e classificação e limitações de tipos disciplinar, temático, etc.” (TN)⁹ Então, após basear criteriosamente um *corpus* diante de tais características, reserva-se o momento para pensar sobre o estudo do léxico através da frequência lexical, que pode ser analisada pela ótica comparativa de duas línguas, como o fazemos nesta monografia, obtendo, assim, os dados necessários para pensar nessas palavras que devem figurar no protótipo, além dos contextos que também são retirados do *corpus*.

Deste modo, o Linguista que empregar a Linguística de *Corpus* como metodologia usará como base as ocorrências linguísticas mais assíduas de um determinado *corpus*. O *corpus* textual coletado e analisado deve estar em formato digital. Esse *corpus* será trabalhado em uma ferramenta que seja capaz de decodificar sua totalidade lexical, através dos *softwares* conhecidos da área de Linguística de *Corpus*, como o *Unitex@*, CEPRIL¹⁰,

⁷ “una recopilación extensa de textos (escritos, orales o de ambos tipos) recogidos con el fin de servir como muestra representativa de una lengua, como conjunto de datos lingüísticos reales que reflejen el uso de la lengua (...) del cual quieren ser representativos.” (CVC)

⁸ “debe ser representativo de la realidad de la lengua que se vaya a estudiar” (PIÑOL, 2012, p. 23)

⁹ “1. Extensión 2. Formato 3. Representatividad 4. Diversificación 5. Marcado o etiquetado 6. Procedencia 7. Tamaño de las muestras 8. Clasificación y adscripciones de tipos disciplinar, temático, etc.” (PARODI, 2010, p. 23)

¹⁰ Centro de Pesquisa, Recursos e Informação em Linguagem.

NILC¹¹, *Wordsmith Tools* etc. Essas ferramentas não são adaptadas para leitura de imagens, negrito e itálico, por exemplo, por isso a necessidade de um arquivo textual (**.txt**), para que o texto, em sua forma pura, seja processado em sua totalidade. Além de operar com *corpus*, as ferramentas também são usadas no desenvolvimento de etiquetadores morfossintáticos e na tradução automática. Para Ribeiro (2004, p. 163), programas como Unitex@ e Wordsmith Tools ajudam na “tradução de fraseologias, colocações e terminologia técnica, (...) que podem ser amplamente beneficiadas pela análise de corpora em língua de origem e língua de chegada, contribuindo, assim, para a produção de um texto mais fluente e natural (...)”

O campo da LC tende a crescer à medida em que houver um avanço nas tecnologias computacionais, que poderão aperfeiçoar e criar novas ferramentas para auxiliar nas análises e manipulação dos *corporas*. Existe a disponibilidade do *download* gratuito dos programas supracitados, com exceção do Wordsmith Tools. Finatto aponta que:

a Lingüística de Corpus assenta-se numa concepção funcionalista de linguagem. Compreende a língua como um sistema probabilístico em que as diferentes combinatórias das unidades lexicais espelharão suas condições gramaticais em um sentido amplo, as quais serão fornecidas pelo de (sic) uso, distribuição, frequência e funcionalidade. (2004, p. 99)

Dessa maneira, partindo da metodologia da Linguística de *Corpus*, é possível a formação de um *corpus*, que é pensado e coletado com um objetivo pré-estabelecido, e, com o uso de *softwares* propícios para a leitura desse *corpus*, há a possibilidade de investigar frequências de uso, combinatórias de unidades léxicas e outras questões acerca da característica lexical, sintática, semântica etc. de determinada língua. Ela não é a única, mas, no entanto, é a metodologia proposta no processo de elaboração de um dicionário por Parodi (2010, p. 14), pois a Linguística de *Corpus* “brinda uma base empírica para o desenvolvimento de materiais educativos e metodológicos de diversa índole assim como para a construção de gramáticas, dicionários e outros, tanto de discursos gerais como especializados, orais e escritos.”¹² (TN)

¹¹ Núcleo Interinstitucional de Lingüística Computacional.

¹² “brinda una base empírica para el desarrollo de materiales educativos y metodológicos de diversa índole así como para la construcción de gramáticas, diccionarios y otros, tanto de discursos generales como especializados, orales y escritos.” (PARODI, 2010, p. 14)

3. METODOLOGIA

O trabalho que realizamos, como aludido, está inserido em um projeto maior que visa a elaboração de um Dicionário Pedagógico Bilingue Português-Espanhol para aprendizes brasileiros do Ensino Médio (DiPEAB). Desse modo, a primeira parte de nossa metodologia é também parte do projeto supramencionado. A natureza de nosso trabalho é quantitativa, na qual fizemos um levantamento de dados dos nossos *corpora* através da inserção deles no *Unitex*, que ilustrou as concordâncias e frequências de verbetes.

Na conjuntura do projeto maior, trabalhamos com dois *corpora* organizados, ao menos inicialmente, para a elaboração do Dicionário em questão e necessitávamos inserir esse conjugado de textos em um programa para que fosse possível gerar as concordâncias e, posteriormente, a seleção e análise dos contextos de uso.

Descrevemos, na sequência, os três procedimentos utilizados em nossa pesquisa, quais sejam: (i) a elaboração na etiquetagem dos textos; (ii) a inserção dos textos etiquetados no *software Unitex@* e testes de seleção e verificação dos contextos de uso e; (iii) a escolha e análise de contextos de uso.

3.1 Sobre o nosso corpus e o processo de etiquetagem

Para o desenvolvimento da pesquisa, usamos o *corpus* textual supramencionado, que conta com aproximadamente 2 milhões de ocorrências, na tentativa de poder contar com mais possibilidades de contextos de uso. Em nossa pesquisa não tivemos a pretensão de que o *corpus* representasse a totalidade da língua espanhola, nem que representasse todos os países hispânicos, pois partimos do princípio de que um *corpus*, como aludido, será sempre representativo de uma parcela de uma realidade, um recorte, seja histórico (diacrônico) ou discursivo (língua comum, usos especializados da língua) etc. Ao longo do período de realização do projeto, que culminou na elaboração dessa monografia, tivemos um objetivo mais modesto, não abrangendo a totalidade de territórios falantes do espanhol; apesar disso, os textos foram coletados em jornais dos sete países (Espanha, Argentina, Paraguai, Uruguai, Chile, México e Venezuela) e em livros didáticos de espanhol para aprendizes brasileiros.

Em suma, o *corpus*, formado por textos retirados de livros didáticos para aprendizes brasileiros do espanhol e de textos jornalísticos, abrangeu as seções contidas no jornal: notícia, carta, entrevista, reportagem, opiniões de leitores enviadas para o jornal, entre outros. A escolha desse *corpus* se justifica pelo fato de que jornais e materiais didáticos

costumam registrar amostras reais da língua com as quais possivelmente os aprendizes interajam em seu cotidiano, seja via educação formal, por meio dos livros didáticos que podem agregar uma diversidade de gêneros textuais, como os literários e descritivos, e também pelas diferentes mídias, como internet, por exemplo, no caso dos jornais, que abrangem uma fração heterogênea de leitores, se pensamos em idade, classe social, gênero, entre outros. O nosso recorte contribuiu de forma relevante por apresentar essa representação da realidade linguística do falante, formando o nosso corpus textual.

Os textos veiculados em livros didáticos haviam sido escaneados e salvos em arquivos *word*, para posteriormente serem convertidos em arquivos *.txt*, e os textos oriundos dos diferentes jornais foram coletados anteriormente em pesquisas de Iniciação Científica de outro membro do Grupo de Pesquisa ao qual estamos vinculados. Coube, portanto, a nós nessa fase, auxiliar na etiquetagem dos textos, na revisão e/ou conferência com os originais quando havia necessidade.

Criou-se, para tanto, um Protocolo Geral de Etiquetagem (PGE – *vide apêndice 1*) para que houvesse um padrão, ou seja, coletar as informações básicas de cada texto. Essa etiquetagem surge então da necessidade de padronização no tratamento dado ao *corpus*, e também de recuperação das referências bibliográficas dos textos.

Como os textos são provenientes de diferentes suportes – livros didáticos e jornais – as etiquetas foram pensadas de forma a contemplar essas diferenças. Assim, para um texto publicado em um livro didático, inserimos a etiqueta:

- **PGE – Livro Didático**

<SIGLA REFERENTE AO NOME DO MATERIAL DIDÁTICO
ACOMPANHADO DO VOLUME SE HOUVER_AUTORES DO MATERIAL
DIDÁTICO_ ANO DE PUBLICAÇÃO DO MATERIAL
DIDÁTICO_PÁGINA DO MATERIAL DIDÁTICO NO QUAL
APARECE_AUTOR DO TEXTO UTILIZADO_DATA DO TEXTO QUANDO
HOUVER_PAÍS DA PUBLICAÇÃO DO TEXTO QUANDO HOUVER_PAÍS
DO AUTOR QUANDO HOUVER>

A título de exemplo, observemos dois pequenos textos publicados em livros didáticos e suas respectivas etiquetas.

Figura 1: Exemplo de texto em livro didático

<LDEALE1_PICANÇO_VILLALBA_2010_p.8>

Identidad

¿Quién eres tú? ¿Cómo eres? ¿Qué haces? ¿Dónde vives? ¿Cuál es tu lengua materna?
¿Te gusta la poesía? ¿Cómo es tu familia?

Latinizad

Hay días en que nos sentimos media perdidos, sin saber exactamente quiénes somos,
¿verdad? A ver qué dice el poema abajo...

</LDEALE1_PICANÇO_VILLALBA_2010_p.8>

Assim, leia-se, da esquerda para a direita, LD-livro didático; EALE1-*El Arte de Leer Español I*; de autoria de Picanço e Villaba, publicado em 2010 à página 8.

O segundo, embora também esteja publicado em livro didático, trata-se de um fragmento de texto literário. Desse modo, sempre que o texto não for de autoria dos autores do livro didático, inseriram-se outras informações sobre o autor, conforme podemos observar na Figura 2.

Figura 2: Exemplo de texto literário em livro didático

<LDEEJB3_OSMANEOUTROS_2010_2009_BENEDETTI_URY>

Poema: Hagamos un trato, del escritor uruguayo Mario Benedetti

Mario Benedetti In BENEDETTI, Mario. *El amor, las mujeres y la vida*. Madrid: Punto de Lectura, 2009.

</LDEEJB3_OSMANEOUTROS_2010_BENEDETTI_2009_URY>

Leia-se, portanto, LD-livro didático; EEJB2-*Español para jóvenes brasileños*; publicado em 2010, de autoria de Osman e outros e que o excerto é de autoria de Mario Benedetti, de 2009; URY-oriundo do Uruguai.

Os dois procedimentos foram adotados em todos os textos coletados em livros didáticos. Sempre que fossem disponibilizadas informações sobre a autoria do texto usado no livro (gêneros literários, jornalísticos etc.), acrescentavam-nas, conforme Figura 2. Quando não havia informações nesse sentido ou informações incompletas, optou-se por incluir somente as informações do livro didático (Figura 1).

- PGE – Jornais

No caso dos textos publicados em diferentes jornais dos sete países hispânicos considerados, criou-se a seguinte etiqueta:

<JORNAL_DIA/MÊS/ANO_PAIS_AUTOR_GÊNERO>

No geral, as etiquetas dos jornais se assemelham às dos livros didáticos, entretanto, por abordar outros gêneros textuais, há a inserção de novas informações (*vide apêndice II*). No jornal, o padrão se concentra em: nome do jornal, data de publicação, país de origem, nome do autor, o gênero ao qual pertence o texto selecionado, conforme Figura 3.

Figura 3: Exemplo de texto em jornal

<CLARÍN_02/06/2011_ARG_OSCAR RODRÍGUEZ_CARTA>

Estacionamientos cada vez más caros

02/06/11 - 01:29

Por OSCAR RODRÍGUEZ

Frecuentemente uso la playa de estacionamiento de avenida Figueroa Alcorta y Pueyrredón, perteneciente a la empresa Dakota, que también explota el servicio de grúas STO.

Recientemente, aumentaron el valor de la estadía de \$ 14 a 24 y la hora de \$ 1,8 a \$ 4,5. Esto implica un aumento del 78% para la estadía y del 170% para la hora. ¿Quién controla? ¿Serán los mismos encargados de controlar el vencimiento del contrato de la grúas?

http://www.clarin.com/opinion/Estacionamientos-vez-caros_0_491950980.html

</CLARÍN_02/06/2011_ARG_OSCAR RODRÍGUEZ_CARTA>

Na figura acima, o texto foi extraído do jornal Clarín, publicado em 02 de Junho de 2011, na Argentina por Oscar Rodríguez e trata-se do gênero “carta¹³”.

Cada texto, portanto, oriundo de livros didáticos ou de jornais, recebeu duas etiquetas: uma no início e outra no fim de cada um deles. Na etiqueta final introduziu-se uma

¹³ É importante ressaltar que a questão do gênero já havia sido definida quando da seleção dos textos nos projetos de PIBIC/CNPq desenvolvidos por Caroline Costa Lima. Os gêneros coletados por ela foram: **carta**, **entrevista**, **nota**, **notícia** e **artigo de opinião**.

barra (/) antes da sigla do livro didático ou do nome do jornal, como podemos observar na Figura 3 acima. O objetivo principal dessa barra foi o de delimitar o fim de um texto. Concluída a etiquetagem, portanto, passamos ao estudo do programa *Unitex@*, o objeto de nossa pesquisa, cuja metodologia descrevemos na seção seguinte.

Conforme descrevemos acima, organizado o *corpus*, realizamos os testes e orientamos, quando necessário, os membros do Grupo quanto ao uso do programa e iniciamos o procedimento de seleção e análise de contextos e preenchimento da ficha lexicográfica, que será explorada na seção de Análises de Dados. Esclarecemos que a proposta de leamário do dicionário foi idealizada pelo Prof. Odair, coordenador do Projeto maior, quem enviava para cada membro do grupo listas de lemas em espanhol para que pudessem realizar a pesquisa nos *corpus*. Desse modo, recebíamos a lista e buscávamos no *corpus* organizado pelo Grupo.

Cumpramos esclarecer que, em nossa pesquisa, priorizamos a seleção dos contextos de uso nos *corpus* de jornais e materiais didáticos supracitados. Entretanto, no conjunto total do dicionário, referente ao projeto maior, usamos também, sempre que necessário, outros *corpus*, como por exemplo, *Corpus de Referencia del Español Actual* (CREA) e o *Corpus del Español del Siglo XXI* (CORPES XXI), ambos da *Real Academia Española*.

Sobre a etiqueta, todos os dados apresentados deveriam seguir o padrão estabelecido no protocolo do projeto maior, entretanto, no decorrer do processo, notamos discrepâncias nas informações. Alguns textos apresentavam todas as informações requeridas, outros apenas o ano de publicação, o que resultou em uma etiqueta sem padrão. O padrão para a etiqueta existia, mas ao produzi-las notamos o problema aludido, pois haviam etiquetas com mais informações do que outras, geradas pela falta de elementos textuais que comporiam a etiqueta. Portanto, fez-se necessária a elaboração de nova formulação e padronização da etiqueta para melhor identificação dos textos, a fim de que todas as informações relevantes pudessem ser encontradas e recuperadas. Alguns exemplos dessas etiquetas antes da reelaboração:

Livro: PALACIOS, M.; CATINO, G. Espanhol para o ensino médio: volume único. São Paulo: Scipione, 2004. – (Série parâmetros)

Etiqueta 1: <LDEEM_2004_*_GEORGINA CATINO>**

Etiqueta 2: <LDEEM_2004_P31_MÓNICA PALACIOS_GEORGINA CATINO>

A necessidade de uma etiqueta que seguisse um padrão tornou-se urgente em razão dos exemplos citados. Na etiqueta 1, temos a sequência: LDEEM (Livro Didático – Espanhol Ensino Médio), o ano de publicação, e então os asteriscos. Os asteriscos sinalizavam a falta de informação dos textos. Nesse caso, o texto coletado não informou de qual página foi retirado. Além desse equívoco na estruturação da etiqueta, apenas uma autora é citada a princípio. Na etiqueta 2, a lacuna de asteriscos é preenchida com a informação requerida (número da página no qual o texto se encontra), sendo a forma adequada na qual pensamos.

Devido a esse problema, fomos levados a repensar e definir uma etiqueta que informasse o mínimo e necessário, para que as informações bibliográficas dos textos fossem possíveis de serem encontradas. Refizemos as etiquetas, com um modelo de etiqueta menor do que a anterior, com as informações relevantes para a nossa pesquisa e coleta.

Livro: ELIAS, N. et al. Enlaces. Español para jóvenes brasileños 1. 2 ed., 3 vol. São Paulo: Macmillan do Brasil Editora, 2010.

Etiqueta padronizada final: <LDEEJB1_OSMANEOUTROS_2010_ARG>

Tiramos informações da etiqueta de livros didáticos como: número da página no qual o texto se encontra, e deixamos como facultativo as informações quando se tratasse de um texto literário. Se caso houvesse a informação de que determinado texto fosse de um autor mencionado, então as informações de autor do texto/país do autor do texto deveriam ser inseridas. As etiquetas de jornais seguiram também um padrão. No exemplo abaixo, tem-se a informação de que o texto foi retirado do jornal *La Nación*, é argentino, foi publicado em 17 de agosto de 2011, escrito por Any Ventura e trata-se de um artigo de *Opinión*.

Etiqueta de Jornal: <JORNAL_DIA/MÊS/ANO_PAIS_AUTOR_GÊNERO>

Exemplo 1: <LANACIÓN_17/08/2011_ARG_ANY VENTURA_OPINIÓN>

3.2 Sobre a inserção no Unitex

Após a revisão e a etiquetagem dos textos, os mesmos foram inseridos no *Unitex@*, que tem por objetivo “processar os textos em línguas naturais utilizando recursos linguísticos. Esses recursos se apresentam na forma de dicionários eletrônicos, de gramáticas

e tabelas de léxico-gramática.” (UNITEX 1.2). Os textos precisaram ser convertidos para a extensão de arquivo *.txt*¹⁴, único formato que o *software* aceita para fazer o processamento de dados. Ainda sobre o *Unitex@*, Almeida e Vale (2008, *apud* BERBER SARDINHA; ALMEIDA, 2008, p. 31) afirmam que:

o software Unitex, desenvolvido na Universidade Marne-La-Vallée (França) por Sébastien Paumier (Paumier, 2002), consiste num conjunto de programas que permite o processamento de grandes quantidades de textos, de diversas línguas. (...) Uma característica que o diferencia de outros programas que trabalham com corpus é o fato de o Unitex funcionar com base de dicionários eletrônicos de cada uma das línguas que o integram. (...) Além disso, o programa também permite que qualquer usuário crie seus próprios dicionários, integrando novas unidades lexicais ou, ainda, acrescentando novas informações morfológicas, sintáticas e semânticas ao léxico já existente ou ainda gerando novas formas a partir de uma forma canônica. Esses dicionários possibilitam ao usuário do programa a realização de buscas pela forma exata, pela forma canônica e também pelas categorias gramaticais.

De posse de todos os textos revisados e convertidos no formato indicado, os inserimos no programa *Unitex@*, organizando assim o *corpus* propriamente dito. Esse *software* permite a seleção de contextos, por meio das denominadas *concordancias*, auxiliando na coleta de informações, bem como no estabelecimento de frequências das palavras no *corpus*, dados que se mostram úteis e indispensáveis no momento de seleção e estabelecimento do leatório, organizado pelo Prof. Dr. Odair Luiz Nadin, coordenador do projeto.

Descrevemos, a partir da sequência de figuras na sequência, mais especificamente da Figura 4 à Figura 10, o funcionamento do programa e o exemplificamos a partir da unidade léxica *pie*.

Figura 4: Funcionamento do Unitex@

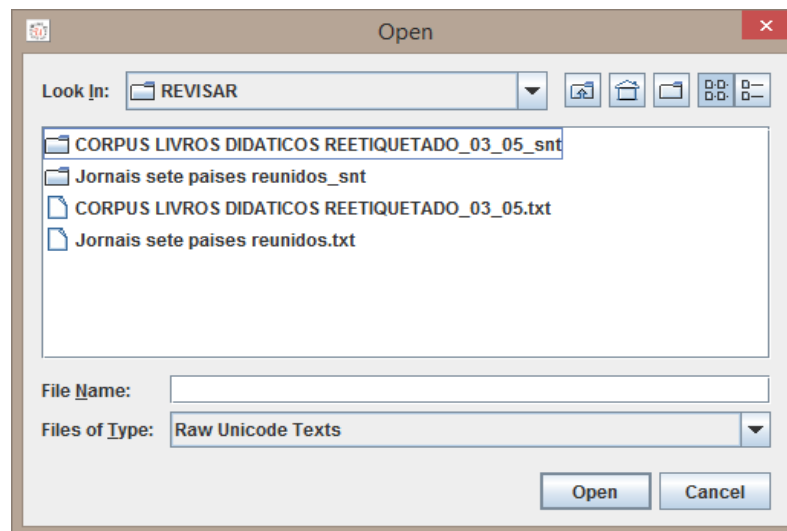


¹⁴ Segundo informação do site IG, o arquivo *.txt* é “arquivo de texto puro”, sendo possível ser processado por qualquer programa de editor de texto. O formato não aceita imagens, apenas caracteres ligados a textos. Disponível em: <<http://somostodosum.ig.com.br/blog-autoconhecimento/informatica-o-que-sao-as-extensoes-de-arquivos-8632.html>>. Acesso: 31 de agosto de 2017.

Quando iniciado o programa, ele solicita o idioma que o usuário deseja operar: *Choose the language you want to work on* (Escolha a linguagem que você deseja trabalhar). Como nosso *corpus* é em língua espanhola, foi selecionada a opção *Spanish* (Espanhol).

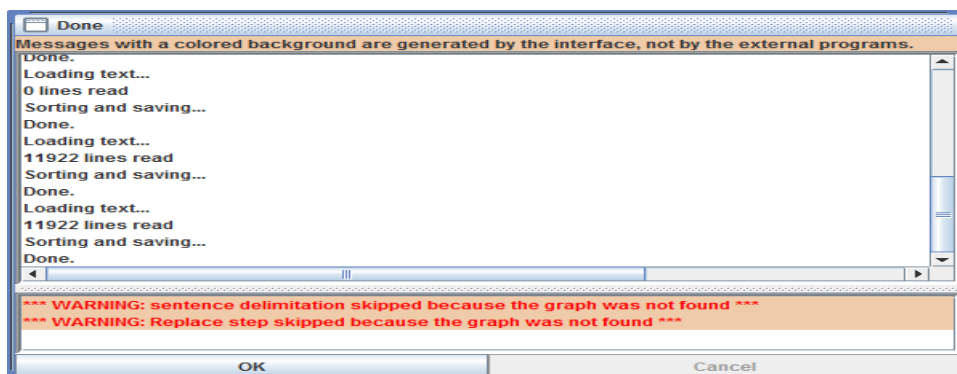
Na sequência, é necessário clicar na barra superior, em *Text* (Texto), e selecionar a opção *Open Tagged Text* (Abrir Texto Etiquetado). Neste caso, é necessário habilitar a opção *Raw Unicode Texts* no bloco *Files of Type* (Tipos de Arquivo). Desse modo, ele mostrará apenas os arquivos convertidos para o tipo *.txt*.

Figura 5: Funcionamento do Unitex@



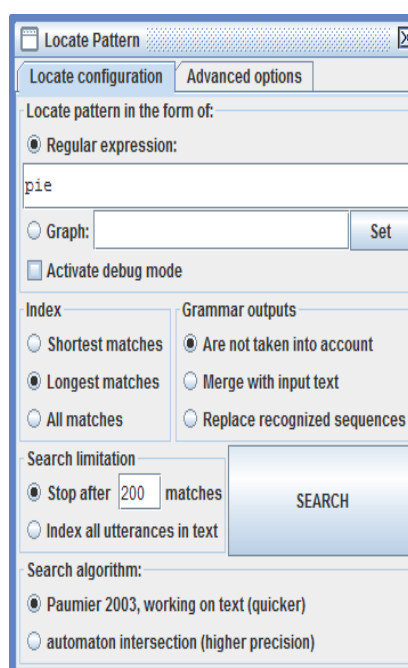
Selecionado o *corpus* desejado, deve-se clicar em *Open* (Abrir). O *corpus* será processado pelo programa, que fará a leitura e acusará caso algum erro aconteça durante o procedimento. Quando a mensagem *Done* (feito) aparece, e a barra *ok* é habilitada, o *corpus* está pronto para ser usado, como demonstramos com a Figura 6.

Figura 6: Funcionamento do Unitex@



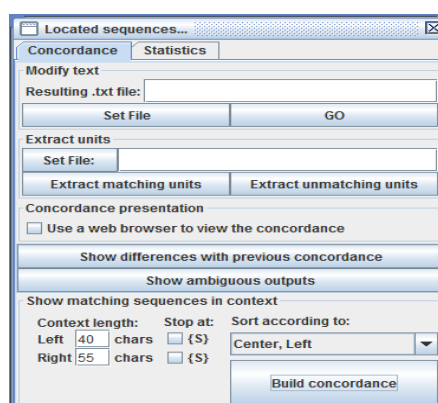
Com o *corpus* já processado pelo programa, é necessária a busca pelas unidades léxicas. Novamente, deve-se clicar na barra superior *Text*, e em seguida na opção *Locate Pattern* (Localizar Padrão). Essa opção é o meio pelo qual procuramos as palavras, e posteriormente, as concordâncias. Abaixo está ilustrado como essa pesquisa é feita. É necessário escrever a palavra dentro da caixa *Regular expression* (Expressão Habitual).

Figura 7: Funcionamento do Unitex@



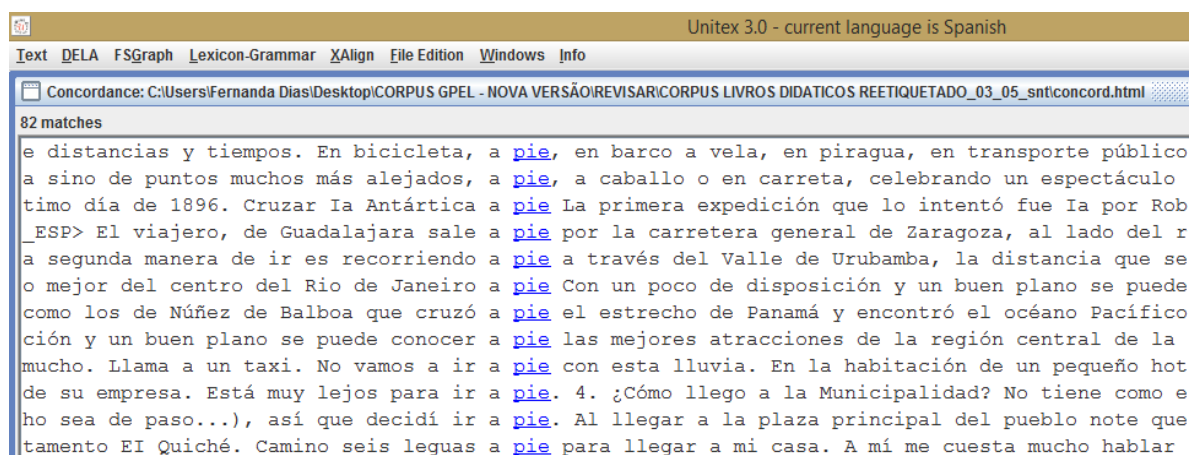
Em seguida, clica-se em *SEARCH* (Procurar). O *corpus* apresenta o total de correspondências, e logo abre outra caixa, devendo clicar em *Build Concordance* (Construir Concordância).

Figura 8: Funcionamento do Unitex@



Ao clicar em *Build Concordance*, o Programa recupera organizado em linhas, denominadas concordâncias, de modo a mostrar os contextos de uso no qual ocorre a palavra pesquisada. No caso elucidado, após clicar em *Build Concordance*, o lema escolhido conta com 82 ocorrências (*matches*) no *corpus*, como a informação mostrada na barra superior.

Figura 9: Funcionamento do Unitex@



Selecionado o contexto, o Programa abre um fragmento do texto a fim de proporcionar melhor análise do contexto e verificar sua pertinência ou não para servir como exemplo lexicográfico. Caso não seja adequado, o conjunto de concordâncias (Figura 9) permanece para novas consultas, sem modificação.

Figura 10: Funcionamento do Unitex@

En bicicleta, a pie, en barco a vela, en
a conocer un territorio.
orridos establecidos y no salir fuera de

Esse procedimento foi realizado para cada um dos 326 lemas que analisamos em nossa pesquisa. As informações coletadas foram inseridas na ficha lexicográfica, conforme ilustração abaixo. Essa ficha contém dados relevantes de contextualização do léxico, incluindo os exemplos lexicográficos em espanhol e as possibilidades de equivalentes no português do Brasil, que deverão compor o verbete do dicionário. Entre esses dados destacam-se a categoria gramatical do lema, a delimitação do significado e informações gramaticas e/ou culturais.

Figura 11: Ficha Lexicográfica

LEMA	CATE GRA	NÚ MER O	SUBLE MA	ACOTA MIENTO	EQUIVAL ENTE	EJEMPLO(S)	REMI SIÓN	INFOCULT	INF OG RA	REFERENCIA DEL EJEMPLO	OBSERVAC ÃO	DUDAS
lavar	v	1			lavar	A partir de ayer se lavan allí con el detergente.				GPEL <CLARÍN_29/08/2010_ARG_SILVIN A HEGUY_NOTÍCIA>		
lavar	v	2			limpar	Godoy Cruz salió decidido a lavarsu imagen.		Godoy Cruz é um time de futebol argentino.		GPEL <LANACIÓN_07/03/2011_ARG_ DANIEL DIMARIA_NOTÍCIA>		
lavar	upv	1	lavado de dinero		lavar dinheiro	Argentina participa de prevención y control de lavado de dinero.				GPEL <PÁGINA12_15/01/2012_ARG_MIG UEL JORQUERA_ENTREVISTA> ADAPTADO		

As frases, que foram retiradas do *corpus*, cumprem os requisitos padrões de exemplo lexicográfico estabelecido para o projeto maior: (i) extensão, (ii) nível de vocabulário e (iii) coerência do contexto. As três frases possuem a extensão adequada: todas têm no máximo 10 palavras¹⁵; o vocabulário encontrado na estrutura das orações é de claro entendimento, e o contexto nos três exemplos diferenciando os equivalentes e a unidade pluriverbal.

Temos acima o lema *lavar*, verbo, com seu equivalente em português **lavar**. Além do verbo, figura também a unidade pluriverbal (upv), correspondente a *lavado de dinero*, que em português significa o ato de *lavar dinheiro*. Sabemos que na primeira frase, o verbo *lavar* está ligado com o substantivo *detergente*, relacionando-se com limpeza e o ato literal de lavar algo; na segunda, escolhemos um sinônimo de *limpar*, usado de uma forma metafórica ainda que o exemplo seja pautado em uma situação real, como indica a coluna *infocult*; e na unidade pluriverbal temos o termo “lavar dinheiro”, expressão que significa ocultar a origem de um ganho financeiro ilícito; ganho que não apresenta origem legal.

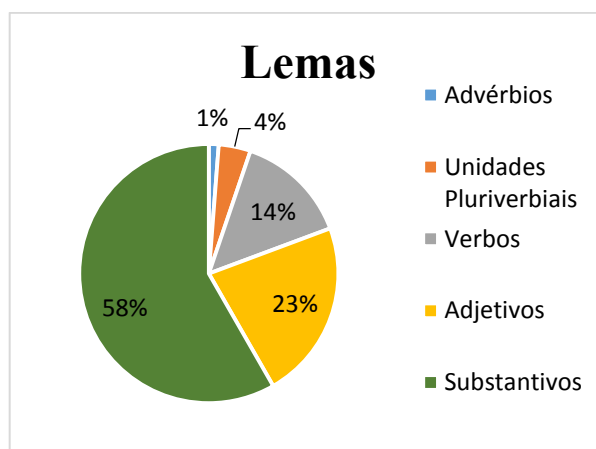
As etapas seguintes consistiram na seleção, descrição e análise de contextos de uso a fim de realizar o preenchimento da ficha lexicográfica. Os contextos selecionados e constantes nas fichas poderão constar como exemplos lexicográficos no dicionário.

¹⁵ Esclarecemos que essa quantidade de dez (10) palavras para cada exemplo foi estabelecida, por questões metodológicas, como padrão no projeto maior. Não se trata, portanto, de uma questão teórica, embora isso possa contribuir para a seleção de exemplos mais coerentes com o modelo de dicionário em elaboração.

4. ANÁLISE DOS DADOS

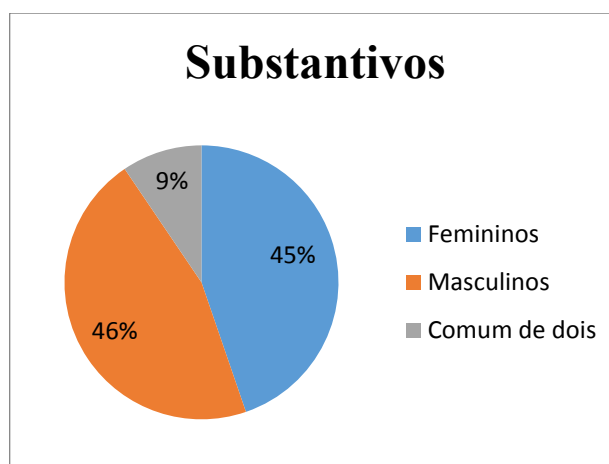
A partir da inserção do *corpus* no *software Unitex@*, foram analisados 326 lemas (vide apêndice III) com os respectivos contextos de uso e demais informações pertinentes presentes na ficha lexicográfica. Entre os lemas analisados, temos: 4 advérbios, 13 unidades pluriverbais, 46 verbos, 73 adjetivos e 190 substantivos. Para melhor elucidar esses dados, utilizamos o Gráfico 1.

Gráfico 1: Categorização de Lemas



Abaixo está ilustrada, em porcentagem, o conjunto de substantivos encontrados nos nossos *corpora*. Dentre eles, 85 femininos, 87 masculinos e 18 substantivos comuns de dois gêneros.

Gráfico 2: Divisão de Substantivos



Todos os lemas figuram na ficha lexicográfica (Figura 12), com as seguintes informações: *lema* (formalização da unidade lexical); *informação gramatical* (coluna destinada para o grupo pertencente de cada lema, sendo *v* para verbo, *adv* para advérbio, *f* e *m* para substantivo, *upv* para unidades pluriverbais e *adj* para adjetivo) etc; *número* (quantas acepções são consideradas possíveis para um único lema, mediante à sua figuração no *corpus*); *sublema* (usado para a colocação das unidades pluriverbais); *delimitação de significado* (usado para a delimitação do significado); *equivalente* (a correspondente na língua portuguesa); *ejemplos* (que são retirados do *corpus* GPEL); *remisión*; *informaciones*; *referencia del ejemplo* (espaço para a etiqueta, possibilitando a identificação do texto); *observación* e *dudas* para qualquer informação relevante.¹⁶ Abaixo, a representação da tabela citada com exemplos:

Figura 12: Exemplos de adjetivo

LEMA	CATEGORIA	NÚMERO	SUBLEMA	ACOTAMIENTO	EQUIVALENTE	EJEMPLO(S)	REMISIÓN	INFOCULT	INFOGRA	REFERENCIA DEL EJEMPLO	OBSERVACIÓN	DUDAS
fino	adj	1			fino	Tenían los mismos labios finos, el mismo cabello crespo.				GPEL <LDEEM_PALACIOS_CATINO_2004_p.167_OCAMPO_ARG>		
fino	adj	2			sutíl, suave, elegante	Puede comer patatas fritas en platos finos decorados.				GPEL <LDEEM_PALACIOS_CATINO_2004_p.292_GENARO_1996_ESP>		

Na Figura 12, temos dois exemplos da mesma palavra. Isso acontece frequentemente no momento de coleta e análise dos contextos. Como vemos no *printscreen*, uma mesma palavra pode ter mais de uma acepção. *Fino*, adjetivo nos dois exemplos, possui significados diferentes dentro dos contextos apontados.

Figura 13: Exemplo de advérbio

LEMA	CATEGORIA	NÚMERO	SUBLEMA	ACOTAMIENTO	EQUIVALENTE	EJEMPLO(S)	REMISIÓN	INFOCULT	INFOGRA	REFERENCIA DEL EJEMPLO	OBSERVACIÓN	DUDAS
adentro	adv	1			dentro	Lo disfruto porque me gusta estar adentro de la cancha.				GPEL <CLARIN_02/01/2012_ARG_MATTIAS BUSTOS MILLA_ENTREVISTA>		
adentro	adv	2			no interior	Adentro están atrapados los 33 mineros.				GPEL <CLARIN_29/08/2010_ARG_SILVINA HEGUY_NOTICIA>		

Na figura 13, temos um advérbio e dois exemplos e dois equivalentes possíveis. Notamos que há exemplos que não necessitam de recorte e/ou adaptação, que já possuem seu sentido total nas condições de extensão exigidas, como os exemplos acima.

¹⁶ Algumas alterações foram feitas durante o processo: as nomenclaturas encontradas nas imagens posteriores (categoría – categorial gramatical – acotamiento, infocult e infogra) são referentes à tabela antiga. Elas foram substituídas por: informação gramatical, delimitação de significado e informações (que contemplam informações gramaticais e culturais), respectivamente.

Figura 14: Exemplos de substantivo

LEMA	CATEGORIA	NÚMERO	SUBLEMA	ACOTAMIENTO	EQUIVALENTE	EJEMPLO(S)	REMISIÓN	INFOCULT	INFOGRA	REFERENCIA DEL EJEMPLO	OBSERVAÇÃO	DUDAS
colega	m/f	1			colega	Me preocupa Pedro. Anda peleándose con sus colegas.				GPEL <GEPP_FANJUL_2005_p.103> ADAPTADO		
letra	f	1			composição	Le hice una nueva letra dedicada a mi hermana Paz.				GPEL <LDTE_MELONE_MENON_2007_p.413>		
pedido	m	1			pedido	Quisiéramos hacer un pedido de galletas de salvado de maíz.				GPEL <LDPS_SOUZA_2003> ADAPTADO		

O lema *colega* é um exemplo de substantivo comum de dois gêneros, ou seja, uma palavra que apresenta uma forma única para os gêneros masculino e feminino. Essa especificidade fica sinalizada com o m/f (masculino/feminino). No caso dos outros dois lemas – letra e pedido – temos respectivamente um exemplo de substantivo feminino e um masculino. Essa informação é relevante para a montagem do verbete, posteriormente, pois as línguas portuguesa e espanhola possuem uma série de substantivos conhecidos como *heterogênicos*, isto é, palavras que diferem o gênero entre uma língua e outra, como por exemplo, “o sorriso/*la sonrisa*”; “a árvore/*el árbol*”.

Figura 15: Exemplo de verbo

LEMA	CATEGORIA	NÚMERO	SUBLEMA	ACOTAMIENTO	EQUIVALENTE	EJEMPLO(S)	REMISIÓN	INFOCULT	INFOGRA	REFERENCIA DEL EJEMPLO	OBSERVAÇÃO	DUDAS
importar	v	1			importar	Pero Julio no conoce las estadísticas. Ni siquiera le importan.				GPEL <PARAGUAY.COM_24/05/2011_PAR_MARIANO_NIN_OPINION>		
importar	v	2		comércio	importar	El navio llevaba matrícula para importar pescado a la UE.				GPEL <ELMUNDO_30/09/10_ESP_INTERNACIONAL_NOTA> ADAPTADO		

A primeira acepção do verbo “importar” trata-se de um verbo intransitivo, cujo significado é, segundo o dicionário *clave*, de “ser conveniente ou ter valor, interesse ou influência”.¹⁷ (TN). Na segunda, trata-se de um verbo transitivo, significa “referido a (...) um produto estrangeiro, introduzi-lo em um país”.¹⁸ (TN). Observemos a coluna *demilitação de significado*¹⁹ com a informação “comércio”, detalhe que o diferencia do primeiro exemplo.

¹⁷ “Ser conveniente o tener valor, interés o influencia” (DCLAVE)

¹⁸ “Referido (...) a un producto extranjero, introducirlo en un país” (DCLAVE)

¹⁹ Entende-se por “acotamiento”, na nomenclatura antiga.

Figura 16: Exemplo de unidade pluriverbal

LEMA	CATEGORIA	NÚMERO	SUBLEMA	ACOTAMIENTOS	EQUIVALENTE	EJEMPLO(S)	REMISION	INFOCUL	INFOGRA	REFERENCIA DEL EJEMPLO	OBSERVAÇÃO	DUDAS
leche	f	1			leite	Voy a pedir un desayuno continental con café con leche.				GPEL <LDHEEA_BRUNO_MEN DONZA_1993>		
leche	upv	1	mala leche		mal humor	Nunca cultives el deporte nacional: la mala leche.				GPEL <ELMUNDO_28/10/09_ES P_FERNANDO SANCHEZ DRAGÓ_ENTREVISTA>		

Para que seja reconhecida como unidade pluriverbal, o lema precisa ser composto de duas ou mais unidades léxicas. No caso acima, a unidade pluriverbal deriva da palavra *leche*, em português, *leite*. No primeiro exemplo, temos um substantivo feminino em espanhol (caso de heterogénico *la leche/o leite*), e abaixo há a sinalização de *upv*, com o número 1, mostrando ser o primeiro exemplo da unidade pluriverbal, como sublema *mala leche* e seu equivalente em português **mal humor**. Esse é um exemplo do que é a unidade pluriverbal e como elas são sinalizadas na ficha lexicográfica.

5. CONSIDERAÇÕES FINAIS

As considerações aqui abordadas são reflexos de uma metodologia que em muito pode auxiliar a área das Ciências do Léxico, especialmente no campo da Lexicografia Pedagógica Bilíngue. O estudo sobre a etiquetagem e o *Unitex@*, aliando-se com a metodologia da Linguística de *Corpus*, mostram como esses dois processos são necessários em uma proposta de dicionário bilíngue.

Com o advento da internet, o dicionário impresso tende a perder relevância, e por vezes é tomado como um material obsoleto, pois não há mais necessidade de saber manuseá-lo. Entretanto é necessário repensar os métodos e sob quais objetivos elas serão produzidas para que essa ferramenta didática continue tendo o seu valor reconhecido. Há de se levar em consideração fatores como o nível linguístico do estudante, a função desse dicionário, monolíngue ou bilíngue, ativo ou passivo, dentre outras questões linguísticas para um material adequado.

Ao pensar na formação de um dicionário bilíngue, por meio da escolha e formação de *corpus*, e sua inserção em uma ferramenta que auxilie na leitura extensa deste *corpus*, é evidente a sua importância. Com o tamanho referido do *corpus* trabalhado, ele só poderia ser trabalhado com o advento tecnológico que surge com a Linguística de *Corpus*. Manualmente não haveria a possibilidade de fazê-lo, pois ainda que a coleta de *corpus* tenha um tamanho finito e seja impossível de representar a língua em sua totalidade, esse processo demandaria um tempo considerável, retardando pesquisas sobre o léxico. Novamente é manifestada a relevância de uma ferramenta que tenha esse caráter facilitador.

Escolhemos um *corpus* textual – o nosso *corpus GPEL* –, que abrangesse a formalidade e a fluidez da língua; o primeiro na forma de livros didáticos, estruturalmente pensados dentro do que a gramática normativa propõe, em razão de serem usados para o ensino e aprendizagem da língua, e o segundo com os textos jornalísticos, com uma linguagem simplificada, do cotidiano, no qual há urgência em comunicar-se, abrangendo diversos gêneros textuais, como a carta, opinião de leitor, reportagem, entre outros.

A diversidade exigida nos *corpus* foi compreendida pensando nos 7 países falantes do espanhol supramencionados, que apresentam diferenças linguísticas. Através da análise de 326 lemas, constatamos a diversidade supramencionada por meio dos dados, que nos revelam advérbios, unidades pluriverbais, verbos, adjetivos e substantivos, femininos, masculinos e comuns de dois gêneros. Ao escolher a ferramenta do *Unitex@*, reforçamos seu caráter gratuito, o que ajuda os pesquisadores diante de suas pesquisas, a possibilidade de

leitura de textos em diversas línguas, como o espanhol, e por cumprir o nosso objetivo, que é o de processamento do *corpus* textual e de coleta de contextos para que figurem no dicionário bilíngue.

Trabalhos vinculados à LC dependem de ferramentas que processem o *corpus* em sua totalidade. Com o avanço da tecnologia, as duas áreas da Linguística trabalhadas no presente trabalho (de *Corpus* e, indiretamente, a Computacional) servirão de respaldo para novos estudos enquanto houver a valorização do ensino de línguas estrangeiras, pensando, evidentemente, em um trabalho comparativo de línguas. Não excluimos aqui a importância das pesquisas na língua portuguesa – as pesquisas continuarão sendo realizadas pois a língua é um fenômeno social, e por isso mutável, porque está em constante mudança, acompanhando as características culturais da sua sociedade. Nos atentamos a destacar as línguas estrangeiras pois, as pesquisas feitas nessa área possuem mais impacto já que, se há oferta de língua estrangeira, há a necessidade de elaboração de materiais adequados para o ensino.

Por meio do PNLD, em 2001, foram distribuídos livros didáticos e dicionários de língua portuguesa para os estudantes de escolas públicas. Com o tempo, através da lei supramencionada, começaram a ser produzidos livros didáticos da língua inglesa e espanhola. Com a lei, uma vez em vigor, nos fez pensar na formulação e produção de materiais didáticos adequados para os estudantes brasileiros de língua espanhola do ensino médio, pois segundo Odair Luiz Nadin, em entrevista para Glauber Moreira (2017), a incidência de dicionários didáticos é baixa, logo, há poucos materiais desenvolvidos para potencializar a aprendizagem desse aluno. Logo, esse trabalho se justifica pela necessidade de se produzir bons materiais didáticos, que possam potencializar o ensino de Língua Espanhola no Brasil.

Em detrimento do atual panorama, esperamos que a nossa pesquisa através da LC possa contribuir para outros protótipos de dicionários, visando uma ferramenta mais completa e adequada para a apreensão de línguas estrangeiras.

6. REFERÊNCIAS BIBLIOGRÁFICAS

- ALUÍSIO, S. M.; ALMEIDA, G. M. de B. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. **Revista Calidoscópico**, Rio Grande do Sul, n. 3, v. 4, p. 156-178, set./dez. 2006. Disponível em: <revistas.unisinos.br/index.php/calidoscopio/article/view/6002/3178>. Acesso: 06 jul. 2016.
- BERBER SARDINHA, T. **Linguística de Corpus**. 1 ed. Barueri: Editora Manole Ltda, 2004.
- BERBER SARDINHA, T. Linguística de Corpus. In: GONÇALVES, A. V.; GÓIS, M. L. de S. (Org.) **Ciências da linguagem: o fazer científico?** Volume 1. 1 ed. Campinas: Mercado de Letras, 2012, p. 321-347.
- BERBER SARDINHA, T.; ALMEIDA, G. M. de B. A Linguística de Corpus no Brasil. In: TAGNIN, S. E. O.; VALE, O. A. (Org.) **Avanços da linguística de corpus no Brasil**. 1 ed. São Paulo: Humanitas, 2008, p. 17-40. Coleção: Pesquisa e Crítica Vol. 3.
- BIBER, D. Representatividade em planejamento de corpus. Tradução de Paula Marcolin. In: FINATTO, M. J. B.; ZILIO, L.; GONÇALVES, F. B. Corpus, Corpora e Dicionários. **Cadernos de Tradução**, Porto Alegre, nº 30, 2012, p. 11-46.
- CENTRO VIRTUAL CERVANTES. **Linguística de corpus**. Disponível em: <https://cvc.cervantes.es/Ensenanza/Biblioteca_Ele/diccio_ele/diccionario/linguisticacorpustm>. Acesso: 14 set. 2017.
- ELIAS, N. et al. **Enlaces. Español para jóvenes brasileños 1**. 2 ed., v. 3. São Paulo: Macmillan do Brasil Editora, 2010.
- FILHO, R. C. **Discurso de recepção - Aurélio Buarque de Holanda**. 1961. Disponível em: <<http://www.academia.org.br/academicos/aurelio-buarque-de-holanda/discurso-de-recepcao>>. Acesso: 11 jul. 2017.
- FINATTO, M. J. B. Terminologia e Linguística de Corpus: da Perspectiva Enunciativa aos Novos Enfoques do Texto Técnico-científico. **Revista Letras de Hoje**, Porto Alegre, n. 4, v. 39, p. 97-106, dez. 2004. Disponível em: <http://www.ufrgs.br/textecc/textquim/arquivos/Terminologia_e_linguistica_de_corpus.pdf>. Acesso: 14 set. 2017.
- IMPORTAR. In: **Diccionario Clave**. Disponível em: <<http://clave.smdiccionarios.com/app.php>>. Acesso: 07 abr. 2017.
- INSTITUTO CERVANTES. **Com 450 mi de falantes, espanhol é o segundo idioma do mundo**. In: Terra Educação. Disponível em: <<https://noticias.terra.com.br/educacao/com-450-mi-de-falantes-espanhol-e-o-segundo-idioma-do-mundo,dee8ec8d7cbea310VgnCLD200000bbcecb0aRCRD.html>>. Acesso em: 04 ago. 2016.
- INSTITUTO CERVANTES. **Relatório aponta que 21 milhões de pessoas estudam espanhol**. In: Exame MUNDO. Disponível em:

<<http://exame.abril.com.br/mundo/noticias/relatorio-aponta-que-21-milhoes-de-pessoas-estudam-espanhol>>. Acesso em: 04 de agosto de 2016.

MALDONADO, C.; NADIN, O. L. **¿Pero es qué todavía es útil usar el diccionario? Un diálogo sobre la importancia de su empleo en la clase de ELE con Concha Maldonado y Odair Nadin.** [jul./set. 2017]. Entrevistador: Glauber Lima Moreira. Florianópolis: Revista Fórum Linguístico, 2017. Disponível em: <<https://periodicos.ufsc.br/index.php/forum/article/view/1984-8412.2017v14n3p2388/35147>>. Acesso: 12 out. 2017.

NADIN, O. L. Dicionários escolares bilíngues de Língua Espanhola: reflexões sobre obras direcionadas ao aprendiz brasileiro. **Revista de Letras**, Curitiba, n. 11, 17 p., 2009. Disponível em: <<https://periodicos.utfpr.edu.br/rl/article/view/2436/1562>>. Acesso: 12 out. 2017.

NOVODVORSKI, A.; FINATTO, M. J. B. Linguística de Corpus no Brasil: uma aventura mais do que adequada. **Revista LETRAS & LETRAS**, n. 2, v. 30, p. 7-16, 2004. Disponível em: <<http://www.seer.ufu.br/index.php/letraseletras/article/viewFile/28516/15799>>. Acesso: 28 set. 2017.

PARODI, G. **Lingüística de Corpus: de la teoría a la empiria.** v. 40. Espanha: Iberoamericana Vervuert, p. 13-36, 2010.

PAUMIER, S. UNITEX 1.2. **Manual do Usuário.** Disponível em: <<http://www.usp.br/gmhp/soft/unitx.pdf>>. Acesso: 06 fev. 2016.

PICANÇO, D. c. de L.; VILLALBA, T. K. B. **El arte de leer español español 1 – contacto.** 1. ed. v. 1. Curitiba: Base Editorial, 2010.

PIÑOL, M. C. **Lingüística de Corpus y Enseñanza del Español como 2/L.** Madrid: Arco/Libros, S.L., 2012, 189 p.

PNLD 2011. **Guia de Livros Didáticos 2011.** Disponível em: <portaldoprofessor.mec.gov.br/storage/materiais/0000016064.pdf>. Acesso em: 12 out. 2017.

PNLD 2012. **Guia de Livros Didáticos 2012.** Disponível em: <<http://www.fn-de.gov.br/centrais-de-conteudos/publicacoes/category/125-guias?download=5510:pnld-2012-lingua-estrangeira>>. Acesso em: 12 out. 2017.

REAL ACADEMÍA ESPAÑOLA. **Corpes XXI.** Disponível em: <<http://www.rae.es/recursos/banco-de-datos/corpes-xxi>>. Acesso: 05 fev. 2016.

REAL ACADEMÍA ESPAÑOLA. **CREA.** Disponível em: <<http://www.rae.es/recursos/banco-de-datos/crea>>. Acesso: 12 mar. 2017.

RIBEIRO, G. C. B. Tradução técnica, terminologia e lingüística de corpus: a ferramenta Wordsmith Tools. **Cadernos de Tradução**, Florianópolis, n. 14, v. 2, p. 159-174, 2004. Disponível em: <<https://periodicos.ufsc.br/index.php/traducao/article/view/6479>>. Acesso: 25 set. 2017.

SANTOS, D. Corporizando algumas questões. In: TAGNIN, S. E. O.; VALE, O. A. **Avanços da linguística de corpus no Brasil**. 1 ed. São Paulo: Humanitas, 2008, 41-68 p. Coleção: Pesquisa e Crítica Vol. 3.

7. BIBLIOGRAFIA CONSULTADA

BELLAY, R. O uso de dicionário por alunos de escola pública no Brasil comparado ao uso de dicionário por estudantes na Espanha. **Revista FACEVV**, Vila Velha, n. 5, jul./dez. 2010, p. 107-117. Disponível em: <<http://facevv.cneec.br/wp-content/uploads/sites/52/2015/10/O-USO-DE-DICION%C3%81RIO-POR-ALUNOS-DE-ESCOLA-P%C3%9ABLICA-NO-BRASIL-COMPARADO-AO-USO-DE.pdf>>. Acesso em: 07 ago. 2017.

COSTA, E. G. de M.; RODRIGUES, F. C.; FREITAS, L. M. A. de. Implantação do Espanhol na Escola Brasileira: Polêmica e Desafios. **Revista Linguasagem**. Disponível em: <<http://www.letras.ufscar.br/linguasagem/edicao10/espanholnaescbr.php>>. Acesso em: 06 jul. 2016.

DIÉGUEZ, I. V. O papel do dicionário no ensino e aprendizagem das línguas. **Revista Exedra, Actas do I EIELP**. Universitat de Barcelona, v. 9, p. 107-110, 2010. Disponível em: <<https://dialnet.unirioja.es/descarga/articulo/3398954.pdf>>. Acesso: 14 mai. 2016.

FROMM, G. O uso de corpora na análise lingüística. **Revista Factus**, São Paulo, n. 1, v. 1, p. 69-76, 2003. Disponível em: <http://comet.fflch.usp.br/sites/comet.fflch.usp.br/files/u30/fromm_corpora.pdf>. Acesso: 06 jul. 2016.

JACOBI, C. C. B. de. **Linguística de Corpus e Ensino de Espanhol a Brasileiros: Descrição de Padrões e Preparação de Atividades Didáticas** (decir/hablar ; mismo; mientras/en cuanto/aunque). 2001. 131 f. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) – Faculdade de Filosofia, Comunicação, Letras e Artes, Pontifícia Universidade Católica de São Paulo, São Paulo, 2001. Disponível em: <http://www.educadores.diaadia.pr.gov.br/arquivos/File/2010/artigos_teses/LinguaEspanhola/Dissertacoes/5dis_claudia.PDF>. Acesso: 06 jul. 2016.

KADER, C. C. C.; RICHTER, M. C. Linguística de corpus: possibilidades e avanços. **Instrumento - Revista de Estudo e Pesquisa em Educação**, Juiz de Fora, n. 1, v. 15, p. 13-23, jan./jun. 2013. Disponível em: <<https://instrumento.ufjf.emnuvens.com.br/revistainstrumento/article/view/2641>>. Acesso em: 13 out. 2017.

OTHERO, G. de A.; AYRES, M. R. Anotação morfológica automática de corpus de língua falada: desafios ao AELIUS. **Revista Texto Livre: Linguagem e Tecnologia**, Belo Horizonte, n. 2, v. 7, p. 44-60, 2014. Disponível em: <www.periodicos.letras.ufmg.br/index.php/textolivre/article/download/6123/5959>. Acesso: 06 jul. 2016.

SAPUCAIA, M. K. de L. **Aquisição e Aprendizado: dois processos no Ensino de uma Segunda Língua**. Centro de Comunicação e Letras, Universidade Presbiteriana Mackenzie, São Paulo. 12 p. Disponível em: <http://www.mackenzie.com.br/fileadmin/Graduacao/CCL/projeto_todasasletras/inicie/MicheleSapucaia.pdf>. Acesso em: 04 ago. 2016.

SCHIERHOLZ, S. J. Lexicografia de Especialidade e Terminografia. Tradução de Leonardo Zilio. In.: FINATTO, M. J. B.; ZILIO, L.; GONÇALVES, F. B. Corpus, Corpora e Dicionários. **Cadernos de Tradução**, Porto Alegre, nº 30, 2012, p. 51-72.

SEIDE, M. S. **Lexicografia pedagógica no Brasil: avanços e desafios**. Unioeste. Disponível em:

<http://150.164.100.248/gtlex/viiiengtlex/pdf/resumos_expandidos/Marcia%20Sipavicius%20Seide.pdf>. Acesso em: 07 ago. 2017.

SILVA, S. T. da. **Conectores Pluriverbais em Espanhol: Proposta de Tratamento Lexicográfico em um Dicionário Pedagógico Semibilíngue**. 2016. 94 f. Dissertação (Mestrado em Linguística e Língua Portuguesa) – Faculdade de Ciências e Letras, Universidade Estadual Paulista “Júlio de Mesquita Filho”, Araraquara, 2016. Disponível em:

<https://repositorio.unesp.br/bitstream/handle/11449/141960/silva_st_me_arafcl.pdf?sequence=3&isAllowed=y>. Acesso em: 08 abr. 2017.

SOARES, D. de A. Introdução à Linguística Aplicada e sua utilidade para as pesquisas em sala de aula de Língua Estrangeira. In: I SIMPÓSIO DE ESTUDOS FILOLÓGICOS E LINGÜÍSTICOS. **Círculo Fluminense de Estudos Filológicos e Lingüísticos**. Universidade Estadual do Rio de Janeiro, Rio de Janeiro, 2008. Disponível em:

<http://www.filologia.org.br/revista/40suple/introdao_a_linguistica%20.pdf>. Acesso: 29 mai. 2016.

8. APÊNDICES

Apêndice I - Protocolo de etiquetagem de textos de livros didáticos

PROTOCOLO DE ETIQUETAGEM

➤ TIPOS DE ETIQUETA

- Modelo Geral:

<SIGLA REFERENTE AO NOME DO MATERIAL DIDÁTICO ACOMPANHADO DO VOLUME SE HOUVER _AUTORES DO MATERIAL DIDÁTICO_ ANO DE PUBLICAÇÃO DO MATERIAL DIDÁTICO__PÁGINA DO MATERIAL DIDÁTICO NO QUAL APARECE _AUTOR DO TEXTO UTILIZADO_ DATA DO TEXTO QUANDO HOUVER_ PAÍS DA PUBLICAÇÃO DO TEXTO QUANDO HOUVER_ PAÍS DO AUTOR QUANDO HOUVER.>

OBSERVAÇÕES:

1. Na etiqueta final põe-se uma barra antes da sigla - </SIGLA...>.
2. No caso de não haver alguma das informações do modelo geral de etiqueta, não colocar nada, nem asterisco, nem traço. Bastam os dados do material didático.
3. No caso de texto retirado de sites, colocar somente os dados do material didático e do autor do texto, quando houver. Não inserir links.
4. A etiqueta é escrita em caixa alta.

➤ REFERÊNCIAS BIBLIOGRÁFICAS E RESPECTIVAS ETIQUETAS

Material Didático	Etiqueta Geral	Quantidade de palavras sem as etiquetas

PALACIOS, M.; CATINO, G. Espanhol para o ensino médio: volume único. São Paulo: Scipione, 2004. – (Série parâmetros).	<LDEEM_PALACIOS_CATINO_2004>	29.724
BRIONES, A. I.; FLAVIAN, E.; FERNÁNDEZ, G. E. <i>Español Ahora 1</i> . São Paulo: Moderna, 2003.	<LDEA1_BRIONES_FLAVIAN_FERNANDEZ_2003>	7.085
BRIONES, A. I.; FLAVIAN, E.; FERNÁNDEZ, G. E. <i>Español Ahora 2</i> . São Paulo: Moderna, 2003.	<LDEA2_BRIONES_FLAVIAN_FERNANDEZ_2003>	15.033
BRIONES, A. I.; FLAVIAN, E.; FERNÁNDEZ, G. E. <i>Español Ahora 3</i> . São Paulo: Moderna, 2003.	<LDEA3_BRIONES_FLAVIAN_FERNANDEZ_2003>	15.747
RÁDIS BATISTA, Livia (org.). <i>Español, Esencial: volume único. Ensino Médio</i> . São Paulo: Moderna, 2008.	<LDEE_RÁDISBATISTA_2008>	40.015
MARTÍN, I. R. <i>Español Serie Brasil</i> . São Paulo: Ática, 2003	<LDESB_MARTÍN_2003>	44.117
BRUNO, F. C.; MENDONZA, M. G. <i>Hacia el Español. Curso de Lengua y Cultura Hispánica. Nivel básico</i> .	<LDHEB_BRUNO_MENDONZA_2004>	19.571

6ed. São Paulo: Saraiva, 2004.		
BRUNO, F. C.; MENDONZA, M. G. Hacia el Español. Curso de Lengua y Cultura Hispánica. Nivel Intermedio. 2ed. São Paulo: Saraiva, 1999	<LDHEI_BRUNO_MENDONZA_1999>	22.566
BRUNO, F. C.; MENDONZA, M. G. Hacia el Español. Curso de Lengua y Cultura Hispánica. Nivel Avanzado. 1ed. São Paulo: Saraiva, 2000	<LDHEEA_BRUNO_MENDONZA_2000 >	32.609
Língua Estrangeira Moderna - Espanhol e Inglês / vários autores. – Curitiba: SEED-PR, 2006. – p. 256	<LSEED_VARIOSAUTORES_2006>	33.259
MILANI, E. M. [et al.] Listo. Español a través de textos. São Paulo: Moderna, 2005.	<LDL_MILANI_2005>	27.707
ALVES, A. N. M; MELLO; A. Mucho: español para brasileños. São Paulo: Moderna, 2000.	<LDM_ALVES_MELLO_2000 >	28.336
SOUZA, Jair de Oliveira. ¡Por Supuesto! Español para brasileños. Ensino Médio. São Paulo: FTD,	<LDPS_SOUZA_2003>	20.426

2003.		
MELONE, Enrique; MENÓN, Lorena. Tiempo español: lengua y cultura. 1.ed. São Paulo: Atual, 2007.	<LDTE_MELONE_MENÓN_2007>	35.597
PICANÇO, D. c. de L.; VILLALBA, T. K. B. El arte de leer español español - Contacto. 1.ed. Volume 1. Curitiba: Base Editorial, 2010.	<LDEALE1_PICANÇO_VILLALBA_20 10>	10.078
PICANÇO, D. c. de L.; VILLALBA, T. K. B. El arte de leer español español - Interacción. 1.ed. Volume 2. Curitiba: Base Editorial, 2010.	<LDEALE2_PICANÇO_VILLALBA_20 10>	11.312
PICANÇO, D. c. de L.; VILLALBA, T. K. B. El arte de leer español 3 - Transformación. 1.ed. Volume 1. Curitiba: Base Editorial, 2010.	<LDEALE3_PICANÇO_VILLALBA_20 10>	14.654
OSMAN, S.; ELIAS, N.; IZQUIERDO, S.; REIS, P.; VALVERDE, J. Enlaces: Español para jóvenes brasileños 1. 1ed. São Paulo: Macmillan, 2010.	<LDEEJB1_OSMANEOUTROS_2010>	9.256
OSMAN, S.; ELIAS, N.; IZQUIERDO, S.; REIS, P.;	<LDEEJB2_OSMANEOUTROS_2010>	11.999

VALVERDE, J. Enlaces: Español para jóvenes brasileños 2. 1ed. São Paulo: Macmillan, 2010.		
OSMAN, S.; ELIAS, N.; IZQUIERDO, S.; REIS, P.; VALVERDE, J. Enlaces: Español para jóvenes brasileños 3. 1ed. São Paulo: Macmillan, 2010.	<LDEEJB3_OSMANEOUTROS_2010>	12.890
MARTIN, I. Síntesis: Curso de lengua española 1. 1 ed., 3 vol. São Paulo: Ática, 2010.	<LDS1_MARTIN_2010>	14.103
MARTIN, I. Síntesis: Curso de lengua española 2. 1 ed., 3 vol. São Paulo: Ática, 2010.	<LDS2_MARTIN_2010>	19.977
MARTIN, I. Síntesis: Curso de lengua española 3. 1 ed., 3 vol. São Paulo: Ática, 2010.	<LDS3_MARTIN_2010>	16.677
MILANI, E. M. Gramática de Espanhol para brasileiros. São Paulo: Saraiva, 1999.	<GEB_MILANI_1999>	3.325
FANJUL, A. (org.). Gramática de Español Paso a Paso. São Paulo: Moderna/Santillana, 2005.	<GEPP_FANJUL_2005>	6.720

Total de ocorrências	502.783
----------------------	---------

Apêndice II - Protocolo de Etiquetagem – Jornais

- Siglas de Países segundo a norma ISO: <http://www.inf.ufrgs.br/~cabral/Paises.html>

ETIQUETA PADRÃO PARA OS JORNAIS

<JORNAL_DIA/MÊS/ANO_PAIS_AUTOR_GÊNERO>

</JORNAL_DIA/MÊS/ANO_PAIS_AUTOR_GÊNERO>

São, no geral, 5 gêneros textuais: *carta*, *entrevista*, *nota*, *notícia* e *opinião*. Contudo, alguns jornais não apresentam todos os gêneros. Nesses casos não há a pasta correspondente no *Dropbox* (site no qual está armazenado todos os *corporas*) e nem o modelo de etiqueta aqui. Assim como feito na etiquetagem dos livros didáticos, a etiqueta deve ser colocada duas vezes no texto, uma no início, sem barra, e outra no final, com barra.

Exemplo de jornais de Uruguai

<ELOBSERVADOR_DIA/MÊS/ANO_URY_AUTOR_ENTREVISTA>

<ELOBSERVADOR_DIA/MÊS/ANO_URY_AUTOR_NOTA>

<ELOBSERVADOR_DIA/MÊS/ANO_URY_AUTOR_NOTÍCIA>

<ELOBSERVADOR_DIA/MÊS/ANO_URY_AUTOR_OPINIÓN>

<LADIARIA_DIA/MÊS/ANO_URY_AUTOR_ENTREVISTA>

<LADIARIA_DIA/MÊS/ANO_URY_AUTOR_NOTA>

<LADIARIA_DIA/MÊS/ANO_URY_AUTOR_NOTÍCIA>

<LADIARIA_DIA/MÊS/ANO_URY_AUTOR_OPINIÓN>

<LARED21_DIA/MÊS/ANO_URY_AUTOR_CARTA>

<LARED21_DIA/MÊS/ANO_URY_AUTOR_ENTREVISTA>

<LARED21_DIA/MÊS/ANO_URY_AUTOR_NOTA>

<LARED21_DIA/MÊS/ANO_URY_AUTOR_NOTÍCIA>

< LARED21_DIA/MÊS/ANO_URY_AUTOR_OPINIÃO>

Apêndice III – Amostras de fichas preenchidas

Tabela 1

LEMA	CATEGORIA	NUMERO	SUBLEMA	ACOTAMENTO	EQUIVALENTE	EMPLEO(S)	REMISION	INFOCULT	INFOGRA	REFERENCIA DEL EMPLEO	OBSERVACION	DUDAS
aceite	m	1			azeite, óleo	Si me gustan, con aceite y vinagre pero poca sal porque pica.				GPEL <LDHEEA_BRUNO_MENDONZA_2000_p.77> ADAPTADO		
aceitunas	f	1			azeitonas	Voy a probar las aceitunas más ricas del mundo.				RAE <Mendoza, Elmer: El amante de Janis Joplin. Barcelona: Tusquets, 2001.>		
acertar	v	1			acertar	A veces tropezarás, a veces acertarás, hay que tener paciencia.				GPEL <ELMUNDO_30/11/10_ESP_AIN HOA ARTETA_ENTREVISTA> ADAPTADO		
alimentación	f	1		empresa	alimentação	El gremio de la alimentación acordó una suba del 35%.				GPEL <CLARÍN_30/05/2010_ARG_NOTICIA>		
alimentación	f	2			alimentação	Con buen desayuno ¡da el primer paso para alimentación saludable!				GPEL <LDEE_RÁDISBATISTA_2008_p.47_ESP> ADAPTADO		
altura	f	1			altura	La altura en la Plaza de Azoguejo es de 28,5 metros.				GPEL <LDHEEA_BRUNO_MENDONZA_2000_p.224_ESP> ADAPTADO		
altura	f	2			nível	A Pablo, a estas alturas, le cargan ya las pantallas.				GPEL <LDHEE_BRUNO_MENDONZA_1999_p.198_ALVAREZ_1996_ESP> ADAPTADO		

Tabela 2

LEMA	CATEGORIA	NUMERO	SUBLEMA	ACOTAMENTO	EQUIVALENTE	EMPLEO(S)	REMISION	INFOCULT	INFOGRA	REFERENCIA DEL EMPLEO	OBSERVACION	DUDAS
acelerador	adj	1			acelerador	Los problemas del TGV se reducen a apretar el acelerador.				GPEL <LDEE_RÁDISBATISTA_2008_p.161_SOPENA_2003_ARG>		
acelerador	adj	2		mecanismo	acelerador	La red es un acelerador de todos los procesos relacionados con la información.				GPEL <PARAGUAY.COM_29/03/12_PARRA *** ENTREVISTA> ADAPTADO		
acelerar	v	1			aumentar	Luego el teléfono acelerará esas comunicaciones.				GPEL <LDEEM_PALACIOS_CATINO_2004_p.30>		
acelerar	v	2			acelerar	La Asamblea Nacional decidió acelerar la aprobación de leyes.				GPEL <ELUNIVERSAL_23/05/2010_VEN_MAYELA ARMAS H_ENTREVISTA>		
acertijo	m	1			adivinha	La naturaleza es juguetona. Le gustan los acertijos y las ironías.				RAE <Gámez, Enrique: Caminitos de plata. 100 cápsulas científicas. México D.F.: McGraw-Hill Interamericana Editores, 2001.>		
ácido	m	1			ácido	La leche materna contiene ácidos grasos esenciales.				RAE <Sociedad Argentina de Pediatría. Comité de Nutrición Argentina. Guía de alimentación para niños sanos de 0 a 2 años. Argentina: Sociedad Argentina de		

Tabela 3

FERNANDA_T4 - Microsoft Word

Arquivo Página Inicial Inserir Layout da Página Referências Correspondências Revisão Exibição Ferramentas de Tabela Design Layout

Recortar Copiar Colar Área de Transferência

Times New Rom 10 A Aa

Fontes

Parágrafo

Estilos

Localizar Substituir Selecionar Edição

LEMA	CATEGORIA	NUMERO	SUBLEMA	ACOTAMIENTOS	EQUIVALENTE	EJEMPLO(S)	REMISION	INFOCULT	INFOGRA	REFERENCIA DEL EJEMPLO	OBSERVAÇÃO	DUDAS
filosófico	adj.	1			filosófico	Esa dimensión de la ciudad cumple con el concepto filosófico antiguo.				GPEL <LDHEEA_BRUNO_MENDONZA_2000_p.187> ADAPTADO		
filósofo	m, f	1			filósofo	Quizás la fórmula sea la que concibió el filósofo John Rawls.		Falecido professor de Harvard, detentor da Teoria da Justiça.		GPEL <CLARÍN_15/08/2010_ARG_RO DOLFO TERRAGNO_OPINIÓN> ADAPTADO		
filtro	m	1			filtro	Carlos Zannini actúa como filtro de la Presidencia.		Carlos Zannini é advogado e político argentino.		GPEL <LANACIÓN_20/03/2011_ARG_MARIANA VERÓN_NOTICIA> ADAPTADO		
fino	adj.	1			fino	Tenían los mismos labios finos, el mismo cabello crespo.				GPEL <LDEEM_PALACIOS_CATINO_2004_p.167_OCAMPO_ARG>		
fino	adj.	2			sutil, suave, elegante	Puede comer patatas fritas en platos finos decorados.				GPEL <LDEEM_PALACIOS_CATINO_2004_p.292_GENARO_1996_ESP>		
firma	f	1			assinatura	Si estás de acuerdo, unid vuestra firma al manifiesto.				GPEL <ELMUNDO_05/10/09_ESP_JAVIER BARDEM_ENTREVISTA>		

Página: 1 de 14 Palabras: 2.480 Español (Espanha - tradicional)

110%

POR 01:24
PTB2 11/04/2017