O Tratamento de Marcadores Discursivos em uma Ferramenta de Apoio à Escrita Acadêmica em Português Para Nativos de Espanhol

Lianet Sepúlveda-Torres, Magali Sanches Duran e Sandra Maria Aluísio

lisepul@gmail.com, magali.duran@uol.com.br,
sandra@icmc.usp.br
Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Matemática e Ciências da Computação, Universidade de São Paulo

Abstract. We report in this paper the development of a module dedicated to discourse markers in HABLA (Hispanofalantes Purchasing an Academic Linguistic Base), a tool designed to support native Spanish speakers in their academic writing in Portuguese. HABLA is conceived to meet the needs of native Spanish speakers who are enrolled in Brazilian federal and state institutions and must write a dissertation or thesis in Portuguese. The diagnosis of difficulties faced by the learners in the use of discourse markers is based on the analysis of a learners' corpus. Part of these difficulties are addressed by two procedures already implemented that identify the problems automatically and present suggestions. The development of the module encompasses the compilation of a bilingual lexicon of discourse markers – Spanish-Portuguese - as well as a list of false friends discourse markers.

Keywords: Portuguese as foreign language. writing support tools. discourse markers. learners' corpus.

1 Introdução

O design de ferramentas de apoio à escrita em língua estrangeira demanda a investigação dos tipos de dificuldades enfrentadas pelos aprendizes nessa atividade. Uma das abordagens para realizar esta tarefa é a análise de um corpus de aprendizes. Os tipos de erros e inadequações detectados servem para inspirar o desenvolvimento das soluções oferecidas pela ferramenta.

Espera-se que um texto escrito corretamente não seja apenas um texto livre de erros lexicais e gramaticais, mas também um texto que apresente as ideias de forma coesa e coerente. Por isso, é importante que o aprendiz utilize corretamente os marcadores discursivos em língua estrangeira. Marcadores discursivos são itens lexicais externos às orações, na sua maioria invariável, que relacionam os enunciados e desempenham um importante papel na coesão dos textos [1] [2]. A simples presença dos marcadores discursivos nos textos não é, porém, garantia de que o texto seja coerente. Aliás, pode ocorrer o contrário: o uso excessivo dos marcadores discursivos pode resultar em um texto com problemas de coerência e coesão [1] [2] [3].

Entre os problemas relacionados ao uso de marcadores discursivos usualmente relatados em trabalhos que analisam córpus de aprendizes [1], [3],[4] estão:

- Emprego incorreto de marcadores discursivos (erros de construção gramatical);
- Sobreuso de marcadores discursivos (os aprendizes aprendem apenas um marcador para cada função e fazem uso recorrente dele);
- Uso de marcadores informais na escrita formal e vice-versa (desconhecimento do registro de cada marcador)
- Uso de marcadores de uma função para outra função (erros de escolha lexical)
- Uso de marcadores inexistentes (gerados por transferência da língua materna para a língua-alvo).

O trabalho que reportamos aqui tem por objetivo informar o módulo de marcadores discursivos do projeto HABLA (Hispanofalantes Adquirindo uma Base Linguística Acadêmica) [5], uma ferramenta de apoio à escrita acadêmica em português para falantes nativos de espanhol. O córpus de aprendizes utilizado neste estudo para diagnosticar dificuldades é o Espanhol-Acadêmico-Br. Esse córpus foi apresentado em [5] e em sua atual versão foi acrescido de 60 novos textos. Ele é constituído de resumos acadêmicos produzidos por diferentes alunos matriculados em programas de pós-graduação da USP, totalizando 403 sentenças e 11098 palavras.

Já o corpus de referência utilizado na implementação dos módulos do projeto HABLA é um córpus paralelo formado por pares de textos da edição online da Revista Pesquisa FAPESP[1]. Esta revista divulga as notícias e reportagens científicas de diferentes áreas do conhecimento, tais como Meio Ambiente, Ciência, Humanidades, Política e Tecnologia. O corpus paralelo foi compilado automaticamente por [6] e contém 2.669 textos paralelos dos pares de línguas português e espanhol. Estes textos foram alinhados de forma sentencial utilizando a ferramenta TCAling [7], obtendo um total de 152,238 pares de sentenças em português e espanhol.

2 Uso dos marcadores discursivos por aprendizes de português, falantes nativos de espanhol

A análise de um corpus de aprendizes constituído de textos acadêmicos escritos em português por falantes nativos do espanhol [5] resultou na identificação de problemas semelhantes aos apontados na literatura e citados na Introdução.

A fim de investigar de forma mais objetiva se o uso de marcadores discursivos por falantes nativos realmente difere do uso feito por aprendizes, escolhemos um grupo de seis marcadores – "no entanto", "porém", "entretanto", "contudo", "todavia", "não obstante" - com a mesma função em português (contraste), levantamos suas ocorrências no corpus de aprendizes e em um corpus de referência (Revista Pesquisa Fapesp) [6]. A Figura 1 mostra a distribuição desses marcadores por posição de uso (no início da sentença e precedido de vírgula), separadamente por nativos e por aprendizes.

[1] http://revistapesquisa.fapesp.br

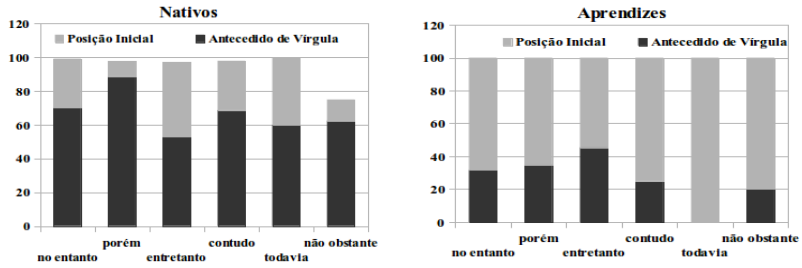

Fig. 1. Distribuição da posição do uso de marcadores por nativos comparada à de aprendizes

Comparando-se a posição de uso de cada marcador, observa-se que os aprendizes utilizam prioritariamente os marcadores em posição inicial, pois é a construção mais simples, ao contrário dos nativos, que usam mais os marcadores entre as duas orações relacionadas. Já na Figura 2, comparamos a frequência de uso dos mesmos marcadores por nativos e aprendizes.

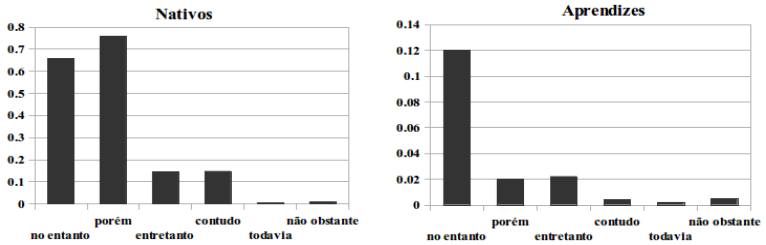

Fig. 2. Frequência de uso de marcadores por nativos e aprendizes

Observa-se que os aprendizes utilizam mais os marcadores "todavia" e "não obstante" do que os nativos, provavelmente porque ambos possuem cognatos no espanhol. Por outro lado, aprendizes usam muito menos "porém" que os nativos. Embora não constituam erros, essas diferenças de uso "marcam" o discurso dos aprendizes, distanciando-o do discurso produzido por nativos.

A proximidade entre o espanhol e o português torna o problema de interferência da língua materna ainda maior, pois a existência de muitos cognatos entre as duas línguas faz com que os aprendizes "arrisquem" traduções literais de marcadores discursivos do espanhol para o português. Embora a estratégia funcione na maioria das vezes, há casos em que ela falha, gerando erros de falsos cognatos. É extremamente comum, por exemplo, encontrarmos "Em concreto...", "De uma parte..., de outra parte, ..." quando os marcadores equivalentes em português são, respectivamente, "Na verdade", "De/Por um lado..., de/por outro lado, ...". Nem sempre tais erros de falsos cognatos impedem a compreensão do texto, mas eles podem causar estranheza aos nativos de língua portuguesa, prejudicando o foco da leitura.

De acordo com [8], falsos cognatos são pares de palavras em duas línguas que são similares segundo sua ortografia ou fonética, mas que têm diferentes significados dado o contexto em que foram empregados. Seguindo essa definição, no projeto HABLA consideramos marcadores discursivos falsos cognatos aqueles cuja tradução literal do espanhol para o português resulta em um item lexical que pode ou não exis-

tir em português, mas nunca exercendo a função de marcador discursivo. Atualmente, no projeto HABLA foram inventariados treze marcadores discursivos como falsos cognatos entre o espanhol e o português, como, por exemplo, "sea como sea" (seja como for) e "a causa de" (por causa de). Estes itens foram levantados com a finalidade de identificar alguns erros produzidos pela interferência do espanhol na escrita do português. Sempre que esse tipo de erro for identificado pela ferramenta do projeto HABLA, será oferecida a sugestão apresentada na Figura 3, ou seja, para um marcador em espanhol é apresentado suas possíveis equivalências em português.

O uso somente de léxicos de marcadores discursivos falsos cognatos para a identificação de erros, não são suficientes porque algumas sequências são ambíguas: podem ser mesmo um falso cognato de marcador (verdadeiro positivo) ou uma sequência perfeitamente correta para outros textos em português (falso positivo), como pode ser visto na Tabela 1.

Tabela 1. Identificação de falsos cognatos de marcadores discursivos

Verdadeiro Positivo	Falso Positivo
Em concreto, aplicamos categorias de análises coerentes com tal perspectiva ao estudo dos processos de construção de conteúdos relativos à contaminação e ao uso da água com uma sala de aula com alunos entre os 15 e os 16 anos.	Até hoje, ninguém desenvolveu o produto para aplicação **em concreto** celular ou para uso como aditivo em argamassa.

O desafio, entretanto, não é somente apontar um possível falso cognato, mas sim identificar o falso cognato na função de marcador discursivo. Somada a esta, há outra dificuldade para a automatização do auxílio aos aprendizes: muitos marcadores apresentam ambiguidade com outros usos do léxico que os compõem [9,10], como os exemplos apresentados na Tabela 2.

Tabela 2. Identificação de marcadores discursivos

Verdadeiro Positivo	Falso Positivo
Eles vieram**, embora** não tenham sido convidados.	Eles vieram **embora** assim que os chamei.
Logo, optar pelo silêncio, esperar pela melhor oportunidade de abordar certo assunto e, assim, atingir seu objetivo sem traumas, é a melhor opção.	Tão **logo** começam a andar, os roedores são capazes de reconhecer os sinais deixados no ambiente pelo predador e perceber quando é hora de sumir.

A Tabela 2 mostra como o item lexical "embora" pode ser ora marcador discursivo, ora advérbio de direção que compõe alguns predicados complexos ("ir embora", "vir embora", "mandar embora"), e o item lexical "logo" pode ser ora um marcador discursivo, ora um advérbio de tempo.

Para identificar os itens lexicais na sua função de marcador discursivo, nós os combinamos com padrões que observamos em um corpus de língua culta do português do Brasil (Revista Pesquisa Fapesp). Esses padrões nos mostram que grande parte dos marcadores discursivos pode ocorrer: a) no início de uma das orações que relaciona, seguido de vírgula; b) entre as duas orações que relaciona, precedido de vírgula ou c) no meio de uma das orações que relaciona, entre vírgulas, como nos exemplos a seguir, criados para fins ilustrativos.

(a) Ele estudou. **No entanto**, não foi aprovado no exame.
(b) Ele estudou, **no entanto** não foi aprovado no exame.

(c) Ele estudou. Não foi, **no entanto**, aprovado no exame.

Uma exceção são os marcadores que se juntam à oração que exprime o primeiro fato no tempo:

(a) **Embora** ele tenha estudado, não foi aprovado no exame.
(b) Ele não foi aprovado no exame**, embora** tenha estudado.

Para simplificarmos, decidimos utilizar na identificação apenas as características comuns aos dois tipos de marcadores, i.e., para ser considerado marcador, o item lexical tem que estar em início de sentença ou ser *precedido* de vírgula. Essas duas características serão utilizadas como parâmetro para duas ações que serão implementadas no módulo de marcadores discursivos do HABLA:

- Apontar marcadores falsos cognatos, apresentando sugestões de equivalentes que possam ser utilizados na correção, como apresentado na Figura 3;
- Apontar marcadores muito recorrentes (que apresentem mais de duas ocorrências em um texto de 500 palavras), apresentando os sinônimos com a mesma função que podem ser utilizados para melhorar a diversidade lexical do texto, como apresentado na Figura 4.

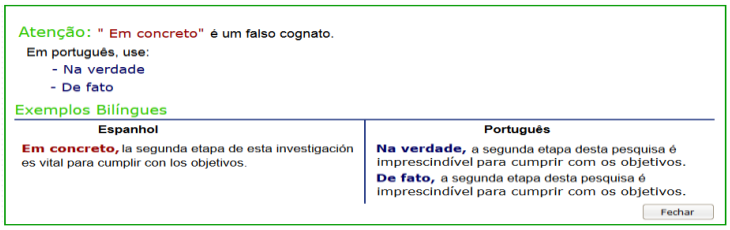

Fig. 3. Tela de sugestão de falso cognato de marcador discursivo no sistema HABLA

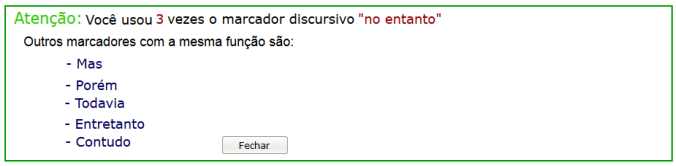

Fig. 4. Tela de sugestão de marcador discursivo muito utilizado

Em textos de aprendizes, contudo, essas características poderão eventualmente não ser suficientes caso eles não façam uso adequado da vírgula. Aliás, como observamos que em textos bem escritos existem regras no uso da vírgula junto a marcadores discursivos, pretendemos em trabalho futuro dar suporte ao aprendiz também nesse aspecto.

Outra dificuldade que ficou evidente na análise do corpus de aprendizes, mas não relatadas na literatura é o uso incorreto do modo verbal exigido por determinados marcadores discursivos. Por exemplo, os marcadores "ainda que" e "embora" exigem que o verbo da oração que os contém esteja no modo subjuntivo (1), enquanto o marcador "apesar de" exige o verbo no infinitivo pessoal (2), como ilustram os exemplos criados a seguir.

1. Ainda que **tenhamos tentado**, não conseguimos marcar nenhum gol.

2. Apesar de **termos tentado**, não conseguimos marcar nenhum gol.

A identificação desse tipo de erro, contudo, exige informações que vão além da superfície dos tokens (i.e. etiquetas de anotação morfossintática) e por isso serão implementadas em um segundo momento.

3 Léxico de marcadores discursivos para uso computacional

Apesar da importância dos marcadores na produção escrita, principalmente na escrita acadêmica, estes itens lexicais não estão totalmente inventariados em português e espanhol. A subcategorização dos marcadores lexicais de acordo com sua função, embora reconhecida, varia muito na literatura das duas línguas [2] [11].

Devido ao fato de existirem vários marcadores discursivos para cada função, é muito importante contarmos com uma subcategorização precisa desses itens lexicais por função em ambas as línguas (língua materna do aprendiz e língua-alvo). Isso porque a tradução dos marcadores discursivos não costuma respeitar uma relação de equivalência item a item, mas sim a relação de equivalência entre as subcategorias de função.

Um estudo foi desenvolvido por [2] com a finalidade de classificar, analisar e comparar os marcadores discursivos em espanhol e em português em 200 textos do gênero jornalístico a partir de uma perspectiva linguístico-pragmática. A autora fez um levantamento de todos os marcadores presentes no corpus e considera os cinco grupos apresentados por [12]: (i) Estruturadores da informação; (ii) Conectores; (iii) Operadores Argumentativos; (iv) Reformuladores e (v) Conversacionais. O léxico resultante desse estudo, porém, não possui as características necessárias para uso computacional. Para automatização das sugestões é necessário que esse léxico esteja inventariado extensivamente em cada língua, agrupado por funções e que cada grupo de uma língua tenha um grupo equivalente na outra língua, como é mostrado na Figura 5.

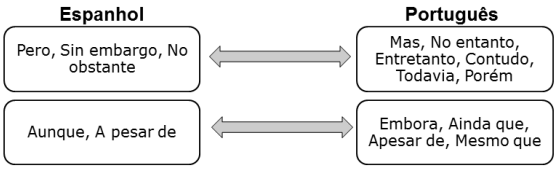

Fig. 5. Léxico bilíngue de marcadores com uma mesma função

A fim de possibilitar a implantação do tratamento de marcadores discursivos no projeto HABLA, construímos um léxico bilíngue de marcadores discursivos, mas ele não contém informações sobre posição, uso de vírgula, tempo verbal exigido e outras possíveis características associadas a cada um dos itens. Esse léxico, somado aos treze marcadores falsos cognatos que inventariamos, foi suficiente para criarmos uma estrutura inicial de suporte automático ao uso de marcadores discursivos na escrita em português como língua estrangeira. É importante ressaltar, contudo, que um estudo contrastivo de marcadores discursivos em português e espanhol com vistas ao tratamento computacional seria desejável para enriquecer o projeto HABLA.

4 Considerações Finais

Reportamos o módulo de apoio ao uso de marcadores discursivos na ferramenta HABLA. Na fase de diagnóstico de dificuldades de aprendizes identificamos oportunidade de incorporar vários possíveis auxílios que dependem da existência de insumos léxicos, gramaticais e pedagógicos que não se encontram disponíveis no momento, como o uso da vírgula associado a cada marcador, as possíveis posições do marcador na sentença e o tempo verbal exigido por cada marcador.

Por ora, estamos utilizando como insumo uma lista bilíngue de marcadores discursivos e uma lista de treze marcadores falsos cognatos, recursos desenvolvidos dentro do próprio projeto. Portanto, o que estamos fazendo é implementar computacionalmente algumas formas de auxílio com recursos léxicos e gramaticais limitados. A utilidade do módulo de marcadores discursivos será aumentada à medida que novos recursos, cujo desenvolvimento extrapola o nosso projeto, estiverem disponíveis para informá-lo.

Atualmente, já foram implementadas as formas de auxílio ilustradas nas Figuras 3 e 4. Foi também concluído o processo que aciona o auxílio da Figura 4, ou seja, o cômputo do sobreuso de um marcador em determinado texto. Falta apenas concluir a identificação automática de falsos cognatos, que aciona o auxílio apresentado na Figura 3. Esta é a parte mais complexa do módulo devido à ambiguidade dos marcadores falsos cognatos com outras categorias gramaticais.

Além disso, pretendemos disponibilizar para consulta os recursos léxicos criados para informar este módulo a fim de beneficiar os aprendizes que estiverem interessados em conhecer as equivalências entre as duas línguas.

O módulo de marcadores discursivos deverá ser testado em breve e ficará disponível no início de 2015, quando a versão inicial do HABLA será lançada.

Referências Bibliográficas

1. Aidinlou, N. A.; Mehr, H. S. The Effect of Discourse Markers Instruction on EFL Learners' Writing. World Journal of Education. (2012)
2. Fernández, S. I. Los Marcadores Discursivos en la Argumentación Escrita: Estudio Comparado en el Español de España y en el Portugués de Brasil. Salamanca: Ediciones Universidad de Salamanca. (2005)
3. Casteele, A.V.; Collewaert, K. The Use of Discourse Markers in Spanish Language Learners' Written Compositions. Procedia - Social and Behavioral Sciences, vol. 95, pp. 550–556. (2013)
4. Jalilifar, A. R. Discourse Markers in Composition Writings: The Case of Iranian Learners of English as a Foreign Language. Journal of CCSE, English Language Teaching, vol.1, no.2 (2008)
5. Sepúlveda-Torres, L.; Rodrigues, R.; Aluísio, S. Espanhol-Acadêmico-Br: A Corpus of Academic Portuguese Learners Produced by Native Speakers of Spanish, In: Aluísio, S. and Tagnin. S. O. (eds.) New Languages Technologies and Linguistic Research: a Two-Way Road. Cambridge Scholars Publishing. (2014)
6. Aziz, W.; Specia, L. Fully automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation. In: The 8th Brazilian Symposium in Information and Human Language Technology, STIL 2011, Cuiabá, MT. (2011)
7. Caseli, H. M. Indução de léxicos bilíngües e regras para a tradução automática. Teses Doutorado, Universidade de São Paulo. (2007)

8. Frunza, O.; Inkpen's D. Identification and Disambiguation of Cognates, False Friends, and Partial Cognates Using Machine Learning Techniques, International Journal of Linguistics, vol. 1, no. 1, p. 1-37, Otawa, Canada. (2009)
9. Heeman, P. A.; Byron, D.; Allen, J. F. Identifying Discourse Markers in Spoken Dialog. (1998)
10. Schourup, L. Discourse Markers. In: Lingua, n. 107, pp. 227--265. (1998)
11. Shanru, Y.: Discourse Markers? An Area of Confusion. Newcastle University, UK. (2012)
12. Martín, Z. M. A.; Portolés, L. J. Los Marcadores del Discurso. In I. Bosque and V. Demonte (eds.), Gramática Descriptiva de la Lengua Española. Tercera parte. Entre la oración y el discurso. Morfología. pp. 4051--4213). Madrid: Espasa Calpe. (1999)