Universidade do Minho
Instituto de Letras e Ciências Humanas

Joana Isabel da Silva Veloso

Identificação/extração semiautomática de colocações utilizando métodos contrastivos

Outubro de 2013

Joana Isabel da Silva Veloso **Identificação/extração semiautomática de colocaçõesutilizando métodos contrastivos**
Uminho | 2013

Universidade do Minho
Instituto de Letras e Ciências Humanas

Joana Isabel da Silva Veloso

Identificação/extração semiautomática de colocações utilizando métodos contrastivos

Tese de Mestrado
Mestrado em Linguística Portuguesa e Comparada

Trabalho efetuado sob a orientação de
Prof. Doutor Álvaro Iriarte Sanromán
e do
Prof. Doutor Alberto Manuel Brandão Simões

Outubro de 2013

DECLARAÇÃO

Nome: Joana Isabel da Silva Veloso
Endereço eletrónico: juana.23@hotmail.com
Número do Bilhete de Identidade: 13716434

Título da tese:

IDENTIFICAÇÃO/EXTRAÇÃO SEMIAUTOMÁTICA DE COLOCAÇÕES UTILIZANDO MÉTODOS CONTRASTIVOS

Orientadores:

Prof. Doutor Álvaro Iriarte Sanromán
Prof. Doutor Alberto Manuel Brandão Simões
Ano de conclusão: 2013

Tese de Mestrado em Linguística Portuguesa e Comparada

Universidade do Minho, 30/10/2013
Assinatura: _____

Aos meus pais

Aos meus avós

Ao Jaime

Agradecimentos

Ao Professor Doutor Álvaro Ririarte Sanromán pela excelente orientação e pelos sábios conselhos dados ao longo de toda a dissertação.

Ao Professor Doutor Alberto Simões pela paciência que teve comigo quando não percebia os aspetos mais técnicos ou mesmo quando bloqueava e precisava dum pequeno empurrão.

A todos os professores do Mestrado de Linguística Portuguesa e Comparada pelos conhecimentos transmitidos nestes últimos dois anos.

Às minhas excelentes colegas de mestrado que tanto apoio me deram durante estes dois últimos anos. Gostaria de agradecer em especial à Micaela Aguiar, que me acompanha desde o meu primeiro ano nesta universidade. Muito obrigada por tudo que partilhámos durante estes cinco anos.

Aos meus amigos de sempre, David, Diana, Janecas, Mário, Hélder, Diana, Joana, Cátia e a Tita por todas as palavras de apoio e por todos os momentos de amizade.

Aos meus pais por estarem sempre lá quando eu preciso. Aos meus irmãos e à minha cunhada por todo o carinho. Aos meus avós por serem o meu porto de abrigo. À minha afilhada por ser a princesa da madrinha.

Ao meu sogro e ao meu cunhado por serem tão atenciosos comigo.

Por último, mas não menos especial, ao Jaime por todo o companheirismo, dedicação, carinho, paciência e acima de tudo amor.

Braga, Outubro de 2013

Identificação/extração semiautomática de colocações utilizando métodos contrastivos

RESUMO

O presente trabalho, que se inscreve nas áreas de investigação da Lexicografia e da Linguística Computacional, pretende investigar da exequibilidade a criação de um algoritmo automático para a extração automática coerente e consistente de colocações a partir da comparação de duas línguas. Para esta experiência escolheram-se as línguas espanhola e a portuguesa e, partindo de um corpus do Banco Central Europeu, obtiveram-se 46,089 pares candidatos a colocação. Estes pares candidatos foram verificados e avaliados de forma manual, utilizando métodos contrastivos. Para além da avaliação fez-se uma análise cuidada dos vários tipos de erros existentes no algoritmo para que mais tarde este possa ser melhorado. Concluindo a análise, comprovou-se que 56,14% dos pares correspondem a combinações lexicais correctamente identificadas, das quais 13,99% eram combinações em que opera algum tipo de restrição lexical e 42,15% eram combinações livres.

Semiautomatic identification/extraction of collocations using contrastive methods

ABSTRACT

This work, which falls in the research areas of Lexicography and Computational Linguistics, pretends to study the practicability of an algorithm for the automatic, coherent and consistent extraction of collocations by comparing translations from two different languages. For this study the chosen languages were Spanish and Portuguese and, starting with the European Central Bank corpus, a total of 46,089 collocation candidates pairs were obtained. This data was verified and evaluated manually, using contrastive methods. Beyond the evaluation, a careful study of the different kind of errors found in the algorithm was made, so that later the algorithm can be improved. The results shown that 56,14% of the extracted pairs are correctly identified as lexical combinations. From these, 13,99% are combinations where exists some kind of lexical restriction, and 42,15% are free combinations.

Identificación/extracción semiautomática de colocaciones utilizando métodos contrastivos

RESUMEN

Este trabajo, que se encuadra dentro de las áreas de investigación de la lingüística computacional y la lexicografía, se propone investigar la viabilidad de un algoritmo para la extracción automática coherente y consistente de colocaciones a partir de la comparación de dos lenguas. Para este experimento se eligieron muestras en español y en portugués, y, a partir de un corpus del Banco Central Europeo, se obtuvieron 46.089 pares de candidatos a colocación. Estos pares de candidatos fueron revisados y evaluados manualmente, utilizando métodos contrastivos. Además de la evaluación, se hizo un análisis detallado de los distintos tipos de errores producidos, para posibilitar correcciones e mejoras posteriores del algoritmo. Concluido el análisis, se comprobó que el 56,14% de los pares corresponden a combinaciones lexicales correctamente identificadas, de las cuales el 13,99% eran combinaciones en que operaba algún tipo de restricción lexical y el 42,15% eran combinaciones libres.

ÍNDICE

ÍNDICE DE ABREVIATURAS

DEC	Dicionário explicativo e combinatório
TST	Teoria Sentido-Texto
MI	Mutual Information
PMI	Pointwise Mutual Information
NPMI	Normalized Pointwise Mutual Information
NMI	Normalized Mutual Information

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

Introdução

> PEOPLE SPEAK IN SET PHRASES – rather than in separate words; hence the crucial importance of set phrases. At the same time, set phrases, or phrasemes, represent one of the major difficulties in theoretical linguistics as well as in dictionary making. (Mel'čuk, 1998: 1)

Nos últimos séculos, tem-se assistido a um interesse crescente da Linguística, mais concretamente da Lexicografia e da Linguística Computacional, pelas unidades fraseológicas da língua, em especial as colocações. Igor Mel'čuk é um dos muitos linguistas que despertaram e impulsionaram o interesse por esta área.

Muitos investigadores defendem que as colocações são meras sequências frequentes de palavras, no entanto a maior parte dos linguistas contesta esta conceção das colocações, defendendo que as colocações são, realmente, combinações frequentes, mas não é isso que as define, visto que existem combinações lexicais livres que também são frequentes.

Semanticamente, as combinações lexicais restritas caracterizam-se pelo facto de o seu significado ser diferente do da soma dos significados dos elementos que as constituem. Porém, a fronteira entre as combinações livres e as combinações restritas não é tão clara como se possa pensar.

A linguística teórica foi estabelecendo, ao longo dos últimos anos, uma série de testes, baseados em critérios morfossintáticos, para ajudar a delimitar o segmento de enunciado que corresponde a uma unidade pluriverbal (combinação lexical restrita), face a outras combinações livres de palavras.

Nos últimos anos, a Linguística Computacional tem-se interessado por esta área da Lexicografia e tem desenvolvido diferentes métodos de extração de colocações para que os investigadores possam identificar, de forma mais rápida, as colocações presentes em corpora de grandes dimensões.

Este projeto de investigação insere-se neste quadro teórico e tem como principal finalidade a identificação e extração semiautomática de colocações em duas línguas, o espanhol e o português, visto que ainda são escassos os trabalhos deste género no nosso país.

Desta forma, o projeto de investigação terá três objetivos principais.

O primeiro objetivo é estabelecer métodos contrastivos para a identificação de colocações num determinado corpus. Por sua vez, o segundo objetivo centrar-se-á na aplicação de testes baseados em métodos contrastivos com corpora bilingues. O terceiro e último objetivo prende-se com a análise da possibilidade do uso de dicionários de tradução para inferência/classificação do tipo de colocação.

A presente dissertação encontra-se estruturada em quatro capítulos.

No primeiro capítulo, intitulado *Revisão da literatura,* farei uma revisão das duas principais abordagens na área das colocações, a abordagem estatística e a abordagem fraseológica. Apresentarei a conceção de colocação em cada uma das abordagens e, ainda, alguns autores representativos destas abordagens.

No segundo capítulo, que tem por título *Delimitação do conceito de colocação,* diferenciarei as colocações das restantes unidades fraseológicas da língua. Numa primeira parte, são apresentadas algumas características das colocações, enunciadas por diversos autores nos últimos anos, que ajudam a delimitar este conceito. De seguida, distinguirei as colocações das combinações livres e das expressões idiomáticas, tendo por base as características apresentadas anteriormente. Por fim, abordarei um conceito muito importante no estudo das combinações lexicais na linguística europeia do século XX, o conceito de solidariedade lexical.

No terceiro capítulo, cujo título é *Técnicas de extração de colocações,* numa primeira parte, abordarei a questão da extração automática de colocações, explicando em que é que se baseia e apresentando alguns conceitos fundamentais nesta área. De seguida, apresentarei, ainda, a constituição e a natureza do corpus de análise, explicando o algoritmo utilizado neste trabalho e as bases em que ele se sustenta. Por último, explicarei como foi feita a análise dos candidatos a colocações, apresentando as diferentes categorias em que estes foram classificados durante a sua análise.

No último capítulo, intitulado *Análise dos resultados obtidos,* farei uma apresentação dos resultados obtidos, divididos por categorias, evidenciando os problemas e erros encontrados na extração automática de colocações. Por fim,

apresentarei as percentagens e os números de ocorrências que cada uma das categorias obteve e uma ilação final acerca dos resultados obtidos.

Por último, na *Conclusão*, apresentarei uma síntese dos principais pontos de interesse desta investigação, os seus limites e problemas e, ainda, as perspetivas de investigação futura dentro da área.

Capítulo I

Revisão da literatura

Ao longo das últimas décadas as colocações têm estado no centro de muitos estudos linguísticos, no entanto muitas são as dúvidas que ainda subsistem em torno destas unidades léxicas, do seu significado, da sua classificação, etc.

Dentro deste quadro teórico existem duas abordagens fundamentais, a primeira abordagem baseia-se, essencialmente, em critérios estatísticos, enquanto a segunda aproximação adota uma perspetiva mais fraseológica, considerando a frequência uma característica secundária e não obrigatória na definição das colocações.

O presente capítulo irá apresentar de maneira sucinta estas duas abordagens.

2.1. Aproximação estatística

A aproximação estatística apareceu com a linguística computacional e a linguística de corpora. As evoluções tecnológicas que se deram nos últimos séculos trouxeram novas dimensões da análise linguística, principalmente a possibilidade de se fazer análises quantitativas de dados, que em muito contribuíram para o aparecimento e desenvolvimento desta abordagem.

Moreno (2009), de quem tomei as descrições que apresento nas páginas seguintes sobre esta aproximação, afirma que esta abordagem sustenta, apoiando-se em dados quantitativos extraídos de corpus informatizados, a ideia de que a colocação é uma combinação de duas ou mais palavras que coocorrem com uma frequência estatística muito significativa, quer isto dizer, combinações frequentes de palavras e relativamente fixas.

Foram sobretudo os linguistas ingleses que adotaram e defenderam esta definição puramente estatística da colocação, entre eles encontram-se nomes como John Firth, Michael Halliday e John Sinclair.

Na área das colocações, John Firth (*apud* Alonso Ramos, 1993: 145) é apontado por muitos investigadores como o primeiro autor a falar de colocações, como objeto de estudo linguístico, para referir-se a combinações habituais de palavras. Muitos chegam mesmo a afirmar que foi Firth que aumentou o interesse por esta área. No entanto, tal como em muitos outros aspetos desta área, não existe unanimidade sobre quem introduziu realmente este termo[1].

Segundo Moreno (2009), as investigações deste linguista inglês centraram-se essencialmente no contexto e no valor fundamental que o significado tem na linguagem. Firth na sua investigação defendeu a ideia de que a língua devia ser estudada como um fenómeno social, tendo em consideração o contexto onde as palavras se inserem e as relações que estabelecem. De forma a sustentar esta ideia, Firth criou uma teoria do significado onde defende que as colocações não devem ser vistas como entidades individuais, mas sim como unidades cujo valor depende das relações que estas estabelecem com as palavras que as rodeiam, visto que, tal como afirma "you shall know a word by the company it keeps" (Firth, 1957: 195 *apud* Moreno, 2009: 16).

John Firth foi considerado por muitos o primeiro percursor da abordagem estatística, pois apesar de na altura em que resolveu estudar as colocações não ter ao seu dispor muitas das ferramentas de processamento automático de texto que apareceram mais tarde, este autor chegou a dividir as colocações em "colocaciones usuales y inusuales", como atesta Alonso Ramos (1993: 142), dando assim a ideia de que existiram colocações mais frequentes do que outras.

Porém, apesar de ter sido um pioneiro nesta área, uma das criticas que foram feitas ao estudo de Firth foi, sem dúvida, o facto de este não ter chegado a dar uma definição clara do que ele próprio entendida por colocação.

As ideias deste autor foram sendo aprofundadas e defendidas por diferentes investigadores ao longo dos anos, contudo as investigações com maiores repercussões foram sem dúvida as de Michael Halliday e depois de John Sinclair.

Para Halliday, as colocações dizem respeito ao domínio do léxico e não ao domínio da gramática. Desta forma este autor defendeu que a teoria lexical não devia

[1] Segundo Mel'čuk (*apud* Alonso Ramos, 1994-1995: 9) foi o russo Vinogradov que chamou atenção para este termo. Por sua vez, segundo Mitchell (*apud* Alonso Ramos, 1994-1995:9), o termo colocação foi introduzido por H.E. Palmer.

fazer parte da teoria gramatical, como se dizia até aquela altura. Para este investigador, a teoria semântica e a teoria gramatical deveriam complementar-se, pois nenhuma destas teorias funciona completamente sozinha.

Muitos têm sido os investigadores que ao longo destes últimos anos têm defendido que a gramática não domina todas as áreas, como se pensava. Entre eles encontra-se Michael Lewis. Lewis (1993) considera que as colocações são apenas um dos muitos aspetos que demonstram que o conhecimento não se pode basear apenas na gramática, o léxico é um aspeto fundamental no ensino duma língua. Aliás, este investigador defende que, no que diz respeito ao ensino duma língua estrangeira, se deve dar primazia ao léxico e não à gramática.

> The Lexical Aproach implies a decreased role for sentence gramar, at least until post-intermediate levels. In contrast, it involves an increased role for word grammar (collocation and cognates) and text grammar (suprassentential features). (Lewis 1993: 3)

> The primary purpose of language is the creation and exchange of meaning. Basic, or proto-language essentially involves nominalization – naming of concepts – and is lexical rather than grammatical. Language consists of grammaticalised lexis, not lexicalised grammar. (Lewis 1993: 51)

Nos seus estudos, Halliday utiliza as colocações como um meio de agrupar o vocabulário em conjuntos lexicais, que podem ser discutidos em termos estatísticos. Para estabelecer o contexto colocacional duma unidade lexical, Halliday, segundo Alonso Ramos (1993), afirma que se deve observar a frequência dessa unidade num determinado contexto em relação com a sua frequência total. Sendo assim, se duas palavras tiverem grande probabilidade de estarem acompanhadas pelos mesmos colocados podemos considerá-las membros do mesmo conjunto léxico.

Segundo esta abordagem, Halliday afirma que uma colocação é

> The syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at *n* removes (a distance of *n* lexical items) from an item *x,* the items *a,b,c,* … Any given item thus enters into a range of collocation, the items with which it is collocated being ranged from more to less probable. (Halliday, 1961:276 *apud* Moreno, 2009: 19)

Quer isto dizer, que as colocações são simples combinações frequentes de palavras onde a sua categoria gramatical não interfere, sendo assim apenas um fenómeno meramente probabilístico.

A influência de Firth nos estudos de Halliday deixou-se notar no facto de este considerar a colocação sob um ponto de vista léxico, mas também no interesse por questões relacionadas com o cotexto e com a análise textual. Este interesse pelos aspetos textuais fez com que mais tarde Halliday e Hasan (*apud* Moreno, 2009: 18) aprofundassem o termo de colocação e afirmassem que esta é:

> a cover term for the cohesion that results from the cooccurrence of lexical items that are in some way or other typically associated with one another, because they tend to occur in similar environments. (Halliday e Hasan, 1976: 287 *apud* Moreno, 2009: 18-19)

Esta definição de colocação suscitou muitas críticas a Halliday, pois muitos investigadores consideravam que o facto de num mesmo texto aparecerem várias palavras relacionadas com um mesmo tema não tinha diretamente a ver com a colocação, mas sim com fatores semânticos e discursivos que são alheios a este fenómeno fraseológico da língua, tal como afirma Coseriu quando diz que:

> La frecuencia de la combinación del adjetivo *blanche* con el substantivo *mouette* depende de nuestro conocimiento de las gaviotas, no de la lengua francesa, ya que 'blanc, blanche' no implica el rasgo distintivo 'para las gaviotas': es por tanto, un hecho de gaviotas, no un hecho de lengua (Coseriu, 1977: 184)

Pode-se, então, concluir que Halliday foi um pioneiro no que diz respeito ao emprego da dimensão paradigmática da colocação para estabelecer relações lexicais numa língua.

John Sinclair foi um dos muitos investigadores que seguiu as ideias introduzidas por Firth e por Halliday. Porém ao contrário dos seus antecessores, Sinclair foi o primeiro linguista a desenvolver uma teoria da colocação, separando-a da teoria contextual onde Firth e Halliday a incluíam, o que lhe atribuiu um estatuto linguístico, que na altura ainda não detinha.

De acordo com Moreno (2009), os estudos de Sinclair foram muito marcados pelos avanços tecnológicos que se deram nos últimos anos. A possibilidade de analisar grandes quantidades de vocabulário com grande rapidez possibilitou uma abordagem muito mais profunda a esta temática do que o vinha a acontecer até então.

Graças a estes desenvolvimentos, este investigador criou um dos maiores projetos computacionais até à atualidade, o projeto COBUILD (*Collins-Birmingham University International Language Database)* que recolhe e analisa cerca de 20 milhões de palavras.

Estas evoluções ajudaram ainda a provar o que Firth e Halliday tinham vindo a afirmar anteriormente, quando diziam que o vocabulário se combinava sistematicamente da mesma maneira e que este se relacionava entre si.

Tal como os autores anteriores, Sinclair utiliza apenas o critério da frequência para determinar o que é uma colocação, deixando de lado as relações sintáticas e semânticas que esta possa estabelecer. Desta forma, Sinclair (1991: 170 *apud* Alonso Ramos, 2010: 2, nota de rodapé 3) afirma que uma "collocation is the occurence of two or more words within a short space of each other in a text".

A este curto espaço Sinclair atribuiu o nome de "span" e fixou-o em quatro palavras para cada lado da palavra-chave, a que apelidou de "node". Contudo, este espaço foi mais tarde aumentado para cinco palavras por diversos investigadores. No que diz respeito às palavras que acompanham o "node", dentro do "span", Sinclair atribuiu o nome de "collocates". Para este investigador todos estes elementos encontram-se ao mesmo nível, contrariamente ao que vai acontecer na abordagem fraseológica, como se verá mais abaixo.

Para Moreno (2009), Sinclair defendeu ainda que uma grande parte da linguagem é constituída por frases semiconstruídas, contrariamente à ideia que se vinha a defender até esta altura, que defendia a criatividade total no sentido de que o falante é sempre livre de selecionar qualquer item ao construir o discurso. Esta visão fez com que este autor estabelece-se dois princípios fundamentais para a organização da língua, o princípio de livre escolha (*"open-choice principle"*) e o princípio idiomático (*"idiom principle"*), princípios que se complementam.

O princípio de livre escolha consiste no facto do falante ter total liberdade para escolher todas as palavras e orações que produz para preencher uma série de lacunas,

mediante uma tomada contínua de decisões, para a qual a única restrição é a gramaticalidade. No que diz respeito ao princípio idiomático, Sinclair diz que este princípio limita as possibilidades de escolha do falante, pois trata-se da faculdade da linguagem através da qual um falante tem ao seu dispor um largo número de frases semiconstruídas.

Apesar de Firth e Halliday terem contribuído para o desenvolvimento da abordagem estatística, o autor mais importante desta abordagem foi, sem dúvida alguma, John Sinclair. Contudo, este autor, tal como outros, foi muito criticado por se basear apenas em dados estatísticos.

Pode-se afirmar, assim, que o grande problema desta abordagem reside no facto desta colocar a frequência como critério único para determinar se se trata duma colocação ou não, pois isto faz com que combinações de palavras que só estão juntas por razões semânticas ou por refletirem situações naturais da vida sejam consideradas colocações. Acerca desta assunto, Coseriu (1977: 160) afirmou que: "la probabilidade estadística general de las combinaciones no tiene prácticamente nada que ver con las solidariedades y no es prueba de su existência".

Outro dos problemas de usar a frequência como critério único, para Moreno (2009), é o facto de algumas colocações, que possam ser interessantes de analisar, passem despercebidas ao investigador por aparecerem com pouca frequência no corpus escolhido, por razões diversas. Porém, existe o reverso da moeda, o facto de esta ter uma frequência muito baixa é de grande utilidade para o ensino de língua.

2.2. Aproximação fraseológica

A aproximação fraseológica opõe-se à abordagem estatística e aos critérios que esta utiliza para definir uma colocação. Os autores que defendem esta teoria defendem que as colocações não são meras associações de palavras quantificáveis estatisticamente, como afirmava a abordagem estatística. Esta abordagem defende que as colocações se caracterizam por um conjunto de propriedades que fazem delas unidades fraseológicas, quer isto dizer, combinações com um certo grau de fixação e arbitrariedade combinatória.

Conforme Moreno (2009), nesta teoria existe, desde sempre, uma procura incessante duma classificação exaustiva de todas as unidades fraseológicas que existem nas línguas, onde se estabelecem diferentes critérios para as diferenciar entre elas.

Muitos foram os investigadores que criticaram a abordagem estatística, principalmente o facto de a probabilidade ser o critério único de seleção duma colocação, e decidiram adotar uma postura diferente, sendo que acreditavam que as colocações não eram só combinações que apareciam frequentemente juntas, mas também combinações que tinham determinadas características sintáticas e semânticas.

O investigador que mais defendeu e divulgou esta abordagem foi, sem dúvida alguma, I. Mel'čuk, investigador que aprofundou os estudos na área das colocações e da lexicografia em geral.

Segundo Alonso Ramos (1993), este investigador criou um modelo lexicográfico que originará o *Dicionário explicativo e combinatório,* daqui em diante DEC, em primeiro lugar aplicado à língua francesa e à língua russa e que está vinculado a uma teoria linguística desenvolvida por Zholkovsky e Mel'čuk num trabalho de 1967. A esta teoria deu-se o nome de Teoria Sentido-Texto (TST) e trata-se duma teoria orientada para o léxico.

O DEC tem como principal característica o facto de estar idealizado para fornecer toda a informação necessária para exprimir um determinado sentido num determinado contexto, sendo desta forma um dicionário orientado para como o falante pretende dizer alguma coisa e não para o que determinada palavra significa. Ao contrário do que se vinha a fazer até aquela altura, o DEC tem por objetivo descrever as lexias[2] da língua, sejam elas lexemas ou frasemas, como afirma Mel'čuk quando diz que:

> Un DEC se doit de décrire non seulement les lexèmes de la langue L mais encore tous ses phrasèmes (dont le nombre est nettement supérieur à celui des lexèmes), en devenant de ce fait un dictionnaire de MOTS et de PHRASÈMES, donc un dictionnaire de LEXIES. (Mel'čuk & Polguère, 1995: 45)

[2] Para saber mais sobre a questão consultar Melcuk & Polguère (1995: 15 – 17)

Este investigador faz a distinção, muito importante, entre dois tipos de combinações lexicais na língua, as combinações livres e as combinações restritas. Afirmando desta forma que todos os lexemas duma língua se podem combinar de forma livre, mas também de forma restrita.

Segundo Alonso Ramos (1993: 190), a coocorrência lexical livre trata-se duma questão essencialmente semântica, enquanto, por sua vez, a coocorrência lexical restrita é uma questão lexical e nem sempre é uma questão semântica.

Para I. Mel'čuk (1998) uma combinação é livre quando:

> A FREE phrase **A⊕B** in language **L** is a phrase composed of lexemes **A** and **B** and satisfying simultaneously the two following conditions:
>
> 1. its signified 'X' = 'A⊕B' is unrestrictedly and regularly constructed on the basis of the given ConceptR – out of the signifieds 'A' and 'B' of the lexemes **A** and **B** of **L;**
>
> 2. its signifier /X/ = /A⊕B/ is unrestrictedly and regularly constructed on the basis of the SemR 'A⊕B' – out of the signifiers /A/ and /B/ of the lexemes **A** and **B**. (Mel'čuk, 1998: 4-5)

Isto quer dizer que uma combinação livre é um sintagma composto por dois ou mais lexemas, cujo significado é resultado da soma regular[3] dos significados dos seus constituintes e o seu significante também é resultado da união linguística dos seus significantes. Estes constituintes podem ser substituídos por expressões sinonímicas desde que respeitem as regras gramaticais e mantenham o significado daquele enunciado.

Mel'čuk (1998, 2013), a quem segui nas páginas seguintes, expôs, nos seus artigos, o que entendia por combinação livre, mas também por combinação restrita.

No que diz respeito às combinações restritas ou frasemas, Mel'čuk (2013: 130) afirma que "un phrasème est un énoncé multilexémique non libre". Sendo assim, um frasema não é nada mais do que uma combinação de dois ou mais lexemas cujo

[3] Para Mel'čuk (1998: 4) a soma regular é:

> The symbol ⊕ is reminiscent of arithmetical summation, but linguistic union is much more complex than simple addition: it presupposes observing ALL general combination rules of L, and this, in conformity with the nature of items being united (signified are united in a different way from signifiers and syntactics, etc.). Thus, X⊕Y denotes the regular union of signs X and Y (= the expression X⊕Y is regularly constructed out of signs X and Y); (X)⊕(Y) is the regular union of signifieds (X) and (Y); etc.

significado e significante não resulta da soma regular dos significados dos seus constituintes, quer isto dizer, que quebra uma ou as duas regras da definição de combinação livre.

Mel'čuk divide as combinações restritas em dois grandes grupos, o grupo dos pragmatemas ou frasemas pragmáticos e o grupo dos frasemas semânticos, que por sua vez podem ser divididos em mais três categorias: os frasemas completos ou expressões idiomáticas, os semifrasemas ou colocações e os quase-frasemas.

No que aos pragmatemas diz respeito, e muito resumidamente, Mel'čuk descreve-os como combinações lexicais cujo significado não é construído livremente, logo os seus constituintes não podem ser substituídos nesse contexto por um sinónimo que seja construído livremente. A restrição destes frasemas é de caráter pragmático visto que são não composicionais ao nível pragmático.

> All ready-made expressions (like greetings, tyical frases used in letters, conversational formulas, technical clichés, proverbs, sayings, etc.), even if semantically and syntactically they are 100 percent compositional, are pragmatemes: they are non-compositional pragmatically. (Mel'čuk, 1998: 6)

Por sua vez, os frasemas semânticos são combinações de dois ou mais lexemas cujo significado não resulta da união linguística dos seus constituintes, mas cujo significante resulta da soma regular dos seus significantes. Com isto Mel'čuk atesta que, ao contrário dos pragmatemas, os frasemas semânticos não têm o seu significado imposto pelo contexto, sendo assim escolhido livremente, mas, por sua vez, a expressão utilizada não é escolhida livremente, sendo a sua seleção restringida pelo significado.

> A semantic phraseme **AB** of L is a set phrase composed of two lexemes **A** and **B** that satisfies simultaneously the following two conditions:
>
> 1. Its signified 'X' is unrestrictedly constructed on the basis of the given ConceptR(SIT) but
>
> **either** it is not regularly constructed out of the signifieds A and B of the lexemes **A** and **B** of **L** ('X' is not a regular sum of 'A' and 'B', i.e., 'X' ≠ 'A ⊕ B'; **or** one of its constituent signifieds is included in the other [for instance, 'A' ⊃ 'B'].

> 2. Its signifier /A ⊕ B/ is not unrestrictedly constructed on the basis of the SemR out of the signifiers /A/ and /B/ of its constituent lexemes A and B (either these signifiers cannot be selected by rules of **L** on the basis of 'X' or the choice of the one is contingent on the other); as for the regularity, in most cases, although far from always, the signifier /A ⊕ B/ is regularly constructed out of /A/ and /B/. [More often than not, /A ⊕ B/ is a regular sum of /A/ and /B/, i.e. /A ⊕ B/ = /A/ ⊕ /B/.]. (Mel'čuk, 1995: 181, *apud* Iriarte, 2001: 173-174)

Esta divisão feita por Mel'čuk é muito similar à que foi feita por Cowie, e posteriormente seguida por Howarth. De acordo com Moreno (2009), Cowie e Howarth dividiram as unidades fraseológicas em dois grandes grupos, as "formulae" e as "composites". Para estes autores, as "formulae" são expressões, geralmente oracionais, cujo significado reflete um valor discursivo, quer isto dizer, que se trata de unidades com uma função essencialmente pragmática, como é o caso dos convites e das saudações. Estas combinações lexicais correspondem aos pragmatemas de Mel'čuk.

Por sua vez, as "composites" são unidades que desenrolam um valor semântico referencial idiomático. É este grupo que corresponde aos frasemas semânticos de Mel'čuk, onde estão inseridas as colocações, e que Howarth também divide em quatro grandes grupos, dispostos ao longo de uma escala dependendo do seu grau de idiomaticidade. No entanto, Howarth[4] denomina estes quatro grandes grupos de "free collocation", "restricted collocation", "figurative idiom" e "pure idiom".

Voltando a Mel'čuk (1998, 2013), este investigador chama de frasemas completos ou expressões idiomáticas às combinações de dois ou mais lexemas cujo significado global não inclui o significado dos seus constituintes, quer isto dizer, o significado global não resulta da união linguística dos seus constituintes. Desta forma, a fórmula utilizada por Mel'čuk para esta combinação é A+B= C[5].

No que diz respeito aos quase-frasemas, o autor define-os como combinações cujo significado global preserva os significados dos seus membros, mas adiciona-lhe ainda novo sentido que não é dedutível através da soma dos significados dos seus lexemas. Quer dizer, que para além do seu significado literal, produto da soma dos seus

[4] Para mais informações consultar Moreno (2009: 30)

[5] A fórmula matemática apresentada por Mel'čuk (2013) é:

'AB' ⊅ 'A' et 'AB' ⊅ 'B'

elementos. A fórmula usada por Mel'čuk que descreve esta combinação é A+B= A+B+C[6].

Por último, e o mais importante para este estudo, Mel'čuk descreve os semifrasemas (ou colocações) como combinações de dois ou mais lexemas, parcialmente fixas, que mantêm o significado inicial de um dos seus constituintes. O elemento que mantém o seu significado inicial é eleito livremente pelo falante, enquanto o outro elemento constituinte é eleito por este e está limitado semanticamente por ele quando aparecem em conjunto. Mel'čuk escolhe a fórmula A+B=A+C[7] para definir esta combinação.

> A COLLOCATION **AB** of **L** is a semantic phraseme of **L** such that its signified 'X' is constructed out of the signified of the one of its two constituent lexemes – say, of **A** – and a signified 'C' ['X'='A⊕C'] such that the lexeme **B** expresses 'C' contingent on **A**. (Mel'čuk, 1998: 7)

Mel'čuk (1998) distingue ainda quatro tipos diferentes de colocações, consoante a natureza do significado 'C'. Nos dois primeiros, o significado que C adota na colocação não aparece no dicionário na aceção de B, enquanto nos dois últimos o lexema B já contém na sua aceção do dicionário o significado 'C'.

O primeiro caso dá-se quando B é uma palavra vazia que é selecionada por A como seu auxiliar, acontece nas colocações formadas por verbos operadores ou verbos-suporte mais um nome. São exemplos deste tipo de colocações: dar um passeio, dar um conselho, etc.

O segundo tipo de colocação é aquele que B tem o seu próprio significado, não está vazio de significado como o primeiro caso, mas só adota o significado 'C' quando combina com o lexema A. Exemplos: sorriso amarelo, chave mestra, etc.

No que diz respeito ao terceiro tipo, Mel'čuk diz que o segundo elemento não está vazio de significado, pois B tem no dicionário o significado 'C' mas este só é

[6] A fórmula matemática apresentada por Mel'čuk (2013) é:

'AB' ⊃ 'A', et 'AB' ⊃ 'B', et 'AB' ⊃ 'C' | 'C' ∩ 'A' = Λ, 'C' ∩ 'B' = Λ

[7] A fórmula matemática apresentada por Mel'čuk (2013) é:

'AB' ⊃ 'A', et 'AB' ⊅ 'B', et 'AB' ⊃ 'C' | 'C' ⋂ 'A' = Λ

atualizado quando está combinado com o lexema A e não pode ser substituído por nenhum sinónimo, fora da colocação B nunca é utilizado com o significado 'C'. São exemplos: ódio mortal, amor fatal, etc.

Por último, temos o caso das colocações onde o significado de B inclui uma parte importante de A sendo então muito específico e totalmente ligado a A, quer isto dizer, que o lexema B não está vazio de significado, mas o sentido 'C' só é atualizado com o lexema A. Os exemplos que representam este tipo são: cabelo louro, nariz aquilino, etc. Segundo Iriarte (2001) e Moreno (2009), este grupo define o que Coseriu (1977: 143-161) chamou de solidariedade lexical, questão que abordarei no próximo capítulo.

Segundo Moreno (2009), Mel'čuk estabeleceu, como já referi acima, a TST, teoria que tinha como principal objetivo criar regras que permitissem obter o maior número de significados expressos por uma só oração. Foi dentro desta teoria que apareceu pela primeira vez a noção de função lexical[8], noção fundamental para Mel'čuk. As funções lexicais foram criadas por Zholkovsky e Mel'čuk e tinham por objetivo descrever as relações léxico-semânticas paradigmáticas e sintagmáticas (*apud* Koike, 2001: 19).

Para Mel'čuk uma função lexical[9] é:

> Du point de vue formel, une ***fonction lexicale*** [=FL] est une fonction au sens mathématique ; elle peut être représentée par la formule traditionnelle :
>
> **f**(x) = y,
>
> où *x* est l'***argument*** de la fonction et *y*, sa ***valeur***. (Mel'čuk & Polguère, 1995: 126)

De acordo com Mel'čuk (1998), a finalidade das funções lexicais é, em primeiro lugar, fazer uma classificação sistemática de todas as colocações existentes, abrindo assim caminho a uma generalização semântica das colocações.

[8] Para mais informações sobre as funções lexicais do DEC consultar Mel'čuk & Polguère (1995: 125-152) e ainda o terceiro e quarto capítulo da tese de Alonso Ramos (1993).

[9] Segundo Iriarte (2001: 179), Mel'čuk & Polguère (1995). apresentam 56 funções lexicais *standard*, enquanto Alonso Ramos apresenta depois, na sua tese de doutoramento, 66 funções *standard*. Para além das funções lexicais *standard,* Mel'čuk & Polguère (1995) estabeleceram ainda algumas funções lexicais não *standard.*

Segundo Moreno (2009), estas funções expressam uma relação de dependência semântica entre a "palavra-chave" e o seu "valor", termos que Mel'čuk utiliza para apelidar aos elementos constituintes duma colocação, que ao contrário do "span" e do "collocate" de Sinclair não se encontram ao mesmo nível. No entanto, nem no que diz respeito à abordagem fraseológica existe consenso quanto aos termos a utilizar para denominar os constituintes duma colocação, visto que autores como Cowie e Howarth, de acordo com Moreno (2009), preferem chamá-los de "base" e "colocado"[10].

Dentro da abordagem fraseológica, conforme Moreno (2009), a "palavra-chave" (ou "base") é caracterizada pela sua autonomia semântica, quer isto dizer, que tem um significado totalmente independente do "valor" (ou "colocado"). A "palavra-chave" é o elemento que é selecionado pelo falante em primeiro lugar e de forma totalmente livre, ao contrário do "valor", cuja eleição está condicionada léxica e semanticamente pela "palavra-chave". Desta forma, pode-se afirmar que a "palavra-chave" é o elemento dominante da colocação visto que é através deste que se seleciona o outro elemento.

Contudo, apesar das funções lexicais terem ajudado muito no estudo das colocações, estas suscitaram muitas críticas, sendo que as maiores críticas se centraram no facto de muitas destas funções produzirem combinações livres em vez de colocações e, ainda, de recolherem relações morfológicas e léxico-semânticas, como é o caso da hiponímia, da sinonímia, entre outras (*apud* Moreno, 2009: 40).

Concluindo, esta abordagem fraseológica veio contrapor as ideias defendidas pela abordagem estatística e tentar provar que a frequência só é um dos critérios para determinar o que é uma colocação. Desta forma, os investigadores que defendem esta abordagem tentaram estabelecer diversos critérios válidos para definir definitivamente as colocações, sendo os mais recorrentes a restrição semântica e a comutabilidade que falarei *a posteriori* (capítulo 2.1).

Tal como a abordagem estatística, esta abordagem também teve algumas críticas e algumas áreas problemáticas, como é o caso dos critérios de qualificação duma colocação, principalmente no que diz respeito ao grau de "obrigatoriedade" destes critérios, segundo Moreno (2009).

[10] Termos introduzidos por Hausmann (*apud* Moreno, 2009: 34) e foram seguidos pela maioria dos autores.

Face a tudo o que foi dito, pode-se concluir que este é um fenómeno muito problemático da língua. Sem dúvida alguma, esta será uma área que, ao longo dos próximos anos, ainda vai dar muito que discutir, devido ao facto de a sua natureza e as suas fronteiras ainda não serem muito claras, o que suscita muito interesse nos mais diversos linguistas.

Capítulo II

Delimitação do conceito de colocação

Apesar de, nas últimas décadas, as colocações terem sido consideradas um fenómeno importante e terem sido estudadas por diversos autores, alguns dos quais referi no capítulo anterior, o conceito de colocação ainda não é aceite unanimemente (Alonso Ramos, 1994-1995: 9) por todos os investigadores que se têm dedicado ao estudo do léxico.

Como vimos, uma colocação pode ser definida como uma unidade fraseológica da língua que é formada por duas (ou mais) unidades lexicais que aparecem regularmente juntas e que tem restrições estabelecidas pelo uso, sendo que um dos seus constituintes tem autonomia semântica (base) e capacidade para selecionar o outro elemento da colocação, o colocativo.

Muito se tem debatido acerca das características formais e funcionais destas unidades fraseológicas, com o intuito de clarificar as diferenças existentes entre estas combinações lexicais e as restantes. Neste capítulo mostrarei alguns critérios utilizados pelos diversos autores para distinguirem as colocações das demais unidades fraseológicas. Depois duma clarificação destes critérios, tentarei distinguir as diferentes unidades fraseológicas entre si. Por último, introduzirei um termo muito importante na área das colocações, o termo de solidariedade lexical, introduzido por Eugenio Coseriu.

2.1. O que são colocações?

Tal como referi no capítulo anterior, Cowie e Howarth defendem, segundo Moreno (2009), que existem dois critérios fundamentais para a identificação e classificação das combinações restritas da língua (*composites*): a transparência semântica e a comutabilidade.

A transparência semântica diz respeito ao facto de as palavras que integram uma combinação terem ou não um significado literal, quer isto dizer, ao facto de o significado duma combinação resultar ou não da soma dos significados dos seus constituintes. Por sua vez, a comutabilidade consiste na possibilidade de substituir um

dos elementos da combinação por outro lexema sem que haja uma alteração no significado dos restantes elementos ou da combinação no seu todo.

Não é nítida a fronteira entre colocações e combinações livres, nem entre as colocações e as expressões idiomáticas. Por esta razão, Corpas Pastor (2001) resolveu descrever um conjunto de características distintivas das combinações lexicais, que se aplicam às colocações e que nos ajudam a estabelecer mais ao menos uma fronteira entre estes fenómenos. Para esta investigadora existem seis características principais: a polilexicalidade, a alta frequência de aparição e coaparição, a institucionalização, a estabilidade (restrição combinatória e especialização semântica), a idiomaticidade, e, por fim, a variação.

Corpas Pastor defende que uma unidade fraseológica é uma "combinación estable de unidades léxicas formadas por al menos dos palabras gráficas, cuyo límite superior se sitúa en el nível de la oración compuesta" (Corpas Pastor, 1998: 167 *apud* Corpas Pastor, 2001: 91).

Desta forma, para Corpas Pastor, as colocações podem ser entendidas como unidades fraseológicas pois são constituídas por pelo menos duas palavras gráficas lexicais, sendo que podem existir ainda colocações constituídas por duas palavras léxicas e uma palavra gramatical ou ainda colocações com mais de duas palavras léxicas.

No que diz respeito à questão da frequência, que é defendida por alguns autores como o critério de seleção mais importante, Corpas Pastor defende que os elementos constituintes duma colocação aparecerem frequentemente juntos no discurso, no entanto, para esta autora, este não é um critério muito importante, tal como Alonso Ramos defende:

> El hecho de que dos lexemas coocurran frecuentemente no es prueba de que exista una colocación. Ya hemos visto que la coocurrencia de muchos lexemas está determinada por su significado y esto es independiente de que ambos lexemas aparezcan frecuentemente en los textos (Alonso Ramos, 1993:146)

Corpas Pastor afirma ainda que para além da frequência com que coaparecem, estas combinações também se caracterizam por serem frequentes no seu conjunto, como unidades.

A terceira característica, a institucionalização, é a característica que marca as colocações como unidades fraseológicas, quer isto dizer, é a fixação na norma destas unidades, em função da sua reprodutibilidade no discurso. Desta forma, as colocações são reconhecidas pelos falantes como combinações familiares e estes utilizam-nas como construções pré-fabricadas. Esta característica, segundo Corpas Pastor (2001: 92), coloca-nos na "dimensíon psicolingüística de la colocación, ante su realidad cognitiva", visto que "las colocaciones de una lengua parecen estar almacenadas como unidades en el lexicón mental de los hablantes".

Por sua vez, a estabilidade divide-se em duas variáveis, a restrição combinatória e a especialização semântica. Tanto num caso como no outro existe, segundo Corpas Pastor (2001), uma escala gradual assimétrica, visto que um dos constituintes apresenta uma maior restrição que o outro. Desta forma, no que diz respeito à restrição combinatória, um dos constituintes da colocação pode-se colocar com o outro constituinte e com um número reduzido de sinónimos nesse mesmo contexto, enquanto o outro elemento se pode combinar com um grupo mais vasto de palavras.

Por outro lado, no que diz respeito à especialização semântica, esta supõe ou uma supressão ou uma adição do significado de um dos constituintes da colocação. Nos casos de supressão de um significado, as combinações estabelecem-se com um verbo deslexicalizado ou de suporte, onde este perde o seu significado original e adquire um "significado general y gramaticalizado, funcional y auxiliar" (Corpas Pastor, 2001: 93), exemplos deste caso são as combinações em espanhol *prestar juramento (*prestar juramento) ou *prestar atención* (prestar atenção). No entanto, o que acontece quando há uma adição de um significado é diferente, visto que existe uma evolução semântica metafórica. Um dos exemplos utilizado por Corpas Pastor é a combinação lexical *levantar castigo,* visto que interpreta que o castigo se conceptualiza como um objeto pesado que impede qualquer movimento.

Uma das propriedades mais importantes das colocações é a idiomaticidade, propriedade semântica das combinações lexicais em que o significado global da combinação lexical não resulta da soma dos significados dos seus constituintes. Ao contrário do que acontece nas expressões idiomáticas, onde o significado nada tem a ver com a soma dos significados dos seus constituintes, nas colocações cada uma das

palavras que a constituem contribui de forma individual para formar o significado global da colocação.

A última característica que Corpas Pastor (2001) atribui às colocações é a variação, que está relacionada com as variantes linguísticas e com a manipulação discursiva. Corpas Pastor afirma que as colocações podem apresentar variantes diatópicas, variantes diafásicas, variantes diastráticas e, ainda, diversas modificações feitas pelos falantes, que alteram de alguma forma a colocação.

Tal como Corpas Pastos, Koike (2001)[11] também definiu que as colocações tinham seis características fundamentais, que as distinguem das outras combinações lexicais: a coocorrência frequente dos seus constituintes, as restrições combinatórias, a composicionalidade formal, o vínculo entre os dois lexemas, a relação típica entre os seus componentes e a precisão semântica da combinação.

Para além da descrição das características que distingam as colocações das restantes combinações lexicais, ao longo destes últimos anos os estudiosos começaram a dividir as colocações em dois grandes grupos: as colocações gramaticais e as colocações lexicais. As primeiras são combinações que na sua base têm uma palavra dominante (verbo, nome ou adjetivo) enquanto o seu colocado é normalmente uma preposição, por sua vez, as colocações lexicais são constituídas por duas palavras da classe aberta, quer isto dizer, são combinações constituídas por verbos, nomes, adjetivos ou advérbios. No entanto, segundo Moreno (2009), as colocações lexicais foram as mais estudadas ao longo destes anos. Isto fez com que fossem criadas diversas classificações para as colocações, quer lexicais quer gramaticais, de forma a tentar organizá-las. Tipologias como as de Benson et al. (1986) e de Hausmann (1989) (*apud* Moreno, 2009) são apenas dois dos muitos exemplos que existem e que representam o imenso interesse por esta área.

[11] Para mais informações sobre estas características ver Koike (2001, 25-29).

2.2. Colocações, combinações livres e expressões idiomáticas

Distinguir colocações, expressões idiomáticas ou mesmo combinações livres é uma das tarefas mais difíceis, até mesmo os mais diversos investigadores que estudam esta área têm imensas dificuldades em fazê-lo.

A colocação é apontada por muitos autores como a categoria que se encontra a meio caminho entre as expressões idiomáticas e as combinações livres, quer isto dizer, a colocação encontra-se entre a comutabilidade e literalidade total e o significado não composicional e a comutabilidade praticamente nula.

2.2.1. Colocações e combinações livres

As diversas características que têm vindo a ser apontadas às colocações ajudaram em muito na tarefa de distinguir as colocações das expressões idiomáticas e mesmo das combinações livres. Desta forma, podemos distinguir as combinações livres das colocações tendo em conta algumas das características atribuídas pelos diversos investigadores.

Koike (2001), de quem tomei as descrições que apresento nos parágrafos seguintes, distingue as colocações das combinações livres tendo em conta as características que tinha delimitado para as colocações.

Em primeiro lugar, os elementos que conformam as combinações livres não coocorrem tão frequentemente como os constituintes das colocações, quer isto dizer, os sintagmas livres são menos estáveis do que as colocações, visto que a coocorrência dos seus componentes está ao cargo da arbitrariedade do falante.

De seguida, as combinações livres apresentam menor grau de restrição do que as colocações, enquanto as colocações apresentam preferências de combinação e restrições impostas pelo uso, as combinações livres são construídas livremente pelo falante a partir das regras gramaticais e semânticas. As colocações distinguem-se então dos sintagmas

livres porque violam a propriedade paradigmática, quer isto dizer, os constituintes das combinações livres são selecionados livremente, não são de forma alguma restringidos, e até podem ser substituídos por outros lexemas, desde que não alterem o significado da combinação livre.

As combinações livres distinguem-se ainda das colocações por apresentarem maior flexibilidade combinatória, morfológica e sintática, ao passo que as colocações são mais flexíveis quando comparadas com as expressões idiomáticas.

Por último, as combinações livres não apresentam nenhuma relação típica entre os seus constituintes, como encontramos nas colocações. Corpas Pastor (2001: 103) afirma que "los miembros de las colocaciones reflejan la relación típica, y, por tanto, verdadera que mantienen los colocados en el mundo real.". Os exemplos que Corpas Pastor (2001) utiliza para mostrar esta diferença são *cargar una pistola* e *lavar una pistola*, visto que o primeiro caso se trata duma colocação e a segunda duma combinação livre, pois entre *lavar* e *pistola* não existe nenhuma relação típica.

Bahns (*apud* Koike, 2001: 30), constata ainda que as colocações são "fáciles de memorizar y psicologicamente destacadas (*psychologically salient)* a diferencia de las combinaciones libres.".

Pode-se então concluir que um binómio léxico poderá formar uma colocação quando existir uma certa restrição arbitrária na sua combinatória, enquanto uma combinação livre será formada unicamente através das leis gramaticas e semânticas da língua e poderá substituir os seus elementos por sinónimos, desde que não altere o significado da combinação. (*apud* Moreno, 2009: 59)

2.2.2. Colocações e expressões idiomáticas

Tal como acontece entre as colocações e as combinações livres, as colocações e as expressões idiomáticas têm aspetos que as distanciam e que nos ajudam de certa forma a separá-las. Porém, segundo Alonso Ramos (1993: 13), "encontramos una gran confusión en la bibliografia entre colocaciones y lo que nosotros llamamos frasemas.".

Koike (2001) aponta várias diferenças entre as colocações e as expressões idiomáticas, dividindo-as em três categorias: as transformações sintáticas, a coocorrência e fixação estrutural e as diferenças semânticas.

No que diz respeito à sintaxe, as colocações apresentam maior flexibilidade formal do que as expressões idiomáticas. Desta forma, as diferenças formais que separam uma colocação duma expressão idiomática são[12]:

a) Modificação adjetival

hacer un aterrizaje – hacer un aterrizaje forzoso

*tomar tierra - *tomar una tierra forzosa*

Contudo, segundo Koike, existem algumas expressões idiomáticas que aceitam certas modificações, que podem ser expressas através de quantificadores ou intensificadores:

Para que la fiesta resulte un éxito hay que pulsar muchas teclas.

b) Pronominalização

Asumió el cargo de alcalde, pero su repentina enfermedad le impidió desempeñarlo. – desempeñar un cargo.

*lavárselas - lavarse las manos

No entanto, convém referir que também existem frasemas completos que admitem pronominalização, como podemos verificar nos exemplos *tomar el pelo (tomárselo)* e *empinar el codo (empinarlo)*, que foram apresentados por Bosque (*apud* Koike, 2001: 32).

c) Relativização

Este libro marca la línea que deben seguir sus partidários. – seguir la línea

**El ojo que acabo de echar a esse vestido…* - echar el ojo a algo

[12] A maior parte dos exemplos foram retirados de Koike (2001: 31-33)

d) Passivação

El órgano fue trasplantado. – trasplantar un órgano

**El bulto fue escurrido.* – escurrir el bulto

e) Nominalização

trasplantar un órgano - *El trasplante del órgano.*

escurrir el bulto - **El escurrimiento del bulto.*

f) Extração de um componente

Llevaba allí una hora. Había intentado anudar una conversación que el Cubano esquivaba por timidez. Quería, sin embargo, retenerle, como si su presencia diese brillo a la taberna, un brillo del que nadie era testigo. – dar brillo

??Si esse niño no deja de hacer ruido, voy a tener que calentarle las orejas, las grandes orejas que tiene. – calentar las orejas.

g) Outros

O facto de as colocações terem mais flexibilidade morfológica e morfossintática que as expressões idiomáticas possibilita, ainda, as seguintes modificações:

trabar amistad – *muchas amistades, amistad trabada*.

pagar el pato – *pagar los *patos, *el pato pagado*.

No entanto, para além destas modificações, Iriarte (2001) refere outras restrições sintáticas que a colocação admite e a expressão idiomática não aceita. Entre elas encontram-se a adjetivação participial, a modificação adverbial, a quantificação[13].

h) Adjetivação participial

perder a cabeça - **A cabeça perdida…*

prestar atenção – *O ministro agradeceu a atenção prestada.*

[13] Os exemplos utilizados foram retirados de Iriarte (2001: 166-167)

i) Modificação adverbial

perder a cabeça - **O João perdeu a cabeça intensamente.*

prestar atenção – *O público prestou atenção ininterruptamente.*

j) Quantificação

perder a cabeça - **O João perdeu muito a cabeça.*

prestar atenção – *O público prestou muita atenção.*

Resumindo, as colocações apresentam maior flexibilidade do que as expressões idiomáticas, no entanto não se pode estabelecer regras, pois existem exceções à regra.

No que diz respeito à coocorrência e fixação estrutural, Koike defende que ao contrário do que acontece nas expressões idiomáticas, onde a coocorrência dos seus constituintes é quase sempre imprescindível para manter a sua idiomaticidade, as colocações admitem algumas alterações nos seus constituintes sem que isso altere o seu sentido.

No entanto, Koike afirma que não se pode utilizar apenas esta característica para distinguir as colocações das expressões idiomáticas, pois existem alguns frasemas que também admitem algumas substituições. Sendo assim, para este investigador, esta característica tem de estar ligada à fixação estrutural que caracteriza as expressões idiomáticas e que é "casi total en la locución, mientras que es parcial en la colocación" (Koike, 2001: 34).

Por último, no que às diferenças semânticas diz respeito, existem quatro características que distinguem uma colocação duma expressão idiomática: o número de lexemas implicados, o grau de composicionalidade semântica, o número de significados e, por fim, as relações típicas.

a) Número de lexemas implicados

Enquanto as colocações são formadas por dois lexemas, as expressões idiomáticas são normalmente constituídas por mais de dois lexemas. Ou seja, a colocação consiste, tipicamente, na união de duas unidades lexicais (por vezes com uma terceira palavra gramatical), enquanto a expressão idiomática pode ter mais de duas unidades lexicais.

b) Grau de composicionalidade semântica

Outro dos traços que distingue as colocações das expressões idiomáticas é a composicionalidade semântica que é própria das colocações e que não existe nas expressões idiomáticas. Tal como Koike (2001: 35) afirma "la composicionalidad semántica es relativa en las colocacións, mientras que la locución carece de ella".

O significado das colocações resulta da soma dos significados dos seus constituintes, enquanto o significado das expressões idiomáticas tem um significado idiomático, como o seu próprio nome indica, logo não resulta dessa mesma soma.

c) Número de significados

As colocações, salvo algumas exceções, têm apenas um significado, que, como referimos acima, resulta da soma dos significados dos seus constituintes. No entanto, as expressões idiomáticas (*como baixar a cabeça* ou *dar na cabeça*) podem apresentar dois significados, um significado literal, que resulta da soma dos significados dos seus elementos constituintes, e um significado idiomático, que não é passível de ser explicado através da soma dos seus constituintes. Sendo assim, a expressão idiomática é semanticamente menos transparente que a colocação.

d) Relações típicas

A última diferença apresentada por Koike que nos ajuda a distinguir uma colocação de uma expressão idiomática é a relação típica. As expressões idiomáticas não estabelecem, ao contrário das colocações, uma relação típica entre os seus constituintes, ao contrário das colocações que o fazem sempre.

Apesar de todas estas diferenças apresentadas, existem algumas combinações que num determinado momento funcionam como colocação e noutro como expressão idiomática. Segundo Koike (2001: 36) "en estes casos las locuciones suelen ser los resultados de una metaforización de sus vínculos colocacionales correspondientes."

Para concluir, devo evidenciar que apesar de todas estas tentativas de separar as colocações das expressões idiomáticas e mesmo das combinações livres, a tarefa de separá-las continua a ser uma tarefa muito complicada, senão a mais difícil de todas na área, devido aos limites difusos que existem entre as mesmas.

3.3. Colocações e solidariedades lexicais

Na área das combinações lexicais restritas, muitos foram os termos usados com o mesmo sentido de colocação, porém um dos mais importantes foi, sem dúvida alguma, o conceito de solidariedade lexical[14]. Este conceito foi introduzido por Eugenio Coseriu.

Eugenio Coseriu (1977) fala de certas implicações sintagmáticas que existem entre as palavras, às quais atribuiu o nome de solidariedades lexicais, noção que, para ele, foi introduzida por Porzig nos seus estudos, no entanto este autor nunca chegou a utilizar este termo.

Na sua obra, este investigador define uma solidariedade lexical como uma:

> determinación semántica de una palabra por medio de una clase, un archilexema o un lexema, precisamente, en el sentido de que una clase determinada, un determinado archilexema o un determinado lexema funciona como rasgo distintivo de la palabra considerada. Dicho de otro modo, se trata del hecho de que una clase, un archilexema o un lexema pertenece a la definición semántica de esa palavra, en el plano de las diferencias semânticas mínimas (rasgos distintivos). (Coseriu, 1977: 148)

Coseriu queria com isto dizer que uma determinada classe, um determinado arquilexema ou um determinado lexema[15] funciona como traço distintivo dessa determinada palavra.

Este conceito designa, para Coseriu, um relação orientada num sentido único, quer isto dizer, o significado duma palavra está contido na outra mas o inverso não acontece, sendo assim uma das palavras implica semanticamente a outra, porém não acontece o inverso.

[14] Para um estudo mais aprofundado sobre as solidariedades lexicais veja-se Coseriu (1977: 143 – 161)

[15] Para mais informações acerca do que Eugenio Coseriu considera classe, arquilexema e lexema consultar Coseriu (1977: 146-147)

As solidariedades lexicais são compostas por uma palavra determinante, cujos traços distintivos fazem parte da outra palavra que forma a solidariedade, e por uma palavra determinada, que recebe os rasgos distintivos da palavra determinante. É este critério que este investigador utiliza para diferenciar os mais diversos tipos de solidariedade.

Primeiramente, Coseriu (1977) divide as solidariedades lexicais em dois tipos: as solidariedades unilaterais e as solidariedades multilaterais[16]. No que diz às solidariedades unilaterais a determinação do lexema determinado pela classe, pelo arquilexema ou pelo conteúdo do lexema determinante trata-se duma determinação interna. Por sua vez, nas solidariedades multilaterais o lexema determinado opõe-se aos outros lexemas através do traço distintivo, o que faz com que a determinação do lexema determinado seja externa. Sendo assim, as solidariedades unilaterais funcionam só ao nível sintagmático e as solidariedades multilaterais constituem paradigmas.

No entanto, a distinção mais importante é a que Coseriu estabelece ao distinguir os três tipos de solidariedades, tendo em conta se a determinação dos lexemas determinados corresponde a uma classe, a um arquilexema ou a um lexema. Desta forma, Coseriu divide as solidariedades em três grupos diferentes: as solidariedades lexicais por afinidade, as solidariedades lexicais por seleção e a solidariedades lexicais por implicação.

No que diz respeito à solidariedade lexical por afinidade, Coseriu fala das relações estabelecidas pelas classes. O segundo tipo, a solidariedade lexical por seleção, diz respeito às relações estabelecidas por um arquilexema. Por último, a solidariedade lexical por implicação refere-se às relações estabelecidas entre lexemas.

Tendo em conta tudo que foi dito, posso concluir que o conceito de solidariedade lexical é muito mais restrito do que o de colocação. Este conceito está intimamente relacionado com a colocação, no entanto não são sinónimos.

Como referi no primeiro capítulo, a solidariedade lexical parece ser um tipo de colocação. Segundo Iriarte (2001), o conceito coseriano de solidariedade lexical diz apenas respeito às colocações em que um determinado lexema está incluído como traço

[16] Para saber mais sobre solidariedades unilaterais e solidariedades multilaterais consultar Coseriu (1977: 152-153)

distintivo na definição do outro lexema. Desta forma, todos os outros tipos de colocação seriam, para Coseriu, realizações da norma e nada mais.

Outra das diferenças entre as colocações e as solidariedades lexicais é a questão da frequência, visto que o próprio Coseriu (1977: 160) afirma na sua obra que "la probabilidade estadística general de las combinaciones no tiene prácticamente nada que ver con las solidaridades y no es prueba de su existência (…)".

Apesar de existirem diferenças entre o conceito de solidariedade lexical e colocação, pode-se verificar que os estudos de Coseriu foram um importante contributo para esta área e para o desenvolvimento da mesma.

Capítulo III

Técnicas de extração de colocações

No presente capítulo abordarei a questão da extracção automática de colocações e de alguns termos importantes nesta área.

Depois duma apresentação geral das técnicas de extracção de colocações, apresentarei o corpus que foi utilizado nesta dissertação e ainda o algoritmo usado e todo o processo efetuado para extrair os 46,089 pares que foram analisados neste trabalho. Por fim, mostrarei em que bases assentaram a sua análise.

3.1. Extração automática de colocações

Ao longo dos últimos anos muitos têm sido os estudos com o objetivo de conseguir uma extracção automática de termos e mesmo de colocações. A extracção automática de termos a partir de corpora textuais especializados constitui uma ferramenta fundamental na análise do vocabulário hoje em dia.

Segundo Corpas Pastor (2001: 100), "la extacción automática de colocaciones a partir de corpus extensos ha supuesto un auténtico revulsivo para los estudos teóricos y aplicados de la colocación".

Na extracção de colocações, a principal tarefa tem sido identificar, num determinado corpus, combinações de palavras que mostrem alguma idiossincrasia na sua distribuição linguística.

As técnicas de extração de colocações que são utilizadas atualmente fundamentam-se na informação linguística, mas também em valores estatísticos baseados nas frequências com que estas são utilizadas num determinado corpus. Isto deve-se ao facto de as colocações não serem escolhidas aleatoriamente, o que "justifica el uso de técnicas estadísticas para la explotación de textos en busca de combinaciones

de palabras cuya frecuencia de aparición puede considerarse que no se debe al azar." (Suaréz et al., 2011: 147).

Sendo assim, a extracção de colocações consiste na comparação da distribuição dos seus constituintes através de uma medida de associação. Essa medida de associação é responsável por avaliar e seleccionar os melhores candidatos a colocações.

Segundo Bouma (2009), existem imensas medidas de associação disponíveis atualmente, porém a literatura que existe acerca da extracção de colocações tem o intuito de descobrir novas medidas, cada vez mais eficazes, visto que ainda não existe nenhuma medida de associação que seja a melhor para extrair todas as colocações. Diferentes medidas encontram de maneira eficaz diferentes tipos de colocações, daí ser necessário e útil conhecer o maior número possível de medidas de associação e o seu comportamento.

Segundo Suaréz et al. (2011), a maioria das medidas de associação baseiam-se na relação entre a probabilidade de ocorrência conjunta $P(x,y)$ e as probabilidades individuais da base e dos colocativos, $P(x)$ e $P(y)$.

$$P(x,y) = P(x) \times P(y)$$

Existem três critérios muito importantes no que diz respeito às medidas de associação são eles: a frequência, a significância e a dimensão do efeito[17].

Gerlof Bouma no seu artigo *Normalized (Pointwise) Mutual Information in Collocation Extraction* (2009) fala-nos da extração de colocações e introduz alguns conceitos importantes na literatura da extração de colocações, tais como a *mutual information,* daqui para a frente tratada como MI, e a *pointwise mutual information,* identificada como PMI daqui adiante.

No que diz respeito à MI, esta trata-se duma medida de associação que calcula as probabilidades de $P(x)$ e $P(y)$, e ainda as probabilidades conjuntas de $P(x,y)$. Quer isto dizer, MI mede a informação partilhada por x e y. Esta medida é definida da seguinte forma:

[17] Slides da apresentação "Mutual Information and Collocations" por "Simon Šuster" no Seminário em Estatística e Metodologia em abril 2011"

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} P(x,y) \log\left(\frac{P(x,y)}{P(x) \times P(y)}\right)$$

Por sua vez, PMI é uma medida de associação que calcula a diferença entre a probabilidade real de uma determinada coocorrência *P(x,y)* e a probabilidade esperada dos constituintes da coocorrência em determinados eventos particulares, quer isto dizer, a probabilidade de *P(x)* × *P(y).* Segundo Bouma, PMI foi introduzida na lexicografia por Church e Hanks (*apud* Bouma, 2009: 34):

$$PMI(x;y) = \log\left(\frac{P(x,y)}{P(x) \times P(y)}\right)$$

No entanto, segundo Bouma, existem alguns problemas a resolver, no que diz respeito à PMI e à MI. Para este investigador, o facto de estas não terem um limite máximo acarreta alguns problemas. No caso da PMI em concreto, impossibilita-nos de perceber se determinada coocorrência está perto da correlação máxima ou não, enquanto o contrário é possível visto que uma combinação de palavras sem correlação alguma recebe a PMI de 0.

De forma a resolver este problema, Bouma, no seu artigo, decidiu normalizar estas duas medidas e atribuir o valor máximo de 1, sendo que 1 será atribuído a uma associação perfeita.

Após normalizar a MI e a PMI, aos quais deu o nome de NMI e NPMI, Bouma avaliou estas duas variantes num estudo experimental e chegou à conclusão que a NPMI comporta-se melhor do que a PMI, pois apresenta melhores resultados na extração de colocações, enquanto na NMI não se sucede o mesmo.

Em forma de conclusão, pode-se afirmar que a extração automática de colocações é um método de análise em crescente, que pode ajudar a melhorar a análise das colocações e ainda pode facilitar a análise das mesmas, visto que uma extração automática diminui em muito o trabalho do investigador.

3.2. O corpus

Para este estudo foi utilizado o corpus do Banco Central Europeu, a partir do projeto Per-Fide, desenvolvido na Universidade do Minho por Sílvia Araújo et al. (2010), em específico do par de línguas Espanhol/Português, e selecionaram-se os pares de palavras onde uma das palavras é um adjetivo e a outra palavra é um nome.

Com a colaboração do Doutor Alberto Simões, que implementou o programa informático que utilizei para extrair as colocações, avaliou-se um método para extrair colocações de corpora paralelos, usando a propriedade de tradução não composicional, que será explicada na secção seguinte. A ideia principal deste método será, então, descobrir sequências de palavras cuja tradução não siga a simples tradução palavra a palavra dos seus constituintes.

De seguida apresentarei com maior detalhe o algoritmo que está na base de todo o estudo efetuado nesta dissertação.

3.2.1. O algoritmo

A hipótese testada nesta dissertação consiste na ideia de avaliar se uma sequência constituída por um adjetivo e por um nome pode ser traduzida por outras duas palavras, em que apenas uma delas tem tradução direta da palavra original. Caso isto aconteça considerar-se-á uma colocação.

Esta hipótese pode ser explicada através da sintaxe matemática. Considere-se que *T* define a função que traduz as palavras do espanhol para o português e o ponto é o operador de concatenação de palavras. Desta forma, a tradução de uma combinação *a* e *b* é considerada composicional se $T(a.b)= T(a).T(b)$.

Contudo, no que diz respeito à área da tradução, este é um dos aspetos mais complicados visto que na maior parte das traduções não é isto que acontece. Sendo

assim, nesta dissertação procuraram-se os pares de palavras (*a,b*) cuja tradução não segue a equação apontada acima, isto é

$$T(a.b) = T(a).c \text{ ou } T(a.b) = c.T(b) \text{ em que } c \text{ não é } T(a) \text{ nem } T(b).$$

O objetivo principal deste algoritmo de extração é, então, testar a hipótese que a extração de uma colocação baseada numa tradução não composicional é possível.

Desta forma, o primeiro passo foi descobrir sequências de palavras constituídas por um adjetivo e um nome num corpus paralelo Espanhol/Português. Para que isto fosse possível foi usado o analisador morfológico do FreeLing, desenvolvido por Padró & Stanilovsky (2012) e abordado por Simões e Carvalho (2012), que é composto por várias funcionalidades de análise de texto. No entanto, nem todos os módulos de etiquetação e desambiguação (*part-of-speech*) foram utilizados, o que levantou alguns problemas, que serão referidos mais à frente.

Quando uma combinação é encontrada, os seus conjuntos de tradução são calculados. É preciso ter em atenção que cada uma das palavras que constitui uma combinação pode ter mais do que uma tradução, daí se ter construído um conjunto de traduções. A tradução destas palavras é feita com base no dicionário espanhol/português do motor de tradução automática Apertium (Corbí-Bellot et al., 2005: 79-86).

Os conjuntos de traduções são pesquisados na língua-alvo, quer isto dizer, na tradução portuguesa.

Se qualquer uma das palavras de ambos os conjuntos de tradução coocorrem, essa combinação é descartada, quer isto dizer, se a tradução das duas palavras se mantiverem fazendo com que o significado da combinação seja composicional, a hipótese que aqui definimos não se coloca.

Por outro lado, se uma das palavras tiver a sua tradução na língua-alvo, mas a outra não tiver, a combinação espanhola é mantida para mais tarde ser analisada manualmente, juntamente com um segmento de palavras portuguesas que se encontre na periferia da tradução encontrada.

Os resultados desta análise foram depois analisados manualmente, de forma a determinar se se tratavam realmente de combinações restritas, mais precisamente colocações, ou não.

3.2.2. Análise dos candidatos a colocações

A análise dos resultados foi feita manualmente, utilizando recursos *on-line* como referência, tais como o IATE (Interactive Terminologia para a Europa) de Johnson & Alastair (2000) e dicionários de Espanhol-Português e Português-Espanhol quer em papel, quer on-line.

Depois de aplicado o algoritmo obteve-se um total de 46,089 pares candidatos a colocações.

Numa fase inicial ficou decidido que cada par de palavras iria ser incluído numa das seguintes categorias: erro, combinação livre, colocação e combinação restrita. No entanto, com o decorrer da análise optei por dividi-las apenas em três categorias, a categorias dos erros, das combinações livres e das combinações restritas. Esta alteração deveu-se ao facto do corpus utilizado ser um corpus técnico e por isso mesmo incluir muitos quase-frasemas (conceito explicado no primeiro capítulo desta dissertação).

Sendo assim, a categoria dos erros incluirá os casos em que a combinação lexical espanhola e a portuguesa não têm nada em comum. O principal problema reside no facto do programa utilizado não ter sido capaz de encontrar/extrair a sequência de palavras que inclui a tradução da combinação de palavras selecionada, como será explicado no próximo capítulo.

Por sua vez, o grupo das combinações livres será constituído pelas combinações que foram corretamente traduzidas, mas que não se tratam de forma alguma de colocações. Explicarei no próximo capítulo o porquê do programa ter extraído estas mesmas combinações.

Por último, nas combinações restritas estarão presentes as combinações que estão corretamente transcritas e que são colocações, mas também os quase-frasemas que foram encontrados no corpus. No entanto, apesar de ter juntado os quase-frasemas e as colocações num mesmo grupo, apresentarei exemplos para os dois casos em separado.

Sendo assim, no capítulo seguinte explicarei com maior detalhe cada uma destas categorias e apresentarei vários exemplos de pares classificados em cada uma delas. Depois falarei dos problemas encontrados ao longo da análise e dos aspetos a melhorar neste método. Por fim, apresentarei algumas estatísticas referentes aos resultados obtidos.

Capítulo IV

Análise dos resultados obtidos

Neste capítulo apresentarei, em detalhe, os resultados obtidos, apresentando também um conjunto de exemplos que ilustram as categorias que foram pré-estabelecidas neste trabalho, a categoria dos erros, a categoria das combinações livres e a categoria das combinações restritas.

4.1. Erros

No decorrer da análise da lista de pares candidatos a colocações foram detetados muitos problemas, de várias naturezas, desde erros no próprio algoritmo até ao dicionário utilizado na tradução das combinações lexicais.

Em primeiro lugar, o algoritmo apresentou alguns problemas no que diz respeito à procura das combinações no contexto da versão traduzida, a versão portuguesa, chegando mesmo a quebrar combinações lexicais, o que tornou impossível a análise de alguns exemplos. Quer isto dizer, que diversas combinações lexicais foram truncadas, sendo que uma das palavras aparecia no contexto extraído e a outra não. No entanto, através do contexto, é possível predizer que essa mesma palavra iria ocorrer de seguida no texto. A maior parte dos casos dizem respeito aos pares de palavras em espanhol que foram transformados em combinações lexicais de três palavras em português, sendo que uma delas se trata duma palavra gramatical.

tabaco crudo	Sector de o **tabaco em**
derechos humanos	promoção de os **direitos de**
productos pesqueiros	mercado de os **produtos de**
Seguridad Alimentaria	Europeia para a **Segurança de**
meses siguientes	prazo de três **meses a**

Tabela 1- Exemplos de combinações truncadas.

Como podemos verificar nos exemplos supracitados, apesar de a combinação não estar completa conseguimos perceber que o segundo constituinte da combinação apareceria de seguida no segmento de texto. Este problema poderá ser resolvido num trabalho posterior, aumentando o número de palavras no segmento de texto da língua-alvo.

Dentro deste problema aconteceram ainda casos em que o contexto na língua-alvo não foi extraído na quantidade pretendida e algumas vezes a combinação lexical nem chega a aparecer nesse segmento de texto, como se pode verificar na tabela seguinte:

siguiente información	**o seguinte** :
Autoridad importadora	; **autoridade (s)**
Ciencias biológicas	, **Ciências de**
Cilindrada máxima	: **Cilindrada** :
Diámetro normal	2 **Diâmetro a**

Tabela 2- Exemplos de erros na extração do contexto da tradução.

Estes casos, aqui mencionados, podem ter acontecido por culpa da aplicação de extração, que poderá ter quebrado frases em sítios errados, o que impossibilitou uma extração correta do contexto, ou de alguma das ferramentas usadas na preparação do corpus.

Outro dos problemas encontrados na lista de pares candidatos a colocações foi o facto de o algoritmo não estar preparado para descobrir todas as combinações de palavras existentes num mesmo segmento. O algoritmo utilizado não funcionou corretamente quando existiam dois pares de palavras similares no mesmo segmento de texto. Por diversas vezes, o algoritmo usou o primeiro par de palavras que encontrava, em vez de procurar o par que correspondia à combinação espanhola, tal como se pode comprovar nos seguintes exemplos:

Comunidad Económica	que institui a **Comunidade Europeia**
cantidad superior	, em uma **quantidade inferior**
tiempo completo	de trabalho a **tempo parcial**
tercera columna	referidas em a **coluna 2**
navegación marítima	afectos a a **navegação aérea**

Tabela 3- Exemplos de erros na extração do segmento correto referente às palavras espanholas.

Como se pode verificar na combinação do espanhol *tiempo completo*, o algoritmo em vez de procurar a combinação correspondente em português selecionou uma combinação parecida, a combinação *tempo parcial.* Provavelmente, o corpus tinha na mesma frase estas duas combinações lexicais e isto fez com que o algoritmo selecionasse a que se encontrava em primeiro lugar. Isto deve-se ao facto de o algoritmo ter procurado a combinação através do nome que a compõe e não ter tido em conta o segundo elemento da combinação, por falta de tradução no dicionário usado, como será discutido a seguir.

Também foram detetados alguns erros na lista de pares candidatos a colocação devido ao facto de na língua-alvo, o português, terem sido usados, algumas vezes, pronomes para se referirem a um nome, que foi utilizado numa frase anterior, ao passo que na versão original, o espanhol, se optou por repetir o mesmo nome nas duas frases. A seguinte tabela apresenta alguns exemplos deste problema:

Estado membro	legislação de **esse Estado** ,
ciertos produtos	regime de **esses produtos** .
valores limite	ou de **esses valores** ,
último caso	. Em **esse caso** ,
segundo Estado	legislação de **esse Estado** ;

Tabela 4- Exemplos de utilizações de pronomes na língua-alvo ao invés da tradução direta dos termos.

Apesar de terem aparecido com pouca frequência, existiram outros problemas na nossa lista de pares candidatos a colocação, entre eles o aparecimento de sequências de palavras que não foram corretamente traduzidas, visto que se manteve o espanhol em vez de se ter traduzido para português a sequência, e ainda o aparecimento de alguns casos de sequências de palavras inglesas. A seguinte tabela apresenta alguns dos exemplos encontrados:

medio ambiente	contaminación del **medio ambiente** .
horas extraordinárias	, incluindo as **horas extraordinárias**
Autoridades nacionales	Lista de las **autoridades nacionales**
medio ambiente	en el **medio ambiente** acuático
Vivo Test	In **Vivo Test** for Chromosomal

Tabela 5- Exemplos de sequências noutras línguas que não a portuguesa.

Por último, o facto de se ter utilizado um analisador morfológico em vez de um etiquetador de *part-of-speech*, ferramenta que associa a cada palavra uma categoria gramatical de forma não ambígua, levou a que aparecessem alguns casos de combinações de palavras com verbos, que foram confundidos com nomes ou adjetivos. No entanto, quando bem traduzidas foram consideradas combinações livres, visto que não se tratavam de erros de tradução. O único problema dessas combinações era o simples facto de não serem a combinação pretendida neste trabalho. Contudo, ao se ter optado por usar apenas o analisador morfológico já se esperava que estes casos surgissem. A tabela que se encontra abaixo mostra alguns dos exemplos encontrados na lista de pares candidatos a colocação:

anexo figura	presente **anexo figura** um modelo	combinação livre
conjunto presente	. Se o **conjunto apresentar**	combinação livre
informe figura	este **relatório consta** de o	combinação livre
certificado falla	verificação de o **certificado falhar**	combinação livre
Anexo figura	em o 2092/91 **figura no**	erro

Tabela 6- Exemplos de combinações com verbos.

Como podemos verificar nestes exemplos, as sequências foram incluídas numa determinada categoria consoante a sua tradução, pois não havia motivos para os considerar erros quando são sequências perfeitamente traduzidas e sem qualquer problema.

Antes de falar das combinações livres é importante referir que o corpus utilizado apresenta alguns problemas, ao nível da tradução, na versão portuguesa, o que pode ter influenciado de alguma maneira os resultados aqui apresentados. Isto pode ser um indicador duma tradução feita com pouco cuidado e pouca atenção e que pode de alguma maneira ter criado mais problemas do que os que seriam de esperar inicialmente neste trabalho, como se pode atestar na tabela subsequente:

presente artículo	de o **presnete artigo** ,
legítimo titular	seu **legítimo titual** a ocupar
zona geográfica	específicas em uma **zona geográfirca**
presente Reglamento	. O **presidente regulamento** é
proyectos transnacionales	acompanhamento de os **projectos trannacionais**

Tabela 7- Exemplos de erros na tradução.

Convém também referir que, como em todas as traduções, alguns erros aconteceram devido a algumas decisões tomadas pelo tradutor, que por vezes optou por usar pronomes ou mesmo omitir palavras porque já as tinha referido anteriormente.

4.2. Combinações livres

O grande dilema desta lista de pares candidatos a colocações foi a ocorrência de imensas combinações livres, combinações que foram bem traduzidas mas que não se tratam de combinações restritas. Ao contrário das combinações restritas, em particular das colocações, as combinações livres ocorreram com muita frequência nesta lista de pares candidatos a colocação, sendo inclusivamente a classe com mais ocorrências.

O principal problema que provocou esta situação, que interferiu com a aplicação do algoritmo, e fez com que o algoritmo extraísse muitas combinações livres foi a falta de entradas no dicionário utilizado. Quer isto dizer, o dicionário utilizado para efetuar a tradução por vezes não continha uma entrada para determinada palavra utilizada no corpus.

Este problema levou a que o algoritmo detetasse combinações de palavras onde pelo menos uma das palavras não tinha entrada no dicionário utilizado para fazer a tradução e a selecionasse para uma posterior análise manual, pois não reconhecia a palavra. A tabela abaixo mostra alguns dos exemplos encontrados ao longo da análise:

Segundo resultado	resultado : **segundo resultado** :
primer trimestre	em o **primeiro trimestre** de
presente Directiva	requisitos de a **presente directiva**
primeros párrafos	os dois **primeiros parágrafos** podem
presente apêndice	o **presente apêndice** , os

Tabela 8- Exemplos de combinações em que uma das palavras constituintes não tem entrada no dicionário.

Um problema semelhante ocorreu com as palavras não lematizadas corretamente, que desta forma não foram encontradas no dicionário utilizado na tradução, como se pode verificar na tabela seguinte:

Segunda página	) (**Segunda página** de
tercera fase	de a **terceira fase** de
primera frase	: A **primeira frase** de
segunda columna	em a **segunda coluna** ,
tercera prorroga	de a **terceira prorrogação** de

Tabela 9- Exemplos de palavras cuja tradução não foi encontrada por não lematização.

Outro problema relacionado com o dicionário foi o facto de este ter poucos sinónimos para a palavra em questão, o que levou o algoritmo a extrair combinações lexicais que não são combinações restritas, mas sim combinações livres de palavras. Como na entrada do dicionário utilizado para fazer a tradução alguma das palavras da combinação não tinha o sinónimo apresentado na versão portuguesa, o programa considerou que essas combinações se tratavam de colocações pois a tradução não era, supostamente, composicional.

Este foi o problema mais recorrente na lista de pares candidatos a colocação. A tabela abaixo mostra alguns dos muitos exemplos que foram encontrados no decorrer da análise.

Texto pertinente	EEE) (**Texto relevante**
siguiente texto	a **seguinte redacção** :
Medidas vigentes	**Medidas existentes**
última fabricación	a **última data de fabrico**
Conocimientos sucintos	**; Conhecimentos sumários** de as

Tabela 10- Exemplos de combinações em que a entrada no dicionário de tradução não contém a tradução usada.

Como se pode verificar nos exemplos apresentados, este problema poderia ser resolvido com um dicionário mais amplo, pois o problema reside no facto de o dicionário utilizado ser demasiado incompleto, pois trata-se de um dicionário para tradução automática, em que é importante diminuir ao máximo a ambiguidade de tradução e, desta forma, diminuir o número de traduções possíveis.

Encontramos outros problemas relativos às combinações livres em casos de deixis textual como *criterios anteriores*, *fórmula anterior* ou mesmo *cuadro anterior*, que foram traduzidos por *critérios acima*, *fórmula acima* e *quadro acima*, respetivamente. Isto criou um entrave ao algoritmo, visto que este considerava que as palavras não eram traduções mútuas, nem significavam o mesmo, logo o par devia ser considerado como combinação restrita.

4.3. Reduções e nomes próprios

Com o decorrer da análise manual foi ainda necessário criar duas novas categorias, devido à existência de alguns casos não esperados, mas que apareceram poucas vezes na lista de pares candidatos a colocação. A primeira classe, chamada de redução, diz respeito às situações onde uma combinação de palavras em espanhol é traduzida em português por uma só palavra. Isto pode acontecer porque em português uma dessas duas palavras é frequentemente omitida ou porque a palavra utilizada traduz o significado completo da combinação espanhola. O exemplo mais recorrente nesta categoria é o de *medio ambiente*, visto que em português na maior parte das vezes apenas se utiliza a palavra *ambiente*. A tabela seguinte apresenta este exemplo e outros do mesmo género:

medio ambiente	protecção de o **ambiente** ,
auditoría medioambientales	de ecogestão e **auditoria** (
titular opositor	parte de o **titular** .
trabajo anual	uma unidade de **trabalho** ,
cigarros pequeños	, **cigarrilhas** e cigarros ,

Tabela 11- Exemplos de reduções.

Estes exemplos foram confirmados na base de dados europeia, a IATE (Interactive Terminologia para a Europa).

A segunda classe encontrada, a que se deu o nome de NP (nome próprio), diz respeito aos casos onde as combinações são nada mais do que nomes próprios. Nesta classe estão incluídos casos como nomes de países ou de instituições, como se pode confirmar na tabela seguinte:

Sudeste Asiático	**Ásia de o Sudeste**
Santa Sede	a **Santa Sé** e a
Santa Rosa	**Santa Rosa**
República Federal	originários de a **República Federativa**
continental español	**Espanha continental**

Tabela 12- Exemplos de nomes próprios.

4.4. Combinações restritas

Por último e o mais importante para esta dissertação, a classe das combinações restritas inclui não só as colocações, mas também os quase-frasemas, como referi no capítulo anterior.

O facto de este corpus se tratar dum corpus técnico aumentou a dificuldade de avaliar o que era ou não uma combinação restrita, tornando a tarefa de distinguir o que era uma combinação livre e uma combinação restrita ainda mais complicada do que o que normalmente acontece.

No que diz respeito aos quase-frasemas, combinações lexicais que para além de terem o seu significado literal acrescentam outro significado ligado com a área técnica com a qual estão relacionadas, foram muitos os exemplos encontrados na lista de candidatos a colocação. Na tabela seguinte pode-se verificar alguns dos exemplos encontrados:

Disposiciones legales	em vigor as **disposições legislativas**
mercado interior	funcionamento de o **mercado interno**
cadena alimentaria	Permanente de a **Cadeia Alimentar**
Derecho interno	as disposições de **direito nacional**
derechos humanos	adoptadas segundo os **direitos fundamentais**
contingentes arancelarios	volumes de os **contingentes pautais**
fronteras interiores	um espaço sem **fronteiras internas**

Tabela 13- Exemplos de quase-frasemas.

Existiram ainda situações de combinações de palavras que foram incluídas na categoria das combinações restritas, mais propriamente ligadas aos quase-frasemas, que faziam parte duma combinação restrita mais longa. Estas situações foram, normalmente, fáceis de detetar dado a especificidade do corpus utilizado e ainda devido ao contexto fornecido na língua-alvo, que inclui mais palavras do que as duas palavras que estão presentes na combinação lexical espanhola. De seguida apresento alguns exemplos destes casos, que também foram confirmados na base de dados europeia IATE:

gestión medioambiental	o **sistema de gestão ambiental**
producción homogénea	A **unidade de produção homogénea**
política agrícola	de a **política agrícola comum**
ejecución forzosa	de **medidas de execução forçada**
residuos radiactivos	Gestão de os **resíduos radioactivos**

Tabela 14- Exemplos de combinações restritas que fazem parte de combinações restritas mais longas.

Em comparação com as categorias dos erros e das combinações livres, as combinações restritas ficaram muito aquém dos resultados alcançados por estas. No que diz respeito às colocações propriamente ditas, o que realmente importa nesta dissertação, não foram muitas as colocações encontradas nesta lista de pares candidatos a colocação. A seguinte tabela enumera algumas das colocações encontradas:

caso necesario	, **se for caso de**
días hábiles	prazo de três **dias úteis**
gran velocidade	passageiros de **alta velocidade** que
persona física	" , a **pessoa singular**
peso neto	100 quilogramas de **peso líquido**
años naturales	período de dois **anos civis**
precio neto	aplicável a o **preço líquido**
personas jurídicas	Indicar se as **pessoas colectivas**
amarillo oscuro	propagação certificados e **amarelo torrado**
segunda mano	automóveis em **segunda mão** ,
petróleo crudo	ou resíduos de **petróleo bruto**

Tabela 15- Exemplos de colocações.

4.5. Considerações finais

O gráfico e a tabela que se seguem sintetizam o que foi discutido neste capítulo, apresentando as percentagens e o número de ocorrências de cada uma das categorias estabelecidas neste trabalho.

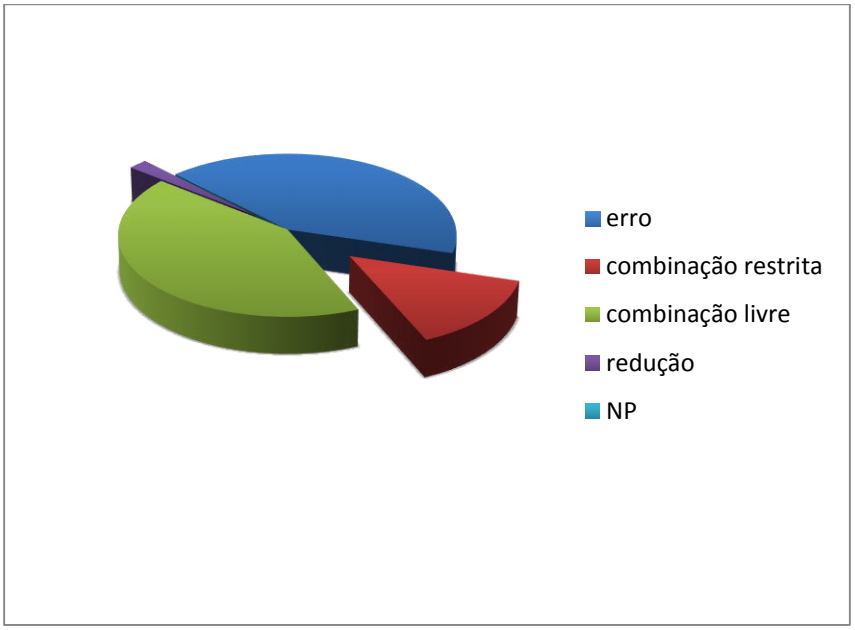

Figura 1 – Distribuição das diferentes categorias.

Categoria	Número de ocorrências	Percentagem
Combinação livre	19 428	42,15 %
Combinação restrita	6 447	13,99 %
Erro	19 281	41,83 %
Redução	914	1,98 %
NP	19	0,04 %

Tabela 16- Valores absolutos e percentuais da distribuição das várias categorias.

Atentando aos dados fornecidos, pode-se concluir que as combinações livres foram as que mais apareceram nesta lista de pares candidatos a colocação, porém convém também referir que a categoria dos erros não ficou muito distante das combinações livres, o que mostra que este algoritmo precisa de ser trabalhado e melhorado para que se possam atingir fins mais satisfatórios em trabalhos futuros, na extração semiautomática de colocações.

Juntando as combinações livres e as combinações restritas podemos perceber que os dois grupos juntos perfazem 25 875 ocorrências, que equivale a 56,14 %.

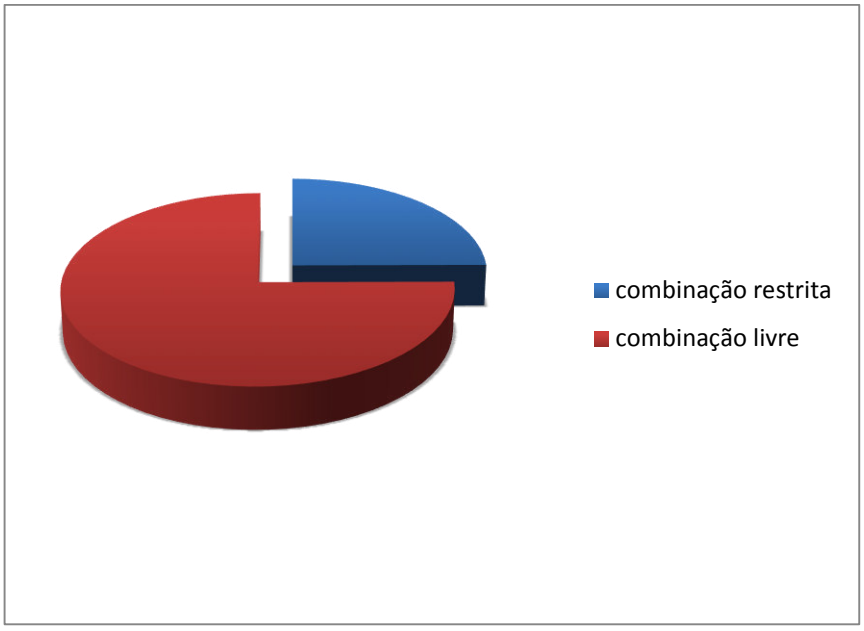

Figura 2- Distribuição das combinações (livres e restritas).

Categoria	Número de ocorrências	Percentagem
Combinação livre	19 428	75,08 %
Combinação restrita	6 447	24,92 %

Tabela 17- Distribuição relativa das combinações livres e das combinações restritas.

Confrontando as duas categorias das combinações, a categoria das combinações livres e das combinações restritas, pode-se concluir que a categoria das combinações restritas ficou muito aquém da categoria das combinações livres. Isto demonstra que apesar do algoritmo ter extraído muitas combinações a maior parte delas são combinações livres, confirmando o que dissemos sobre o dicionário utilizado na tradução e a necessidade de mudar de dicionário numa próxima investigação.

Depois desta análise pode-se concluir que, evidentemente, o algoritmo utilizado apresenta diversos problemas. Porém, tal como foi dito ao longo da análise, muitos destes problemas podem ser eliminados fazendo algumas alterações ao algoritmo.

Em primeiro lugar, será necessário retificar o algoritmo de forma a ele não cortar combinações lexicais a meio ou, então, aumentar o número de palavras selecionadas no contexto que rodeia a combinação lexical.

De seguida, seria necessário arranjar uma maneira de o algoritmo detetar diferentes combinações num mesmo segmento de texto, pois muitas combinações lexicais perderam-se devido a trocas feitas pelo algoritmo aquando da sua extração.

Porém, nem todos os erros encontrados nesta lista de pares candidatos a colocação foram culpa do algoritmo e é preciso ter em atenção que o dicionário utilizado para fazer a tradução do espanhol para o português apresentou muitas falhas também.

Seria então necessário mudar de dicionário, pois o dicionário utilizado é pouco rico em termos de sinónimos, visto tratar-se de uma ferramenta de tradução automática, em que a inexistência de ambiguidade é benéfica.

No entanto, pode-se concluir que melhorando e aperfeiçoando este algoritmo muitos serão os estudos sobre colocações que poderão ser feitos.

Conclusão

Este trabalho teve como principal objetivo a identificação e extração semiautomática de colocações usando métodos contrastivos.

Em primeiro lugar, através duma revisão da literatura especializada, apresentei as duas principais abordagens ao conceito de colocação, de forma a demonstrar as diferentes formas de entender o que é uma colocação. De seguida, fiz uma delimitação dos diferentes tipos de combinações lexicais, tendo em conta algumas características que lhes foram sendo atribuídas ao longo dos últimos anos. Porém, esta delimitação não é em nada fácil visto que as fronteiras entre as diferentes unidades fraseológicas são pouco claras.

Numa segunda parte desta dissertação, introduzi a questão da extração automática de colocações, explicando como esta é feita e apresentando alguns termos importantes dentro da extração automática. Depois de feita uma introdução a esta temática, foi introduzido o corpus, o algoritmo e o método de análise utilizados neste projeto de investigação.

Numa última parte, foi feita a análise detalhada dos resultados obtidos através da extração semiautomática de colocações, quer isto dizer, uma extração que em primeira instância foi feita de forma automática, mas que depois foi analisada manualmente.

Atendendo aos resultados obtidos, a primeira reação a estes resultados é de descontentamento, pois foram encontradas muitas combinações livres na lista de pares candidatos a colocações e, ainda, bastantes erros. O algoritmo acabou por encontrar poucas combinações restritas, especialmente colocações.

À medida que a lista de pares candidatos a colocações foi sendo analisada tornaram-se claros os problemas e as limitações desta extração semiautomática de colocações, em especial do algoritmo utilizado, apresentado no terceiro capítulo desta dissertação.

Em primeiro lugar, chego à conclusão que o dicionário utilizado para fazer a extração precisa de ser alterado em futuras experiências, visto que este trouxe muitos

problemas, desde o facto de não ter algumas entradas até ao facto de ter poucos sinónimos nas entradas existentes.

Desta forma, em investigações futuras, este problema pode ser resolvido utilizando outros dicionários ou até dicionários de tradução baseados em probabilidades, como é o caso dos dicionários apresentados em Simões & Almeida (2003) e Simões, Almeida & Carvalho (2013), de forma a aumentar o tamanho do dicionário e, ao mesmo tempo, a quantidade de traduções por palavra.

Em segundo lugar, o algoritmo utilizado nesta dissertação apresentou problemas de diversa natureza. Porém, muitos destes problemas poderão ser ultrapassados fazendo algumas alterações no algoritmo, o que pode aportar melhores resultados em novas experiências.

O primeiro aspeto que precisa de ser resolvido tem que ver com a grande quantidade de combinações truncadas. Para a sua resolução terá de se encontrar em primeiro lugar a razão real de tal acontecer: se é problema do algoritmo de extração ou de algum passo prévio na preparação do corpus. Se o problema for do algoritmo bastará aumentar o número de palavras no contexto que rodeia a combinação lexical. Quer dizer, em vez de o algoritmo extrair quatro palavras em redor da combinação, terá, por exemplo, de extrair cinco ou mais palavras.

O seguinte aspeto prende-se com o facto de o algoritmo não ter detetado algumas combinações lexicais por terem no mesmo segmento de texto uma outra combinação lexical semelhante. Isto fez com que o número de erros encontrados nesta análise fosse maior e se perdessem muitas combinações lexicais. De forma a resolver este problema será necessário que o algoritmo sofra algumas alterações para que seja capaz de distinguir duas combinações lexicais diferentes, que isto dizer, ter noção do número de vezes que cada palavra e suas possíveis traduções ocorrem no segmento, de modo a extrair uma relação correta. Também se poderá tirar partido de ferramentas de alinhamento à palavra para facilitar este trabalho.

Por último, convém referir que o facto de determinadas palavras não serem encontradas no dicionário por terem sido lematizadas de forma errada pode ser resolvido melhorando o algoritmo e utilizando um etiquetador de *part-of-speech*.

Pode-se concluir que, apesar deste projeto de investigação não ter tido os resultados esperados em termos de precisão, acabou por ser produtivo nas conclusões e ilações que se podem aplicar em futuras experiências de extração de colocações.

Bibliografia

Alonso Ramos, M. (1993). *Las funciones léxicas en el modelo lexicográfico de I. Mel'čuk,* Tesis doctoral, Madrid: UNED.

Alonso Ramos, M. (1994-1995). Hacia una definición del concepto de colocación: de J. R. Firth a I. A. Mel'čuk. *Revista de Lexicografía*, 1, 9-28.

Alonso Ramos, M. (2010). No importa se llamas o no colocación, descríbela. In Mellado, C. et al. (eds.), *Nuevas perspectivas de la fraseología del siglo XXI*. Berlin: Frank & Timme, 55-80.

Araújo, Sílvia, Almeida, J. J., Simões, Alberto & Dias, Idalete (2010). Apresentação do projecto Per-Fide: Paralelizando o português com seis outras línguas. *Linguamática*, 2, 71-74, Junho, 2010.

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. *German Society for Computational Linguistics and Language Technology (GSCL) Conference* (pp. 31-40).

Corbí-Bellot, et al. (2005). An open-source shallow-transfer machine translation engine for the romance languages of Spain. *Proceedings of the European Associtation for Machine Translation*, *10th Annual Conference*. Budapest: Hungary, 30-31.05.2005, 79-86.

Corpas Pastor, Gloria (2001). *En torno al concepto de colocación.* [em linha] *Euskera*, 46, Bilbao, Real Academia de la Lengua Vasca, 89-108. Consultado em 24 de Outubro de 2013 em http://www.euskaltzaindia.net/dok/euskera/11643.pdf

Coseriu, E. (1977). *Principios de semántica estructural.* Madrid: Gredos.

Cowie, A. P. (ed.) (1998). *Phraseology. Theory, Analysis and Applications.* Oxford: Clarendon Press.

Johnson, Ian, & Macphail, Alastair (2000). IATE–Inter-Agency Terminology Exchange: Development of a Single Central Terminology Database for the Institutions and Agencies of the European Union. *Workshop on Terminology resources and computation, LREC 2000 Conference*. Athènes: Grèce

Iriarte Sanromán, Álvaro (2001). *A Unidade Lexicográfica. Colocações, Frasemas, Pragmatemas.* Braga: Centro de Estudos Humanísticos – Universidade do Minho.

Koike, Kazumi (2001). *Colocaciones léxicas en el español actual: estudio formal y léxico-semántico.* Alcalá de Henares: Universidad de Alcalá/ Takushoku University.

Lewis, Michael (1993). *The Lexical Approach. The State of ELT and the Way forward.* Londres: Language Teaching Publications.

Mel'čuk, I., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire.* Louvain-la-Neuve: Duculot.

Mel'čuk, I. (1998). *Collocations and Lexical Functions*. Em Cowie (ed.) (1998), 23-53 [utilizei aqui uma fotocópia da versão dactilografada].

Mel'cuk, I. (2013). Tout ce que nous voulions savoir sur les phrasèmes, mais... *Cahiers de lexicologie*, 102, 2013, 129-149.

Moreno Jaén, M. (2009*). Recopilación, desarrollo pedagógico y evaluación de un banco de colocaciones frecuentes de la lengua inglesa a través de la linguística de corpus y computacional.* Granada: Editorial de la Universidad de Granada

Padró, Lluís & Stanilovsky, Evgeny (2012). FreeLing 3.0: Towards Wider Multilinguality**.** *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*. Istanbul: Turkey.

Simões, Alberto & Almeida, J. J. (2003). NATools -- a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31: 217-224.

Simões, Alberto & Carvalho, N. (2012). Desenvolvimento de aplicações em Perl com FreeLing 3. *Linguamática*, 4(2): 87-92.

Simões, Alberto, Almeida, J.J., & Carvalho, Nuno (2013). Defining a probabilistic translation dictionaries algebra. In Luís Correia, Luís Paulo Reis, José Cascalho, Luís Gomes, Hélia Guerra, and Pedro Cardoso, editors, *XVI Portuguese Conference on Artificial Inteligence – EPIA* (pp. 444-455). Angra do Heroísmo.

Suárez, Octavio, Aguiar, José, Berriel, Isabel & Rodríguez, Virginia (2011). Extracción automática de colocaciones terminológicas en un corpus extenso de lengua general. *Procesamiento del Lenguaje Natural*, 47: 145- 152.