**MASARYKOVA UNIVERZITA**
**FILOZOFICKÁ FAKULTA**

**Ústav románských jazyků a literatur**
Portugalský jazyk a literatura

**Alena Vašíčková**

# COLOCAÇÕES

Magisterská diplomová práce

Vedoucí práce: Mgr. Iva Svobodová, Ph.D.

**2009**

*Prohlašuji, že jsem diplomovou práci vypracovala samostatně a pouze s využitím uvedených pramenů a literatury.*

*Prohlašuji, že tištěná verze práce je totožná s verzí elektronickou.*

……..

Zde bych chtěla upřímně poděkovat vedoucí práce Mgr. Ivě Svobodové, Ph.D. za podnětné připomínky a energii a čas, který mé práci věnovala.
Děkuji i PhDr. Jarmile Fictumové z Katedry anglistiky a amerikanistiky FF MU za poskytnutí cenných materiálů.
Zároveň děkuji RNDr. Ondřeji Bojarovi, Ph.D. z Ústavu formální a matematické lingvistiky, bez jehož pomoci by nevznikl paralelní česko-portugalský korpus a nástroje pro jeho další rozšiřování.
Dík patří i Ing. Pavlu Erlebachovi za trpělivost s mými dotazy ohledně PHP+MySQL a za užitečné návrhy pro psaní skriptů.

# Índice

1. Introdução ..... 5
2. A importância das colocações ..... 8
2.1 Colocações na aquisição da linguagem e no processamento psicológico ..... 8
2.2 Colocações no ensino ..... 11
2.3 Colocações na tradução ..... 14
2.4 Colocações no processamento computacional e na linguística computacional ..... 14
3. Colocações na teoria linguística ..... 17
3.1 Colocações na tradição linguística anglo-saxónica ..... 17
3.1.1. Princípios do termo 'colocação' em linguística ..... 17
3.1.2 J. R. Firth e a escola firthiana ..... 19
3.1.3 Colocações na era de linguística de corpus ..... 24
3.1.4. Abordagem frasal ..... 27
3.2 Co-ocorrência lexical restrita na tradição linguística portuguesa ..... 31
3.2.1 Álvaro Iriarte Sanromán ..... 33
3.2.2 Herculano de Carvalho ..... 42
3.2.3 Bernard Pottier ..... 43
3.3 Colocações na tradição linguística checa ..... 44
3.4 Síntese das teorias e princípios da selecção de colocações para o dicionário de colocações checo-português ..... 51
4. Dicionários de colocações ..... 54
4.1 O problema de metalexicografia ..... 54
4.2. Dicionários existentes de colocações e de combinações não livres ..... 55
5. Criação de um dicionário de colocações checo-português ..... 60
5.1 Métodos de detecção e extracção de colocações existentes ..... 60
5.1.1 Colocações monolingues ..... 60
5.1.2 Pares bilingues ..... 61
5.2 Métodos utilizados na criação do dicionário de colocações checo-português ..... 64
5.3 Versão electrónica do dicionário de colocações checo-português ..... 66
5.4 Trabalho futuro ..... 67
6. Conclusão ..... 68
Bibliografia ..... 70
Apêndice 1: Lista das colocações ..... 79
Apêndice 2: Código fonte da página web do dicionário de colocações online ..... 85

*We talk of high mountains and tall trees, but not usually of tall mountains and high trees. Similarly a man can be tall but never high (except in the sense of being intoxicated!), whereas a ceiling can only be high, not tall. A window can be both tall or high, but a tall window is not the same as a high window. We get old and tired, but we go bald or grey. We get sick but we fall ill. A big house, a large house and a great house have the same meaning, but a great man is not the same as a big man or a large man. You can make a big mistake or a great mistake, but you cannot make a large mistake. You can be a little sad but not a little happy. We say very pleased and very tiny, but we do not say very delighted or very huge.*[1]

*In order to speak natural English, you need to be familiar with collocations. You need to know, for example, that you say 'a heavy smoker' because heavy (NOT big) collocates with smoker, and that you say 'free of charge' because free of collocates with charge (NOT cost, payment, etc.). If you not choose the right collocation, you will probably be understood but you will not sound natural.*[2]

*La combinación *introducir ganas no existe en español, pero sí entrar ganas; lo mismo ocurre con *albergar un chasco. En español, albergamos esperanzas e incluso odios, pero no chascos. Estos, los tenemos o nos los llevamos.*[3]

*…o esp. pan blanco se opõe ao pan negro, que não é negro, e água [sic] salada a água [sic] dulce, que é simplesmente não-salgada. Trata-se sempre de oposições na norma, que caracterizam os idiomas a que pertencem; assim, o esp. vino tinto é vermelho em italiano (vino rosso) e preto em servo-croata (crno vino).*[4]

[1] http://esl.fis.edu/grammar/easy/colloc.htm
[2] Longman Dictionary of Contemporary English (1987: 193).
[3] http://www.dicesp.com
[4] Coseriu, 1979: 68, em Sanromán Iriarte (2002: 190).

# 1. Introdução

Todos os dias, usamos expressões aparentemente ilógicas: *vinho branco* é dourado e *pão preto* é castanho. Um falante nativo sabe que deve dizer *vinho tinto* e *primeiros socorros* mas um estrangeiro que está a aprender a língua talvez não ache estranho *vinho vermelho* nem *primeira ajuda*. Em checo «damos» atenção, em inglês «pagamos» atenção e em português prestamo-la. Todas essas peculiaridades devem-se ao fenómeno da colocação e as expressões chamam-se colocações (do latim *collocare*, colocar junto).

Analisando a nossa experiência pessoal com a aprendizagem e o ensino de línguas e com a tradução, chegámos à conclusão de que a problemática das colocações é extremamente importante; porém, não lhe é prestada devida atenção. Este trabalho tem como objectivo mudar esta situação desfavorável e demonstrar a importância das colocações.

Quem decidir dedicar-se ao estudo das colocações, notará logo que – tal como disse Nesselhauf (2004, em Antunes, 2007) – há tantas respostas à pergunta 'o que é uma colocação?' quantos os autores que escrevem sobre o assunto. Segundo Bartsch (2004: 27-28), esta situação deve-se aos seguintes factos:

(i) a dificuldade encontrada em delimitar os diferentes tipos de combinações de palavras;

(ii) o facto de as combinatórias consistirem num fenómeno linguístico que parece situar-se na fronteira entre a gramática e o léxico (o que dificulta a elaboração de uma definição formal e sistemática do ponto de vista da teoria linguística, uma vez que parece não ser possível uma explicação baseada unicamente em princípios sintácticos e semânticos já estabelecidos);

(iii) o facto de existirem diferentes abordagens deste fenómeno, nomeadamente a abordagem baseada em frequências e a abordagem fraseológica.[5]

[5] Traduzido por Sandra Antunes (2007).

Bartsch acrescenta que a insuficiente especificidade das definições do termo 'colocação' domina nos estudos sobre as colocações desde a introdução do termo em linguística. Este é ainda hoje em dia um problema com o qual a linguística não consegue lidar efectivamente.

Na literatura consultada encontrámos o termo e o conceito de 'colocação' utilizado em vários sentidos diferentes:

(i) denominação do facto de algumas palavras ocorrerem juntas formando locuções;
(ii) o facto de algumas palavras poderem ocorrer juntas, i. e. a capacidade de algumas palavras ocorrerem juntas (neste sentido é também utilizado o termo 'colocabilidade');
(iii) combinações frequentes ou usuais de palavras;
(iv) combinações de palavras aparentemente livres, onde actua qualquer tipo de restrição lexical determinada pela norma e uso.

Apesar destas dificuldades, tentaremos esboçar o desenvolvimento no campo dos estudos sobre as colocações, traçando o seu uso desde as primeiras menções até ao dia de hoje. Devido ao facto de o campo ter sido liderado pelos autores anglo-saxónicos, abordaremos com teorias e conceitos nascidos na Grã-Bretanha, mencionando, ocasionalmente, também autores não anglo-saxónicos que escrevem em inglês. A seguir, pesquisaremos as obras de linguistas portugueses e checos com a finalidade de descobrir se eles trazem abordagens novas e originais ou apenas adoptam os conceitos da linguística anglo-saxónica. Como durante a nossa pesquisa não encontrámos evidência de que os autores que escrevem sobre colocações se agrupassem ou formassem escolas complexas linguísticas que merecessem registo[6], vamos tratar dos autores mais importantes separadamente.

O conhecimento adquirido servir-nos-á como a base teórica para a criação de um futuro dicionário de colocações checo-português. O segundo objectivo do presente trabalho

[6] Há seguidores de alguns autores que formam algo que pode ser considerado uma escola linguística, esta contudo baseia-se em abordagens semelhantes da problemática de colocações. De nenhuma maneira podemos falar duma corrente linguística complexa. Ao contrário, podemos dizer que alguns autores pertencem ao estruturalismo, à pragmática, etc. ou pelo menos foram influenciados por eles.

é então explorar os métodos de criação de tal dicionário. Os resultados, i. e. amostras de colocações que poderão ser incluídas no futuro dicionário, serão publicadas em anexo deste trabalho e na Internet.

Demonstraremos também a importância das colocações na aquisição e processamento de linguagem, ensino de línguas, tradução, linguística computacional e lexicografia.

# 2. A importância das colocações

## 2.1 Colocações na aquisição da linguagem e no processamento psicológico

Os estudos recentes confirmam que devido ao facto de as frases lexicais ('lexical phrases', termo introduzido por Nattinger) prefabricadas formarem uma parte da linguagem dum adulto tão ubíqua, podemos assumir que também fazem parte da linguagem duma criança, constituindo uma parte do sistema da linguagem que esta está a adquirir (cf. Nattinger-DeCarrico, 1992: 24).

Esses estudos mostram que as crianças passam por uma frase em que usam um grande número de sequências não analisadas de língua ('unanalyzed chunks of language') [7] em certos contextos sociais previsíveis. Por outras palavras, elas usam frequentemente uma linguagem «prefabricada» em situações apropriadas. Nattinger-DeCarrico dão o seguinte exemplo: Quanto a criança pergunta a questão frequente *what is that?*, usa-a como se os três morfemas fossem uma unidade única e não segmentada, *what-is-that?*, como qualquer palavra do seu vocabulário. Tratam estas três palavras como se fossem uma unidade não analisada, frequentemente reduzindo muitos dos seus sons ([hwəsdæt], [hwədæ], etc.) e pronunciando o conjunto sob apenas um acento; aparentemente não reconhecem, pelo menos nas fases iniciais da aquisição, que se trata duma frase com componentes lexicais separados.

Muitas investigações iniciais atribuíram a frequência dessas sequências à relevância da imitação e à necessidade da memorização na aprendizagem da linguagem. Viam-nos como um «atoleiro» na aquisição das regras regulares e sintácticas. Consideravam essas sequências prefabricadas distintas (e de certa maneira periféricas) do corpo fundamental da linguagem (Ibidem).

Naturalmente, há também estudos sobre a aquisição da L2, a segunda língua, ou seja, a língua não materna. Hakuta (1974, em Nattinger-DeCarrico, 2002: 24-25) foi dos primeiros a sugerir que afinal essas sequências talvez não fossem tão incidentais. Num

[7] Como a maioria dos estudos sobre as colocações está escrita em inglês, enfrentámos numerosos problemas terminológicos. Sempre procurámos encontrar um equivalente português, mas às vezes era necessário incluir a nossa tradução.

estudo de crianças japonesas que aprendiam inglês fez uma distinção importante entre as rotinas ('routines') prefabricadas que descreveu como sequências da língua invariáveis ('unvarying chunks of language') e os padrões ('patterns') prefabricados que descreveu como segmentos de frase que operam em conjunto com um componente movível, como inserção duma frase nominal ou duma frase verbal, como por exemplo *this-is-a X*. Ele também sugeriu que essas sequências têm um papel importante no processo da formação de regras – as crianças não os utilizam apenas como fórmulas memorizadas mas sim como matéria-prima da segmentação e análise no desenvolvimento das regras da sintaxe.

As observações de Hakuta foram confirmadas por Lily Wong-Fillmore (1976, em Nattinger-DeCarrico, 2002), cuja pesquisa é reconhecida como uma das mais complexas no campo da linguagem prefabricada na aquisição da L2. Ela até põe essas sequências no próprio centro da aquisição duma língua:

> The strategy of acquiring formulaic speech is central to the learning of language. [...] Routines and patterns learnt in the language acquisition process evolve directly into creative language.

Eis a palavra-chave, 'formulaica', ou seja, etiqueta que é atribuída a expressões que não são formadas segundo as regras regulares e abrange também idiomas e colocações.

Por outro lado, há quem ponha em dúvida as conclusões de Wong-Fillmore: segundo Krashen e Scarcella, rotinas e padrões apenas têm um papel marginal na aquisição da linguagem e são completamente diferentes do processo da construção criativa considerada somente uma questão da análise das regras sintácticas. Sentem que essas sequências de língua são «useful in establishing and maintaining relations, [but] do not serve a primary role in language acquisition» (Krashen-Scarcella, 1978: 283-300, em Nattinger-DeCarrico, 1992: 26).

Contudo, a maioria dos investigadores concorda que os aprendentes usam no processo da aquisição da linguagem um número significante das expressões prefabricadas (Nattinger-DeCarrico, 1992: 26). Onde as opiniões divergem, é na questão da importância destes. O conhecimento linguístico dum adulto parece conter um *continuum* de construções linguísticas de vários níveis da complexidade e abstracção abrangendo itens concretos e

particulares, classes mais abstractas de itens ou combinações complexas de sequências de língua concretas e abstractas[8].

Contudo, é necessário ter em conta que apenas a linguagem dum adulto contém esta mistura e tem esta natureza. Esta estrutura emerge passo a passo, portanto as crianças não aprendem a L1 como se tivessem em mente uma gramática inata[9]. Mais propriamente, aprendem com a ajuda de unidades psicolinguísticas diferentes das que os adultos usam, adquirem linguagem na base da frequência, recorrência e imitação[10].

Antes de falarmos sobre a resultante importância das colocações (e outras formas da linguagem *formulaica*) no ensino, mencionemos brevemente o papel da linguagem prefabricada no processamento psicológico, estreitamente relacionado com a aquisição da linguagem. Há sugestões de que a capacidade da memória é vasta – contrariamente à velocidade do processamento. Por isso temos que aprender «atalhos» com o propósito de tornar o tempo de processamento o mais eficiente possível (cf. Nattinger-DeCarrico, 1992: 31). Vários estudos do processamento da linguagem mostram que a linguagem está depositada redundantemente. As palavras, por exemplo, estão depositadas não apenas como morfemas individuas senão como partes de frases, ou como sequências de língua longas e memorizadas ('longer memorized chunks of speech') e estas são frequentemente recuperadas da memória numa forma dessas «pré-juntadas» sequências ('pre-assembled chunks') (Ibidem).

Como afirmam Nattinger e DeCarrico (Ibid.: 32), «é a nossa capacidade de usar frases lexicais que nos ajuda a falar fluentemente.»[11]. Uma opinião semelhante – relevante às colocações – pode ser encontrada nos livros de aprendizagem de colocações: destaca-se o facto de que a fluência que resulta do uso correcto das colocações deveria ser um dos objectivos do ensino de línguas.

[8] [Adult language knowledge consists of] concrete and particular items (as in words and idioms), more abstract classes of items (as in word classes and abstract constructions), or complex combinations of concrete and abstract pieces of language (such as mixed constructions) (Tomasello, 2000: 99, em Collocations and Idioms 1: 87).
[9] Children do not learn their L1 on the basis of an innate Universal Grammar (Eskildsen-Cadierno, 2007: 87).
[10] Children learn language in an item-based fashion heavily reliant on frequency, recurrence and imitation (Ibidem).
[11] It is our ability to use lexical phrases [...] that helps us speak with fluency.

## 2.2 Colocações no ensino

Em vista do que temos dito sobre a aquisição e o processamento de linguagem, vejamos agora o problema das colocações no ensino da L2. Muitos estudos sobre esta matéria têm sido efectuados, conduzindo, em alguns casos, a resultados parciais diferentes; não obstante, há unanimidade sobre as seguintes conclusões:

(i) a produção das colocações representa um problema para os aprendentes, e estes problemas são maiores do que os que os aprendentes têm com o uso de vocabulário geral;

(ii) os aprendentes usam menos colocações do que os falantes nativos;

(iii) os aprendentes não têm ciência nem das restrições colocacionais, nem do potencial combinatório dos itens lexicais (Eskildsen-Cadierno, 2007: 87-89).

Segundo as estatísticas, o uso das colocações pelos aprendentes é apenas 25 %[12] (Nesselhauf, 2005, em Eskildsen-Cadierno, 2007: 89). Uma das possíveis interpretações desta indicação é que os aprendentes da L2 adquirem a competência colocacional com dificuldades. Esta interpretação é apoiada pela vista tradicional de a linguagem *formulaica* ser divergente das normas duma língua gerada gramaticalmente.

> This is a result of the underlying view of FL [formulaic language] being deviant from the norm of gramatically generated language. It is even implied in Wray[13] that "learning formulaic language is not 'real' language learning" (Ibid.: 89).

A aprendizagem «real» pressupõe computação baseada nas regras sintácticas e a capacidade de analisar e de compor unidades lexicográficas. A linguagem *formulaica*, porém, foge das regras da combinatória gramática; há inconsistências entre elas. Por outras palavras, a linguagem *formulaica* é considerada *fomulaica*, ou seja fixa ou «congelada»[14]porque a competência linguística nunca podia ter produzido tais combinações (Ibidem).

[12] Contudo, os autores não especificam o que exactamente o número significa.
[13] Wray, Alison (2002). *Formulaic language and the lexicon*. New York: Cambridge University Press. p196.
[14] Os autores anglo-saxónicos utilizam o termo *frozen*. Pereira e Nascimento (1996) empregam o termo «cristalizada».

Daqui resulta a necessidade de aprender colocações (bem como outras formas irregulares e «congeladas») duma maneira diferente da aprendizagem de fenómenos regulares. Infelizmente para os aprendentes, a aprendizagem de colocações consiste em memorização automática, pois a forma da colocação na língua estrangeira não pode ser deduzida. Ás vezes a estrutura é semelhante (*kvalifikovaná většina – maioria qualificada*) mas é necessário verificar o uso exacto (*ozonová díra – buraco DO ozono* vs. *ozonová vrstva – camada DE ozono*).

O estudo de materiais de aprendizagem de várias línguas levou-nos à conclusão de que a única língua em cujo ensino é dada devida atenção às colocações é o inglês. Este facto provavelmente resulta de um durável interesse dos linguistas anglo-saxónicos pelas colocações e também do facto de o inglês ser a língua mundial mais importante, a cujo ensino é prestada muita atenção, sendo os métodos já bem elaborados. Existe uma série de livros dedicados exclusivamente às colocações (*English Collocations in Use*) mas as colocações não fazem falta praticamente em nenhum bom livro de aprendizagem de inglês. A situação no ensino de outras línguas é diferente: não existem livros especializados em colocações e há poucos materiais que tratem delas (ou os poucos existentes não as nomeiam explicitamente). A importância das colocações é subvalorizada ou a problemática é omitida completamente, apesar de as colocações serem imprescindíveis no ensino de estudantes avançados: sabendo já como se dizem palavras separadas, devem aprender juntá-las duma forma natural como o fazem os falantes nativos. Para além disso, «precise thought is made easier by well-defined terms» (Lewis, 1993: 1).

Begoña Sanromán Vilas e Margarita Alonso Ramos (2007) observaram que até os estudantes mais avançados faziam muitos erros no uso das colocações em espanhol e afirmam que um dicionário de colocações é um instrumento necessário de aprendizagem. A ideia de combinar um dicionário de colocações com exercícios não é nova: alguns exercícios são inseridos no *Oxford Collocations Dictionary for Students of English*. Também o DiCE (*Diccionario de colocaciones del español*) inclui tanto os exercícios de produção como os de compreensão.

Mencionemos alguns dos exemplos apresentados[15]:

[15] http://www.dicesp.com

A. Rodea la respuesta que mejor se corresponda con la glosa ‘causar pánico en varias personas’:

a. esparciar el pánico;

b. reproducir el pánico;

c. sembrar el pánico;

d. extender el pánico.

B. Transforma las siguientes oraciones de manera que se pueda sustituir el verbo en negrita por una expresión equivalente: un verbo de apoyo seguido del nombre que aparece entre paréntesis:

a. **Firme** este documento, por favor. (FIRMA)

b. **Castigarán** a los culpables. (CASTIGO)

c. No se preocupe: su vida no peligra. (PELIGRO)

C. Encuentra sinónimos para los verbos que aparece en cursiva:

a. *Sentí* una gran ALEGRÍA, que todavía ahora no podría explicar.

b. *Sufrió* una grave DOLENCIA que le mantuvo retirado durante un tiempo.

c. Le *dio* un PUNTAPÉ a un objeto duro que rodó unos pasos delante suyo.

D. Rellena los huecos valiéndote de las glosas entre paréntesis:

a. Pedro le _____ MIEDO a los aviones.

b. No deberías _____le RENCOR por tan cosa.

c. De repente, le _____ el RESPETO. (Ibid.: 289-290)

Quanto ao segundo grupo – exercícios de compreensão – é possível produzir exercícios como o seguinte:

E. Si leemos en el periódico “el gobierno de EEUU tiene fuertes sospechas de que Irán ha desarrollado programas de armas bacteriológicas”, entenderemos que las **sospechas** son FUERTES porque...

a. se refieren a un hecho considerado muy malo;

b. el Gobierno está muy seguro de el hecho es cierto;

c. duran mucho tiempo;

d. son compartidas por varias personas. (Ibid.: 292)

A resposta correcta é b, pois um suspeito dum facto considerado negativo seria *grave* ou *terrible*, um suspeito presente por muito tempo seria *permanente* e um suspeito comúm a muitas pessoas seria *extendido* ou *generalizado*.

Contudo, apesar de uma inegável importância da aprendizagem das colocações, achamos que não é importante aprender milhares de colocações de cor. O que importa são os métodos: os estudantes avançados (entre os quais há futuros tradutores) deveriam conhecer as ferramentas que facilitam a procura das colocações numa língua estrangeira e a transferência das colocações duma língua para outra (algumas ferramentas úteis serão mencionadas em 5.1). Assim estamos a chegar à problemática das colocações na tradução.

## 2.3 Colocações na tradução

Como já mencionámos, as colocações causam problemas sobretudo na hora de produzir um texto numa língua estrangeira, seja em forma de capturar nossas próprias ideias seja em forma de traduzir ideias e textos de outros. Temos problemas em traduzir correctamente colocações de tipos B-D na classificação de Mel'čuk/Iriarte Sanromán (*prudká reakce – reacção *afiada vs. violenta; drtivá většina – *maioria plausível vs. esmagadora maioria; single words – palavras (*)únicas vs. (*)individuais vs. (*)separadas vs. isoladas*) bem como colocações-unidades lexicais (*daňový poplatník – contribuinte, slovní hříčka – trocadilho, zdrojový text – texto fonte*) e outros tipos da linguagem prefabricada.

Naturalmente, um ser humano nunca poderá aprender todas as colocações duma língua estrangeira (e os tradutores confirmarão que às vezes, influenciados pelo texto fonte, até duvidam das colocações em língua materna) mas, como já mencionámos no capítulo anterior, é importante que aprenda como encontrar as colocações em textos disponíveis; no caso de textos técnicos, as colocações frequentemente são idênticas à terminologia.

## 2.4 Colocações no processamento computacional e na linguística computacional

Sob a influência das gramáticas de Chomsky, o léxico computacional tinha sido visto como uma lista de palavras que possui propriedades sintácticas e semânticas muito específicas e

que é sujeite às regras combinatórias dum *parser*/*generator*. Contudo, o léxico passou a ser considerado um conjunto de vários tipos do conhecimento linguístico, não apenas de propriedades de palavras isoladas (Nattinger-DeCarrico, 1992: 22).

Becker (1974, em Nattinger-DeCarrico, 1992: 22) foi dos primeiros a propor que frases como *let alone, as well as* e *so much for* que de muitas maneiras fogem das análises nas gramáticas tradicionais simplesmente não podem ser ignoradas: são ubíquas. Becker pretende tratar essa classe de frases idiossincráticas duma forma sistemática. Como notam Nattinger e DeCarrico, a opinião prevalente entre os linguistas computacionais nos anos noventa (quando foi publicado o seu livro *Lexical Phrases and Language Teaching*) era a de que o conhecimento linguístico não pode ser dividido rigorosamente entre regras gramaticais e itens lexicais. Mais provavelmente, há uma escala inteira de itens, dos quais alguns são específicos e pertencem a um número pequeno de casos, enquanto os outros são muito gerais e pertencem a um grande número de casos. Os primeiros chamam-se frequentemente 'itens lexicais', enquanto os segundos 'regras gramáticas', mas dado que os elementos existem a todos os níveis da generalidade, é impossível distinguir rigorosamente entres eles.

Em 1984, Wilensky *et al.* (1984, 574-593, em Nattinger-DeCarrico, 1992: 23) propôs a abordagem frasal ('phrasal approach') que incorporou frases inteiras bem como palavras separadas. Em 1987, Zernick and Dyer (1987: 308-327, em Nattinger-DeCarrico, 1992: 23) aprofundaram a teoria incluindo, no seu conceito de léxico, as seguintes unidades:

(i) palavras isoladas ('single words');
(ii) frases fixas ('fixed phrases');
(iii) todos os casos de certas frases variáveis ('all instances of certain variable phrases');
(iv) itens isolados e generalizados que abrangem esses séries de frases ('single generalized entries which encompass these sets of phrases').

Quanto à aplicação prática, as colocações podem ser de grande importância nas seguintes áreas:

**Criação de dicionários,** seja monolingues, seja bilingues. As colocações frequentemente formam um significado semântico novo e a extracção automática facilitará a criação de dicionários gerais, dicionários de termos, etc.

**Tradução automática.** As primeiras máquinas de tradução utilizavam o mecanismo de tradução à letra, perdendo assim uma parte do significado. As colocações reconhecidas podem ser traduzidas correctamente, mantendo o sentido do texto fonte.

**Verificadores ortográficos**. Durante o controlo, não apenas as palavras mas também as combinações de palavras podem ser verificadas. Assim descobre-se uso errado de palavras (**vinho vermelho*).

**OCR, reconhecimento de texto.** Geralmente trata-se das situações de predição de palavras num texto, dependendo das palavras antecedentes. Uma palavra dificilmente reconhecível será reconhecida com maior sucesso se as palavras antecedentes (ou o contexto em geral) predicarem ocorrência duma colocação.

**Motores de pesquisa, classificação de documentos e sistemas de indexação.** As palavras adquirem novos sentidos se formam parte duma colocação e conhecendo o «novo» significado específico será mais fácil e mais preciso determinar o conteúdo do documento. Se formos capazes de extrair as colocações dum documento, seremos capazes de utilizá-las (junto com os termos univerbais) como a base para a indexação. Se reconhecermos as colocações também na consulta de pesquisa, obteremos resultados melhores dos que obtemos através da pesquisa de palavras isoladas e indexação clássica, i. e. univerbal. Com volumes de dados tão rapidamente crescentes é cada vez mais importante a especificação dos parâmetros e a capacidade de interpretar correctamente as expressões pluriverbais. Igualmente, se utilizarmos colocações e não palavras únicas na **procura das palavras-chave** podemos esperar melhores resultados. Ao utilizarmos um dicionário de colocações na **síntese de texto**, podemos evitar uso errado dalgumas palavras.

**Recuperação de informações.** Durante o processo da desambiguação semântica distinguimos qual o significado (de todos os que a palavra tem) no texto concreto. Em muitos casos o significado depende do contexto e do conhecimento de possíveis colocações de uma determinada palavra; é porque as colocações são frequentemente formadas por palavras polissémicas e o sentido apenas é determinado através da colocação.

# 3. Colocações na teoria linguística

## 3.1 Colocações na tradição linguística anglo-saxónica

### 3.1.1. Princípios do termo 'colocação' em linguística

Ao consultar a segunda edição do conspícuo OED, *Oxford English Dictionary*, vemos que o primeiro registo do termo 'collocation', num contexto distintamente linguístico, se data já do século XVIII:

> **1750** HARRIS *Hermes* II. iv. Wks. (1841) 197 The accusative..in modern languages..being subsequent to its verb, in the collocation of the words.

Na citação de Harris, o termo colocação é utilizado num sentido mais próximo ao que hoje em dia chamamos coligação, i. e. a justaposição gramática das palavras numa frase (trataremos o termo em detalhe em 3.1.2). Não há nenhuma menção sobre o carácter lexical que o termo assumiu ao decorrer do tempo e com o qual está ligado na terminologia da linguística contemporânea.

Numa citação de Trager de 1940 (na mesma entrada no OED), o termo 'collocation' denota as propriedades gerais combinatórias dos 'elementos linguísticos' (i. e. não apenas itens lexicais).

> **1940** G. L. TRAGER in *Language* XVI. 301 Collocation establishes categories by stating the elements with which the element being studied enters into possible combinations. *Ibid.* 303 It is now necessary to establish the collocations of the various forms to see what their functions are.

Esta apreensão concebe a colocação como um conjunto de propriedades combinatórias sintáticas e semânticas que regem a combinação de itens lexicais individuais e de formas individuais das palavras.

Não é evidente quem foi o primeiro a introduzir o termo 'collocation' no sentido de uma combinação de palavras fixa e relativamente recorrente. Já em 1917, Otto Jespersen

(em Bartsch, 2004: detalhes bibliográficos não mencionados) nota que algumas expressões colocam frequentemente com outras:

> *Little* and *few* are also incomplete negatives: note the frequent collocation with *no*: *there is little or no* danger.

Na obra *A Grammar of English Words* (1938: *x,* em Bartsch, 2004), Harold E. Palmer, além de ter incluído o termo colocação no subtítulo do capítulo *How the vocabulary is set out* explica o que é uma colocação e como ele resolveu o problema das colocações na sua 'gramática das palavras' ('grammar of words'). As colocações são concebidas como unidades lexicais.

COLLOCATIONS

> When a word forms an important element of a "collocation" (a succession of two or more words that may best be learnt as if it were a single word) the collocation is shown in bold type (...). The collocations are entered so far as possible under the appropriate semantic variety of the word (...).

O que é interessante sobre esta definição é o facto de que Palmer a aprofunda mais do que muitos autores posteriores, afirmando que, em princípio, não há restrições quanto ao número das palavras («a succession of two or more words»). Nos exemplos apresentados vemos que Palmer considera como constituintes de colocações palavras semanticamente autónomas bem como as não autônomas (*make up, make a fool of, ask about, ask for, a good many*)[16]. Realce-se, ao mesmo tempo, a importância da distinção entre 'colocações' e 'frases' ('phrases') que, de novo, reflecte o conceito da colocação como uma unidade lexical:

PHRASES

> Phrases are different from collocations. While collocations are comparable in meaning and function to ordinary single "words", (and indeed are often translated by single words in the

[16] Alguns autores (e.g. Hausmann, 1995) reconhecem como constituintes apenas palavras semanticamente autônomas..

> student's mother-tongue), phrases are more in the nature of conversational formulas, sayings, proverbs, etc (Ibid.: *xi*).

Esta distinção entre as colocações e as frases, i. e. provérbios e frases feitas, é uma das que se mantiveram até ao dia de hoje. Contudo, o que Palmer rejeita é a distinção entre os idiomas e as colocações, considerando as colocações apenas uma subcategoria dos idiomas.

> The term 'idiom' is as superfluous as (...) what are usually called idioms are generally nothing other than (a) collocations, (b) phrases and sayings, (c) rarer semantic varieties of words and collocations, (d) peculiar construction patterns and, in short, any word or form of wording that is likely to puzzle the foreign student (Ibid.: *xii*).

### 3.1.2 J. R. Firth e a escola firthiana

J. R. Firth (1890-1960), o primeiro professor de Linguística em Inglaterra, foi um dos iniciadores da chamada *London School of Linguistics* que seguia a linha estruturalista prevalente naquela época. Além disso, copiou algumas ideias e conceitos também do funcionalismo. Sob a forte influência do antropólogo Bronislav Malinowski[17], Firth assumiu a ideia de que a língua devia ser estudada como um fenómeno social, sendo necessário levar em consideração, além dos factos puramente linguísticos, também o contexto social**.** Firth foi o primeiro a tratar das colocações mais amplamente e contribuiu para o aumento do interesse por este fenómeno[18]. Firth baseou a sua teoria do significado[19] em 'sentido através da colocação' ('meaning by collocation'), ou seja, na pressuposição de que o significado duma palavra é determinado pelas relações existentes entre ela e as outras palavras.

> I propose to bring forward as a technical term, meaning by 'collocation', and to apply the test of 'collocability' (1957: 194).

[17] Bronislaw Kasper Malinowski (1882-1942), antropólogo britânico, considerado fundador da escola 'funcional' de antropologia que propunha que instituições humanas devessem ser examinadas no contexto da cultura como um todo.
[18] Contudo, Firth pretende pesquisar o fenómeno de colocações apenas com fins estilísticos.
[19] theory of meaning

Na teoria do significado contextual[20], a colocação tem um papel central no sentido determinado pelo contexto. No ensaio *The Technique of Semantics* de 1934, Firth afirma que as palavras adquirem o seu sentido tanto através de co-ocorrências típicas com outras palavras num ambiente linguístico imediato como através da sua posição nos enunciados que pertencem à linguagem general ou a uma particular variedade domínio-específica[21]. Firth defende que o significado tem quatro níveis:

(i) contexto situacional;
(ii) colocações;
(iii) sintaxe;
(iv) fonologia e fonética.

A primazia atribuída às relações sintagmáticas entre itens linguísticos e a resultante importância da descoberta dos limites sintagmáticos para além dos quais a escolha linguística é impredicável levou Firth a marcar distinção entre os conceitos de 'colocação' e de 'coligação'. Enquanto o conceito de colocação é usado para denotar a escolha lexical restringida (as colocações correspondem a co-ocorrências significativas de determinadas palavras com uma certa proximidade), a coligação refere-se à formação regular de estruturas gramáticas como preferência dalguns verbos em certas estruturas; as coligações correspondem a combinações de palavras sintacticamente restritas do ponto de vista da sub-categorização (*afraid of*).

Firth não nos dá outras referências sobre a distância das palavras que constituem uma colocação, constatando apenas que como colocações podem ser consideradas quer associações contíguas de palavras (*hold life in contempt*), quer palavras que não ocorrem contiguamente (*one dark and cold night; the night is dark*).

A contribuição de Firth é digna de elogio dado que não teve à sua disposição nenhumas ferramentas de processamento automático do texto. Mesmo assim, efectuou uma pesquisa empírica de itens lexicais em contexto baseada em textos autênticos e não em exemplos inventados e sem contexto. Apesar de não haver, na sua época, ferramentas

[20] contextual meaning theory
[21] through their being embedded in utterances identified as belonging to either the general language or to a particular domain-specific language variant (Ibid.: 10)

adequadas de processamento do texto, Firth sugeriu uma metodologia ideal do estudo da linguagem que pôde ser aplicada quando finalmente apareceram ferramentas apropriadas; facilitou assim a subida da linguística de corpus. A «ressureição» do interesse pelas colocações decorreu nos anos 80 graças aos *corpora* já elaborados[22].

Contudo, houve autores que escreveram sobre as colocações nos anos 60 e 70 e algumas notas sobre os seguidores de Firth – T. F. Mitchell, Michael Halliday e Sidney Greenbaum – merecem ser listadas aqui.

**T. F. Mitchell** (1919-2007) começou a sua carreira linguística em Londres: depois da segunda guerra mundial foi nomeado director da *School of Oriental and African Studies* por J. R. Firth. Junto com Sydney Greenbaum é considerado uma das personagens centrais da atitude britânica na área de estudos lexicográficos.

Como Mitchell nota no seu estudo *Principles of Firthian linguistics*, a linguística firthiana[23] começou a dar prevalência ao estudo das relações sintagmáticas entre itens linguísticos. Para além disso, a linguística firthiana (com a excepção de Halliday) reconheceu também a interdependência da gramática e do léxico. Mitchell próprio descreve esta interdependência em termos da influência mútua de sentido formal e estrutural[24] e até os estudos recentes mostram que este meme – que palavras devem ser estudadas no contexto de outras palavras com as quais co-ocorrem em relações sintácticas directas e que essas combinações de palavras devem ser vistas em luz do seu contexto de ocorrência mais amplo (i. e. textual, situacional) – sobreviveu.

Mitchell examinou principalmente todas as variedades da delimitação sintagmática e a inter-relação entre a gramática e o léxico. Contribuiu para os estudos das colocações com as observações sobre a posição dos constituintes: configurações que contêm os mesmos morfemas lexicais nem sempre significam o mesmo se estes constituintes são deslocados ou inflectidos. Por exemplo, *hard* em *hard work* tem significado diferente do em *hard-working*. De maneira semelhante, *goings-on* não equivale a *on-going*.

[22] Os primeiros *corpora* surgem já nos anos 60, mas os métodos de processamento automático do texto estão longe de ser de qualidade pelo menos até aos anos 80.
[23] Usamos este termo aqui para designar o conjunto de teorias dos seguidores de Firth, não de ele próprio.
[24] Lexical particularities are considered to derive their formal meaning not only from contextual extension of a lexical kind but also from the generalized grammatical patterns within which they appear, and conversely, the recognition of general patterns is seen as only justifiable in response to selected comparisons of lexical combinations (1975: 114).

A seguir, devido ao interesse de Mitchell pelas relações gramaticais[25], vemos que nem todas as palavras duma colocação ocorrem em todos os padrões morfo-sintáticos teoricamente possíveis:

> *Heavy damage* is possible as well as *damage heavily* and *heavily damaged*, but not **heavy damaging*, whereas *heavy drinking* is possible as well as *drink heavily*, but not **heavily drunk* (1971: 52).

Segundo Mitchell, uma colocação é constituída por um lexema (ou o seu raiz) e por todas as suas formas derivadas (expressões como *a strong argument, he argued strongly, the strength of the argument,* e *his argument was strengthened* são consideradas instâncias da mesma combinatória).

Quanto à distância dos componentes, Mitchell (bem como Halliday que será analisado mais tarde) afirma que uma combinação de palavras pode ultrapassar fronteiras frásicas, dando como exemplo: *He didn't want the job. I don't think he even applied*.

**Sidney Greenbaum** (1929-1996) foi professor de inglês na *London University College* e *University of London* e autor de várias gramáticas e livros sobre uso de inglês.

Devido à confusão que havia no uso do termo colocação, Greenbaum propôs a distinção entre 'colocabilidade' ('collocability'), para potenciais co-ocorrências de palavras, e 'colocação' ('collocation'), para expressões que co-ocorrem frequentemente.

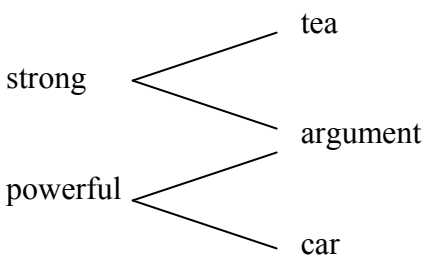
Greenbaum distingue entre dois tipos de restrição da colocabilidade dum item com outro em «certas relações sintácticas». Primeiro, a colocabilidade pode ser restringida pelas características semânticas dos constituintes[26]. **The boy may frighten sincerity* está incorrecto; ao trocarmos os substantivos, a frase fica correcta: *Sincerity may frighten the boy*. O segundo tipo de restrições não pode, porém, ser exprimido em termos gerais. É então necessário enumerar itens concretos que podem colocar com o item em questão. Greenbaum coloca os seguintes exemplos: **The man badly wished them to leave* em contraste com *The man badly wanted them to leave.* Apesar de *want, wish* e *desire* serem sinónimos, não criam colocações idênticas.

[25] Contudo, apesar de considerar as relações gramaticais como as mais importantes, este autor tem também em atenção as relações semânticas no estudo das colocações, tentando estabelecer um *continuum* entre as colocações, os compostos e as expressões idiomáticas.
[26] Este fenómeno já foi descrito por Chomsky (1965).

Quanto à consideração de factores semânticos e sintácticos, o autor defende uma abordagem integrada, i. e. o estudo de colocações deve ter em conta os dois[27]: «for an analysis of collocations that is divorced from syntax, [...] it does not seem possible to establish a criterion for determining whether two items are collocating» (1970: 11).

**Michael Halliday** (Michael Alexander Kirkwood Halliday, nascido em 1952) é um dos linguistas mais importantes do século passado, fundador do influente modelo gramatical, gramática sistémico-funcional[28].

Halliday interessa-se por dois problemas: o da colocação e o da influência que ela exerce em membros de conjuntos lexicais ('lexical sets'). Halliday viabilizou uma separação mais completa entre a semântica e a gramática, com o qual rematou as intenções de Firth. Como um exemplo da lexicalidade da colocação, Halliday compara a colocabilidade diferente de *strong* e *powerful*.




*Tea* coloca apenas com *strong* e *car* com powerful, enquanto *argument* com os dois. Além disso, esta relação mantém-se em várias configurações gramaticais: *He argued strongly against...; the strength of his argument; This car has more power,* etc. Em vista disso, o léxico opera independentemente das restrições gramaticais. *Strong, strength, strongly, strenghten* representam a «dispersão» ('scatter') do mesmo item lexical.

O facto de *strong* e *powerful* colocarem com *argument* permite-lhes entrar num conjunto idêntico. Contudo, cada um deles também entrará em conjuntos diferentes por as suas colocações (com *tea* e *car*) não se sobreporem ('non-overlapping collocations'). O

[27] Greenbaum chama a abordagem contrária (i. e. a que trata as colocações como um nível puramente idependente) 'item-orientated'.
[28] Halliday deixou rastro em várias áreas de linguística e semiótica social. As três metafunções de língua (ideacional, interpessoal e textual) pertencem aos seus conceitos mais citados e mencionados.

item, o conjunto e a colocação definem-se mutuamente. A colocação e o conjunto, como termos da descrição lexical, criam uma analogia com a estrutura ('structure') e o sistema ('system') na teoria gramatical: a diferença é que a colocação é uma relação de co-ocorrências *prováveis*, e os conjuntos são abertos. Um conjunto é um agrupamento de itens com um semelhante privilégio da ocorrência em colocação[29].

A seguir, Halliday marca a distinção entre 'colocadores fortes' e 'colocadores fracos' ('strong collocator' vs. 'weak collocator'). O colocador fraco combina-se, potencialmente, com quase qualquer substantivo comum. Como um exemplo serve o artigo: coloca com a maioria de substantivos. Contrariamente, o colocador forte como *blond* é altamente restringido: a *hair* e poucas palavras relacionadas (*tresses, wig,* etc.). O artigo deve ser deixado para que os gramáticos o descrevam; encontra-se num fim do *continuum* que passa do gramatical ao lexical: *blond* fica noutro fim. Os colocadores fortes podem, até certo ponto, predicar o seu próprio ambiente. Alguns itens predicam certas ocorrências doutros: quando a predicabilidade é 100 por cento (por exemplo *fro* sempre predica *to and*, e *kith* sempre predica *and kin*), podemos declarar que esta ocorrência fixa constitui um apenas item lexical.

Como é evidente, Halliday inclina a observações da criação das padrões lexicais baseadas em dados e frequências. Esta área foi, em seguida, desenvolvida por John Sinclair.

### 3.1.3 Colocações na era de linguística de corpus

A divulgação dos *corpora* e o aperfeiçoamento de técnicas de processamento automático do texto trouxeram consigo novas dimensões da análise linguística. John Sinclair e Michael Halliday foram dos primeiros linguistas que se aperceberam (e começaram a tirar proveito) das possibilidades que os *corpora* de grandes quantidades de textos autênticos legíveis pela máquina oferecem à análise linguística[30]. Podemos dizer que a «ressurreição» do interesse

[29] grouping of items with like privilege of occurrence in collocation

[30] Sinclair comenta sobre as possibilidades que os *corpora* trazem da seguinte maneira: «The big difference has been the availability of data. The tradition of linguistics has been limited to what a single individual could experience and remember. Instrumentation was confined to the boffin end of phonetics research, and there was virtually no indirect observation or measurement. The situation was similar to that of the physical sciences some 250 years ago» (1991: 1).

pelas colocações decorreu nos anos 80 graças a John Sinclair e o seu esforço na área de linguística de corpus.

**John McHardy Sinclair** (1933-2007), professor da Língua Inglesa Moderna na *Birmingham University*, é uma das personagens mais importantes no campo da linguística de corpus. É conhecido por ideias não convencionais e inovadoras que empurraram a linguística de corpus para frente[31].

Quanto ao estudo das colocações, Sinclair segue a linha neo-firthiana. Vamos tratar aqui apenas as suas obras recentes devido ao facto de que apesar de ele ter publicado livros já nos anos 60 e 70, reconsiderou as suas opiniões com a chegada da linguística de corpus[32]. Sinclair é o fundador do projecto COBUILD: além de ter recolhido e analisado 20 milhões de palavras (uma actividade que resultou no famoso dicionário COBUILD de 1987), chegou também a um entendimento mais profundo da co-ocorrência lexical restrita. Uma das observações mais importantes foi a da relação entre o significado e a estrutura; significados diferentes dum item foram frequentemente seguidos por preferidas configurações estruturais. Segundo uma outra observação, uma parte considerável da linguagem natural ocorre em 'frases semi-preconstruídas que constituem escolhas únicas apesar de parecer que estas podem ser divididas em segmentos pela análise'[33]. Esta observação transforma o conceito do léxico de um grande repositório de palavras em um grande repositório de sentidos cujos itens abrangem várias palavras ou frases inteiras. Para além disso, Sinclair também observou que diferenças mínimas no significado de uma palavra (incluindo sinónimos quase equivalentes) correspondem, normalmente, a diferentes padrões combinatoriais.

Sinclair estabelece dois princípios organizadores da língua, dois modelos de interpretação do significado das palavras: o 'princípio de livre escolha' ('the open-choice principle') e o 'princípio idiomático' ('the idiom principle'), simultaneamente alternativos e complementares[34].

---

[31] Sinclair comenta: «Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered quite possible but still lunatic. Today it is very popular» (1991: 1).

[32] Como ele próprio comenta no livro *Corpus, Concordance, Collocation*, «[we will describe the occurence of words] as we currently conceive it».

[33] semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments (1987)

[34] It is contended here that in order to explain the way in which meaning arises from language text, we have to

O princípio da livre escolha é uma maneira de ver o texto como um grande número de escolhas complexas. Sempre quando uma unidade (uma palavra, uma frase) está a ser completada, abre-se uma vasta gama de escolhas e a única restrição é a gramaticalidade. Isto é provavelmente a maneira normal de ver e descrever a língua, frequentemente chamada 'slot-and-filler model', ou seja, o texto é uma série de lacunas que têm que ser preenchidas pelo léxico que satisfaz as restrições locais. Em cada lacuna pode ocorrer praticamente qualquer palavra. Dado que a língua supostamente opera simultaneamente em vários níveis, um padrão de escolhas muito complexo está presente em cada momento; contudo, o princípio subjacente é bastante simples.

Como é óbvio, as palavras não ocorrem aleatoriamente e o princípio da livre escolha não é responsável por todas as restrições. Não seríamos capazes de produzir um texto apenas por procedermos de acordo com o princípio da livre escolha. Apesar de aplicarmos as restrições exigidas pelo tema e registo, ainda existem os demais candidatos que poderão preencher algumas das lacunas. Neste ponto é que entra em cena o princípio idiomático: o falante tem à sua disposição um grande número de grupos de palavras pré-construídos apesar de estes poderem apresentar alguma variação, nomeadamente a nível lexical, flexional ou de ordem das palavras.

A colocação, para Sinclair, representa o princípio idiomático. Em algumas ocasiões as palavras parecem ser escolhidas em pares ou grupos e estes não são necessariamente adjacentes. A colocação, então, pode ser definida como co-ocorrência de duas ou mais palavras dentro de um curto espaço num texto:

> Collocation is the occurence of two or more words within a short space of each other in a text (1991: 170).

Ao mesmo tempo, a colocação – como concebida por Sinclair em 1991 – reconhece apenas a co-ocorrência lexical das palavras. Considera-se que o falante utiliza as capacidades de memória e as experiências da vida rotineira e os seus discursos são preferencialmente constituídos por selecções correspondentes ao princípio idiomático.

---

advance two different principles of interpretation. One is not enough. No single principle has been advanced which accounts for the evidence in a satisfactory way (1991: 109).

A distância em que podem ocorrer as palavras de uma colocação é restringida por Sinclair a quatro palavras como máximo.[35]

John Sinclair ajudou a transformar a linguística numa disciplina moderna que tira proveito dos avanços técnicos. A sua contribução é evidente sobretudo na área de lexicografia: desde a criação do dicionário COBUILD, um *corpus* tem sido utilizado na criação de muitos dicionários, alguns autores até afirmam que a criação dum dicionário sem um *corpus* como a base é quase impensável.[36]

### 3.1.4. Abordagem frasal

Entre os autores fortemente influenciados por uma corrente linguística – neste caso por pragmática – contam-se **James R. Nattinger** e **Jeanette S. DeCarrico** da *Portland State University, Oregon*.

Como já mencionámos e como notam também Nattinger e DeCarrico, os linguistas têm recentemente à sua disposição um instrumento extremamente eficaz: os *corpora*. O facto de incluirem a linguagem autêntica em enormes volumes traz – ao lado das vantagens óbvias – mais um pró: a inclusão de fenómenos marginais.

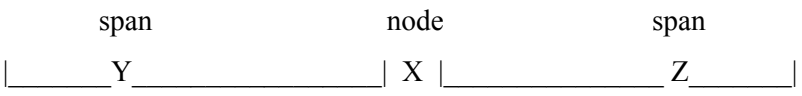
> One interesting aspect of these corpora is that they consist of authentic material, full of unexpected and diverse constructions, which are often treated as too peripheral or ill formed to be of much interest for theoretical grammars, and for this reason often require unconventional categories of description. Computers scan all these data, whether central or peripheral, for *collocations* (1992: 20).

Nattinger e DeCarrico defendem que enquanto a sintaxe rege combinações das classes gerais das palavras, as colocações desrevem itens lexicais específicos e a frequência com a qual ocorrem com outros itens lexicais[37]. Aqui poderíamos polemicar com o emprego do termo ‘colocações’ na formulação «descrevem itens lexicais específicos», pois não se trata

[35] The usual measure of proximity is a maximum of four words intervening (1991: 170).
[36] Lexicography is one of the linguistic applications that most have profited from corpus-based analyses. The appearance of COBUILD dictionary has given rise to a number of corpus-based dictionaries not only in English but in other languages as well. Today it is nearly unconceivable to think of building or expanding a dictionary without the help of a large corpus and computer tools (Uzeda-Garrão – Carmelita Dias, 2002: 357).
[37] Whereas *syntax* deals with general classes of words and their combinations, *collocations* describe specific lexical items and the frequency with which these items occur with other lexical items (1992: 20).

nem da colocação no sentido da combinação de palavras que ocorrem frequentemente juntas (de facto, trata-se de itens singulares) nem da colocação como fenómeno de frequente co-ocorrência de palavras (esta está descrita na continuação da frase: «frequência com a qual ocorrem com outros itens lexicais»).

Nem mais adiante está completamente claro o que os autores concebem sob o termo 'colocações', mas parece que as 'colocações' correspondem ao fenómeno da co-ocorrência permitida ou possível e a 'unidade colocacional' a palavras que co-ocorrem: «As colocações são definidas ao longo da dimensão sintagmática, aliás horizontal, e da dimensão paradigmática, aliás vertical. Isto é, uma unidade colocacional consiste de um núcleo ('node') e das entidades com as quais o núcleo se pode relacionar e que ocorrem de cada lado ('spans'). Os *spans* constam de classes de palavras preenchidas por itens específicos»[38]. O conceito pode ser representado da seguinte maneira:




Os *spans* dos quais escolhemos um (ou mais) para complementar o *node* encontram-se no eixo paradigmático e o resultado no eixo sintagmático. Consequentemente, as colocações podem ser definidas com base em frequência com a qual co-ocorrem os *nods* com certos *spans*; Nattinger e DeCarrico empregam a delimitação «a frequência maior que a coincidência»[39]. Contudo, nem aqui resulta claro qual é o significado da palavra 'colocação' a que se refere a definição.

A seguir, os autores afirmam que quanto maior a certeza de as palavras em *span* co-ocorrerem com *node*, tanto mais fixa e idiomática a colocação será[40]. No caso dos idiomas e lugares-comuns podemos falar das colocações completamente fixas, a expectativa mútua tornou-se fixa, bem sintagmática como paradigmaticamente, o que resultou em perda do

[38] Collocations are defined along a syntagmatic, or horizontal, dimension, and a paradigmatic, or vertical, dimension. That is, a collocational unit consists of a 'node' that co-occurs with a 'span' of word on either side. The span consists of particular word classes filled by specific items (Ibid.: 20).
[39] If it is the case that the node word occurs with a span of particular words at a frequency greater than chance would predict, then the result is a collocation (Ibidem).
[40] The more certain the words in the span are to co-occur with the node, the more fixed and idiomatic the collocation (Ibidem).

sentido por causa da eliminação do elemento da escolha[41]. Pelo contrário, as colocações menos fixas induzem maior variedade e mais possibilidades de combinação[42].

Nattinger e DeCarrico distinguem as colocações das 'sequências sintáticas' ('syntactic strings') num lado e 'frases lexicais' ('lexical phrases') noutro. Apresentemos a citação em inglês para que seja mais evidente a inconsistência no uso do termo 'collocation':

> 1. *Syntactic strings* are strings of category symbols, such as 'NP + Aux + VP', which are generated by syntactic competence and which underlie all grammatical (canonical) structures of the language.
> 2. *Collocations* are strings of specific lexical items, such as *rancid butter* and *curry favor*, that co-occur with a mutual expectancy greater than chance. These strings have not been assigned particular pragmatic functions by pragmatic competence.
> 3. *Lexical phrases* are collocations, such as *how do you do?* and *for example*, that *have* been assigned pragmatic functions [...] (Ibid: 36).

Sequências sintáticas são os padrões subjacentes e regulares. Colocações no ponto 2 são idênticas ao conceito apresentado por Sinclair. Colocações no ponto 3 têm um significado mais geral e o uso do termo depois da definição apresentada no ponto 2 cria confusão. Frases lexicais nem podem ser um sub-grupo das colocações por umas não terem funções pragmáticas particulares e as outras sim.

As frases lexicais ainda podem ser divididas em dois grupos,

(i) sequências de itens que não permitem nenhuma substituição (*what on earth, at any rate, by and large, as it were*)[43];

(ii) estruturas gerais que têm uma função pragmática e que são completadas por um componente escolhido por paradigma (*a ── ago, would you pass the ──?*)[44].

[41] With completely fixed collocations such as many idioms and clichés, mutual expectancy has become fixed, syntagmatically and paradigmatically ossified, which results in loss of meaning because of elimination of an element of choice (Ibidem).
[42] As collocations become less fixed, that is, as more variation becomes possible along both axes, predictability lessens and meaning increases (Ibidem).
[43] Strings of specific (non-productive) lexical items, which allow no paradigmatic or syntagmatic substitution. These strings can be both canonical (conforming to a syntactic string) and non-canonical (Ibid.: 36).

Segundo Nattinger e DeCarrico, as frases lexicais são frequentemente consideradas unidades lexicais apesar de serem pluriverbais e apesar de serem derivadas das regras sintácticas regulares, bem como outras frases. Os autores consideram-nas intermediários entre os níveis lexical e gramatical. O seu uso é governado pela competência pragmática, a qual também escolhe e atribui funções particulares às frases lexicais. O que é importante, segundo Nattinger e DeCarrico, as colocações (contrariamente às frases lexicais) não têm nenhuma função pragmática[45].

Como notam Nattinger e DeCarrico, os colocacionistas tendem para não ver as colocações em preto e branco e admitem que existe uma escala que abrange todos os possíveis graus entre a regularidade e idiomaticidade. Os autores adoptam a escala apresentada por M. Wood (1986). Segundo Wood, não existe ruptura qualitativa entre a sintaxe prefabricada e criativa; ela afirma que estas existem apenas em dois fins opostos do *continuum*. O seu estudo adianta mais do que os restantes em termos de proposição dos critérios sintácticos e semânticos na avaliação da «congelação» da forma. Enquanto a maioria dos colocacionistas emprega apenas o critério semântico para decidir se a combinação é plenamente composicional[46] ou não (isto é, se o sentido da colocação é completamente predicável através do significado dos elementos que a compõem), ela emprega também o critério sintáctico, perguntando se a forma da combinação é inteiramente produtiva ou não, ou seja, se a forma é única enquanto à estrutura, se é inteiramente obrigatória ou se está no meio entre os dois pólos (cf. Ibid.: 177).

No que respeita à relação com outras estruturas «congeladas», Wood (Ibid.: 178) cria o seguinte eixo:

| | | | | | | |
|---|---|---|---|---|---|---|
| idioms | – | collocation | – | colligations | – | free combinations |
| by and large | – | kick the bucket | – | off with his head | – | see the river |

[44] Generalized (productive) frames (by far the largest group), consisting of strings of category symbols (or otherwise generally specified syntactic/semantic features) and specific lexical items, which have been assigned a pragmatic function. Examples would be '*a* + N [+ time] + ago', and 'Modal + you + VP' (Ibid.: 37).

[45] Prefabricated phrases are collocations if they are sets of lexical items with no particular pragmatic functions; they are lexical phrases if they have such pragmatic functions (Ibid.: 37).

[46] fully compositional

Idiomas são únicos quanto à composição e à produtividade, sendo totalmente impredicáveis no que diz respeito à sua forma e seu sentido. Idiomas verdadeiros (*by and large, hell for leather, happy go lucky*) estão completamente «congelados» e há relativamente poucos dado que a maioria das frases permite um certo nível da composicionalidade e produtividade.

Colocações são predicáveis aproximadamente mas mesmo assim são restringidas a certos itens específicos e por isso podem ser denominadas «palavras». Um exemplo apresentado é *take umbrage*: trata-se duma colocação altamente restringida chamada 'interpretação composicional' ('compositional interpretation'), mas o substantivo bastante raro e o verbo bastante vago nos atrapalham. A substituição do substantivo é possível em muitas colocações; assim é que surgem famílias de idiomas[47] (*pay heed/attention, open-and-shut case/issue/problem*).

Falamos de coligações quando a substituição é limitada apenas pela categoria sintáctica e traços semânticos. Em geral, coligações são classes de colocações, nas quais pelo menos um dos constituintes é especificado pela categoria em vez de como um item lexical distinto; mistura-se então o ponto de vista formal e funcional, a restrição gramática e lexical. Como exemplo servem os verbos frasais (*tear/lope/race etc. ... up/along/across*).

## 3.2 Co-ocorrência lexical restrita na tradição linguística portuguesa

Os estudos sobre as colocações, tão enraizados na tradição linguística anglo-saxónica, não têm sido igualmente populares e divulgados na tradição linguística portuguesa, ou, melhor dito, não encontramos muitos textos que tratem explicitamente das colocações. O termo tem sido utilizado com maior frequência apenas nas obras recentes, sobretudo na parte de introdução das obras relativas à linguística de corpus. Mesmo assim, os conceitos mencionados são os de Firth, Sinclair e outros autores anglo-saxónicos.

Não obstante, se concebermos o fenómeno mais geralmente, encontraremos logo vários estudos sobre a co-ocorrência lexical restrita; na maioria dos casos aparecem no

[47] idiom families

âmbito da fraseologia. Outra observação que fizemos é a de que nisto há uma semelhança notável entre as linguísticas portuguesa e espanhola; por isso mencionaremos ocasionalmente também alguns autores espanhóis.

Neste capítulo apresentaremos mais concretamente e mais em detalhe algumas classificações correntes na linguística portuguesa. Não há unanimidade quanto à classificação da co-ocorrência lexical restrita e reparamos numa multiplicidade terminológica bastante grande. Encontramos termos como *frasema, colocação, solidariedade lexical, modismo, locução, frase feita, expressão idiomática, idiomatismo, expressão fixa, lexia complexa, unidade fraseológica, fraseologismo, sintagma, expressão* ou *construção fossilizada, etc.* que são utilizados para referir-se ou ao mesmo conceito ou a conceitos diferentes (cf. Iriarte Sanromán, 2001: 180). Não é propósito deste trabalho descrever a situação duma forma exaustiva, sendo mencionados apenas os nomes dos autores que nos parecem mais importantes e mais complexos. Assim trataremos sobretudo Herculano de Carvalho e Álvaro Iriarte Sanromán mas também os autores não portugueses que influenciaram a linguística portuguesa duma maneira significante, Bernard Pottier[48] e Igor Mel'čuk[49].

Antes de procedermos à procura da definição da colocação (aliás do conceito equivalente descrito pela fraseologia), será necessário definir o que entendemos por 'co-ocorrência lexical' e 'co-ocorrência lexical restrita'. Iriarte Sanromán (2001: 17) distingue-as da seguinte maneira:

> Entendemos por co-ocorrência lexical a capacidade das unidades lexicais para se combinarem em sintagmas de modo a exprimirem um determinado sentido. A co-ocorrência lexical é livre quando a combinação das unidades lexicais é feita segundo as regras gramaticais de uma língua.
>
> Entendemos por co-ocorrência lexical restrita qualquer tipo de combinação de duas ou mais palavras, sejam estas da classe que forem, em cuja construção intervém, para além das regras sintáticas e semânticas da língua, qualquer tipo de restrição puramente lexical, isto é, quando duas unidades lexicais não podem combinar-se sem haver qualquer regra gramatical

[48] Nascido em Paris em 1924, professor de linguística de várias universidades francesas.
[49] Igor Aleksandrovič Mel'čuk, nascido na Rússia em 1932, professor reformado da Universidade de Montréal. Trabalhou também no Instituto das Ciências da Linguagem em Moscovo onde colaborou na criação da TST, Teoria Sentido Texto ('Meaning-Text Theory', MMT).

que o impeça, como acontece com **ódio cego (cf. ódio mortal)* ou **fazer um passeio (cf. dar um passeio)*.

### 3.2.1 Álvaro Iriarte Sanromán

O autor dedicou-se ao estudo deste problemática já na sua tese de doutoramento, denominada *A Unidade Lexicográfica. Palavras, Colocações, Frasemas, Pragmatemas*, obra que pode ser caracterizada como uma síntese do que tinha sido escrito sobre a co-ocorrência restrita em linguísticas portuguesa e espanhola. Contudo, muitas teorias baseiam-se nas teorias inventadas por Igor Mel'čuk.

No presente trabalho serão apresentados vários conceitos tratados pelo dito autor. Manteremos também alguns equivalentes espanhóis que o autor, cuja especialização são dicionários bilingues português-espanhol, menciona. Assim veremos que todas as línguas – mesmo sendo de semelhança tão estreita como são as línguas portuguesa e espanhola – têm combinações restritas próprias. Os exemplos justapostos seguintes ilustram claramente que apesar de uma semelhança tão grande que existe entre o português e o espanhol, nem sempre podemos garantir a equivalência entre as combinações.

Em primeiro lugar, voltemos à divisão entre a co-ocorrência lexical livre e restrita. Duma maneira semelhante é também possível classificar as expressões pluriverbais. Iriarte Sanromán (tendo como o ponto de partida a distinção apresentada por Mel'čuk) distingue as expressões pluriverbais livres das não livres. Uma expressão pluriverbal é <u>livre</u> (não restrita) quando os seus componentes são seleccionados de acordo com as regras de selecção da língua arbitrariamente escolhidas. Nenhuma das regras é preceptiva, no sentido de que o falante pode aplicar quaisquer outras regras da língua para produzir um enunciado equivalente, tendo sempre a possibilidade de seleccionar significados e unidades lexicais, a possibilidade de livre escolha entre expressões e significados independentes (quase)equivalentes (cf. Ibid.: 182).

Por outro lado, uma construção é <u>regular</u> quando tanto o significante como o significado são construídos segundo as regras gerais da língua. A construção *ver um filme* é livre porque *um filme* pode ser substituído por qualquer objecto (dado que obedece as

regras semânticas, morfológicas etc.) e é também regular por o significado ser determinado pela simples união dos significados dos componentes.

Um sintagma não livre (combinação restrita, ou frasema) AB é uma combinação de dois ou mais lexemas A e B, cujo significante e cujo significado não podem ser construídos livre e regularmente por meio da soma regular ou união linguística dos seus componentes. Estruturas de tipo *perder a cabeça, baixar a cabeça, andar à nora, ser o braço direito, dar um passeio, ódio mortal, mudança radical, leite gordo, etc.* são exemplos de combinações não livres, cuja característica principal é que são estruturas holísticas ou não composicionais (cf. Ibid.: 183).

Como apresenta Iriarte Sanromán, a investigação e a descrição linguística relativa à combinatória lexical centra-se principalmente no que poderíamos chamar pelo nome genérico de unidade fraseológica (ou frase idiomática), entendida como o conjunto ou combinação de palavras (AB) cujo significado não pode ser deduzido do significado dos diferentes elementos lexicais que a compõem (por exemplo estruturas do tipo *perder a cabeça, baixar a cabeça*)[50]:

(A) + (B): 'AB' ≠ 'A' + 'B'

Mas existe outro tipo de combinações de palavras que poderiam ser consideradas como sendo expressões livres, mas que, especialmente quando comparadas com outras expressões de sentido equivalente noutras línguas, têm um pequeno componente idiomático que frequentemente escapa ao não nativo na hora de produzir um texto: são as colocações ou semi-frasemas (analisadas em detalhe mais adiante neste capítulo).

> Se queremos falar em 'leite ao que se não lhe retirou nata, ou gordura', deveremos dizer *leite gordo;* se queremos referirmo-nos ao 'cabelo que tem uma cor intermédia entre o dourado e o castanho-claro', dizemos *cabelo louro.* [...] Estamos perante um tipo de combinações lexicais fixas, uma vez que não temos liberdade para escolher qualquer adjectivo que possa exprimir a mesma ideia para acompanhar estes substantivos (**leite gorduroso, *leite oleoso, *leite untuoso*, **leite integral* ou **leite inteiro,* como acontece,

[50] Lyons (2005) usa a terminologia 'expressões sintagmáticas lexicalmente compostas' ('lexically composite phrasal expressions') e 'expressões sintagmáticas lexicalmente simples' ('lexically simple phrasal expressions'). Expressões lexicalmente simples caracterizam-se pelo facto de o seu significado ser diferente do da soma dos significados dos lexemas constituintes, a informação que nos é proposta não se pode deduzir dos respectivos componentes do grupo sintático correspondente.

> neste último caso com o espanhol *leche entera*). Contudo, não estamos perante expressões idiomáticas, uma vez que *leite gordo, cabelo louro* [...] são semanticamente composicionais, isto é, podemos saber o significado da expressão simplesmente pela soma dos significados dos seus componentes (Ibid.: 164).

Aqui vemos uma possível diferença entre os idiomas e as colocações. *Esticar o pernil* é uma unidade lexical e não combinação de unidades lexicais (como *comprar um livro*), trata-se dum idioma. Por outro lado, *leite gordo, ódio mortal, correr um risco,* não são unidades lexicais como no caso das expressões idiomáticas. Trata-se da ocorrência de duas ou mais unidades lexicais, como se pode constatar pelos seguintes exemplos, em que inserimos outros elementos linguísticos entre as duas unidades lexicais que formam a combinação:

> *mudança radical*: A mudança não foi assim tão radical como ele tinha anunciado.
> *dar um passeio*: Dar frequentemente um passeio.
> *leite gordo*: O leite é um bom alimento para as crianças, especialmente se for gordo.
> (Ibid.: 139)

Contudo, é óbvio que as combinações lexicais deste tipo não são totalmente livres, como é evidente se quisermos substituir algum dos componentes (**fazer um passeio*).

As restrições aplicam-se a várias estruturas. Por exemplo, nas estruturas substantivo+adjectivo, diferentes substantivos seleccionam diferentes adjectivos para exprimir a mesma ideia:

> *'muito' interesse* – vivo interesse;
> *'muita' mudança* – mudança radical;
> *'muita' obediência* – obediência cega;
> *'muito' desejo* – desejo ardente;
> etc. (Ibid.: 143)

O mesmo acontece na estrutura verbo + advérbio ou verbo + locução adverbial:

*recusar 'muito'* – recusar firmemente;
*acreditar 'muito'* – acreditar piamente;
*chover 'muito'* – chover a cântaros;
*trabalhar 'muito'* – trabalhar como um negro;
etc. (Ibidem)

Além das estruturas já mencionadas, à co-ocorrência restrita pertencem também casos de regime preposicional; estes têm um lugar importante nos dicionários codificadores e bilingues:

à tarde: por la tarde;
ir às compras: ir de compras;
espera por mim: espérame;
dar ordem para: dar orden de;
dar para o mar: dar al mar;
é parecido com: se parece a;
na realidade: en realidad.
(Ibid.: 19)

A fronteira entre as expressões livres e restritas às vezes não é clara, como documentam os seguintes casos onde a combinação de palavras pertence aos dois tipos:

*passar bem*: por algum lugar vs. cumprimento;
*fazer a barba*: de um boneco vs. barbear(-se);
*fazer uma cama*: um marceneiro vs. com os lençóis etc. (Ibid.: 141)

Alonso Ramos (1993: 190) distingue-as do ponto de vista semântico, destacando que para exprimir um determinado sentido, os lexemas podem combinar-se de maneira livre ou restrita. A co-corrência lexical livre é uma questão semântica, enquanto a co-ocorrência lexical restrita é uma questão lexical e nem sempre semântica.

No presente trabalho serão tratados apenas os fenómenos de combinatória de tipo lexical e sendo as de tipo semântico postas à parte. Contudo, para entendermos melhor a restrição lexical, contrastaremos esta com a restrição semântica.

A questão das restrições semânticas foi abordada já por Chomsky (1965). As restrições de selecção, como expressão e como conceito, foram originariamente introduzidas por ele através das seguintes designações: regras, restrições e restrição de co-ocorrência[51]. O objectivo delas é evitar a produção de frases agramaticais como:

*O rapaz atemoriza a sinceridade.

*A sinceridade admira o rapaz.

*O João resolveu o cachimbo.[52]

Segundo Vilela (1994: 133), estas regras fazem parte do componente básico da gramática, sob a forma de regras sintácticas. As regras de restrição devem 'bloquear' a formação de tais frases e, analogicamente com frases gramaticais, fornecer elementos que possibilitem a interpretação de certas construções produzidas com transgressão (das referidas restrições de selecção).

Neste ponto estamos em condições de apresentar em detalhe os tipos da co-ocorrência lexical restrita. Como introduz Iriarte Sanromán (2001: 17), «ao falarmos aqui em co-ocorrência ou combinatória lexical, falaremos em combinações que intuitivamente qualquer falante chamaria palavras. Assim, interessa-nos tanto o problema das colocações (semi-frasemas) [...], como o da fraseologia, ou expressões idiomáticas, em geral (frasemas completos) [...] do mesmo modo que (especialmente nos dicionários bilingues) vários tipos de combinações pluriverbais lexicalizadas e habitualizadas (quase-frasemas e pragmatemas) [...].» Neste ponto permitimo-nos a polemicar com a afirmação de que qualquer falante chamaria essas expressões palavras. Como já mencionámos, segundo o autor as colocações são compostas por duas ou mais unidades lexicais, então dizer que as colocações equivalem a uma palavra (=uma unidade lexical) é uma contradição.

A distinção entre as colocações/semi-frasemas, quase-frasemas, frasemas completos e pragmatemas foi elaborada por Iriarte Sanromán mas baseia-se originalmente na divisão apresentada por Mel'čuk (1995). Já que é uma distinção muito complexa (e frequentemente adoptada pelos autores portugueses e espanhóis), analisá-la-emos agora mais em detalhe.

Dentro das combinações restritas podemos distinguir os frasemas pragmáticos ou pragmatemas dos frasemas semânticos que são divididos em três tipos:

[51] selectional rules, selectional restrictions, restriction of coocurence
[52] Os exemplos tirados de Vilela (1994: 133).

(i) frasemas completos ou expressões idiomáticas;

(ii) semi-frasemas ou colocações;

(iii) quase-frasemas.

Distingamos primeiro os frasemas pragmáticos dos semânticos. Um frasema pragmático (pragmatema) é

> uma estrutura cujo significado 'X' não é construído livremente (embora possa ser regular) a partir de uma Representação Conceptual determinada (que o falante quer verbalizar), quer dizer, não pode ser substituído por qualquer outro significado sinónimo 'Y' construído livremente, por meio das regras gerais da língua, a partir dessa Representação Conceptual, para esse mesmo contexto situacional (Ibid.: 276).

Como exemplo servem *Consumir de preferência antes do fim de, Volto já, Pré-pagamento*, *Parabéns,* etc.

Um frasema semântico AB é

> uma combinação de dois ou mais lexemas A e B, cujo significante é a soma regular dos significantes dos lexemas constituintes /A + B/, mas cujo significado, é diferente do da soma dos significados dos lexemas constituintes. A diferença dos frasemas pragmáticos, num frasema semântico o significado é escolhido livremente, não é imposto pela situação; mas a expressão para este significado não é escolhida livremente, a sua selecção (a nível lexical) é parcial ou totalmente limitada, restringida, por este significado, enquanto morfológica e sintacticamente (gramaticalmente) pode ser uma expressão regular (Ibid.: 184).

Como já dissemos, temos três tipos de frasemas semânticos: frasemas completos, semi-frasemas/colocações e quase-frasemas. Ao querermos distinguir entre eles, é preciso ter em conta o tipo de restrição que opera na selecção (a nível lexical) dos lexemas que compõem o frasema.

Um frasema completo ou uma expressão idiomática (Sanromán Iriarte usa os dois termos) AB ([ser o] *braço direito*) é uma combinação de dois ou mais lexemas A (*braço*) e

B (*direito*), cujo significante é a soma regular dos significantes dos lexemas constituintes /A + B/ (*braço* + *direito*), mas cujo significado não é a esperada união regular de A e B ('A + B'), isto é 'braço direito', mas um significado diferente 'C' ('[ser o] 'auxiliar principal' ou 'principal colaborador'), que não inclui nem 'A' nem 'B' (cf. Mel'čuk 1995, 1998, Iriarte Sanromán 2001: 185). Algumas características gerais dos frasemas completos são destacadas por Alonso Ramos (1993: 182):

> - pertenecen al lenguaje literal, esto es, deben ser reproducidos en sus propios términos;
> - son no composicionales semánticamente: la suma del sentido de sus constituyentes no es igual a su sentido global;
> - son cohesivos: sus elementos constituyentes están exigidos unos por otros;
> - resisten, con diferentes grados, a la variación formal;
> - pueden ser ambiguos: algunos tienen una contrapartida homófona composicional;
> - algunos presuntos frasemas son, de hecho, colocaciones; deben ser disueltos en un lexema separado (palabra llave) y un valor de la FL aplicada a ese lexema;
> - su carácter como unidades semánticas y su conservación de algunos rasgos del sintagma los convierte en unidades difíciles de tratar en un modelo lingüístico.

Como exemplos podemos mencionar *levantar a cabeça* ('prosperar'), *baixar a cabeça* ('obedecer'), [andar] *à nora* ('[andar] desorientado'), [ser] *o braço direito* ('[ser] o principal auxiliar'), *colete-de-forças* ('peça de roupa empregada para dominar os movimentos dos braços'), *mercado negro* ('comércio ilegal ou clandestino'), *mesa-redonda* ('debate em que os participantes se encontram ao mesmo nível'), *ponte aérea* (comunicação regular entre dois pontos por meio de aviões), *pele-vermelha* ('índio indígena norteamericano') etc.

Antes de analisarmos as colocações, falemos brevemente dos quase-frasemas. Os quase-frasemas são frasemas em que, para além de se conservarem os sentidos dos lexemas que os constituem, acrescenta-se um novo sentido que não é dedutível da simples soma dos sentidos dos lexemas constituintes. Como exemplos servem *tecto falso* (onde, para além dos sentidos 'tecto' e 'falso' temos também o sentido 'para isolar acústica e termicamente'), *cinturão negro* (para além dos sentidos 'cinto' e 'negro' temos também o

sentido ‘grau de conhecimento ou habilidade em artes marciais’), *centro comercial* (‘lugar onde são agrupadas determinadas actividades’ + ‘relativo ao comercio’ + ‘com muitas lojas, serviços, parques, etc.’) (cf. Iriarte Sanromán 2001: 194).

Neste ponto estamos a chegar a semi-frasemas ou colocações como os concebe Iriarte Sanromán na sua obra *A Unidade Lexicográfica*. O autor não pretende descrever a história do termo e do conceito, mas sim analisar a validade deste conceito na descrição lexicográfica duma língua[53]. Como o ponto de partida serve, por ser considerada coerente e completa, uma concepção de colocação utilizada no modelo lexicográfico de Igor Mel’čuk, descrita e apresentada para o espanhol por Alonso Ramos (1993).

Se quisermos distinguir as colocações dos frasemas completos por um lado e das combinações livres por outro lado, veremos que a fronteira não é clara: há casos em que é difícil classificar determinadas combinações de palavras como sendo combinações livres ou restritas. Contudo, «há uma aspecto que hoje já não se pode pôr em causa: o facto de que as chamadas colocações não são combinações livres de palavras, mas antes um tipo de unidades pluriverbais lexicalizadas e habitualizadas» (Iriarte Sanromán 2001: 187). Segundo Alonso Ramos (1993: 183), a característica mais importante que diferencia um frasema completo de uma colocação é o facto de que

> en un frasema, ninguna de sus propiedades semánticas ni sintácticas son deducibles de los lexemas constituyentes. Sin embargo, en la colocación, al menos algunas propiedades son deducibles de uno de los lexemas. En *actividad febril* o en *dar un paseo*, el nombre guarda las mismas propiedades que tiene fuera de la combinación.

Uma colocação, ou semi-frasema, AB é

> uma combinação de dois ou mais lexemas A e B, cujo significante é a soma regular dos significantes dos lexemas constituintes /A + B/, e cujo significado ‘X’ inclui o significado do lexema A mais um significado ‘C’ ('X’ = ‘A + C’), de tal maneira que o lexema B que exprime ‘C’ não é seleccionado livremente. Numa colocação, pensemos por exemplo em *ódio mortal,* um dos seus elementos constituintes, A (*ódio*), é seleccionado pelo falante por causa do seu significado, que é conservado intacto; mas o

[53] Em *A Unidade Lexicográfica*, a teoria serve apenas como base para os estudos metalexicográficos.

> segundo elemento constituinte, B (*mortal*), significa 'C' ('intenso'), diferente de 'B' ('que causa ou pode causar a morte'). Fora da colocação AB, B (*mortal*) não seria usado para exprimir 'C' ('intenso') (Iriarte Sanromán, 2001: 188).

Consoante a natureza de 'C' podemos distinguir quatro tipos de colocações[54] :

A. Colocações formadas por verbo operador (ou seja, uma palavra funcional verbal) ou verbo-suporte[55] mais nome, por exemplo *dar um passeio*. O lexema A, que conserva intacto o seu significado (*passeio*) é acompanhado por outro lexema, B (*dar*), que se esvazia de significado, funcionando apenas como um verbo operador. O lexema A (sublinhado) é também chamado 'base' ou 'palavra-chave' da colocação. Outros exemplos são *dar um conselho, infligir uma derrota, reinar o silêncio, tomar em consideração, pôr em dúvida*, etc.

B. Este tipo é representado por estruturas de tipo *ódio mortal*. Neste tipo de colocações, o segundo elemento constituinte – B (*mortal*) – não está vazio de significado (como no caso anterior), ou seja, B tem no dicionário o correspondente significado 'C' mas este sentido ('intenso, vivo') apenas é actualizado em combinação com o lexema A (*ódio*) e não pode ser exprimido por qualquer sinónimo de B. Fora da colocação AB, B (*mortal*) não seria usado para exprimir 'C' ('intenso'). As colocações formadas com substantivos mais adjectivos intensificadores, ou verbos mais advérbios, são exemplos deste tipo de semifrasemas: *ódio mortal, mudança radical, vontade louca, confessar abertamente, proibir terminantemente,* etc.

C. Neste caso, o segundo elemento constituinte do semi-frasema – B (*amarelo* em *sorriso amarelo*) – não está vazio de significado, como no primeiro caso, mas o sentido 'C' que exprime ('forçado, contrafeito') não aparece no dicionário como acepção do lexema B, senão que apenas é actualizado em combinação com o lexema A (*sorriso*) ou com muito

---

[54] Esta distinção foi, outra vez, proposta por Mel'čuk (1995): 1) **either** 'C' ≠ 'B' i. e. does not have (in the dictionary) the corresponding signified; **and [ a.** 'C' is empty, that is, the lexeme **B** is, so to speak, a semi-auxiliary used to support a syntactic configuration; **or b.** 'C' is not empty but the lexeme **B** expresses 'C' only in combination with **A** (or with a few other similar lexemes); 2) **or** 'C' = 'B', i.e, **B** has (in the dictionary) the corresponding signified; **and [ a.** 'B' cannot be expressed by any otherwise possible synonym; **or b.** 'B' includes (an important part of) the signified 'A', that is, it is utterly specific**]**.

[55] Um verbo semanticamente vazio que combinado com um deverbal forma uma estrutura que mantém uma relação de paráfrase com o verbo de que deriva o nome, como, por exemplo: *dar um passeio = passear; tirar uma conclusão = concluir*, etc.

poucos lexemas mais. Quer dizer, *amarelo* não tem no dicionário, entre as suas diferentes acepções, o sentido de 'forçado, contrafeito' porque realiza este sentido apenas com *sorriso.* Outros exemplos deste tipo de colocações são: *imprensa amarela* (ou, no Brasil, *imprensa marrom*), *ponte levadiça, ponte branca, chave mestra, parede mestra,* etc.

D. As colocações do último tipo são também chamadas *solidariedades lexicais*. Neste tipo de colocações, o segundo elemento constituinte – B (*aquilino* em *nariz aquilino*) – não está vazio de significado, como no primeiro caso, mas o sentido 'C' que exprime ('curvo como bico de águia') aparece no dicionário como acepção do lexema B, mas apenas é actualizado em combinação com o lexema A (*nariz*). O lexema *nariz* está incluído como traço semântico na definição de *aquilino*. Este adjectivo apenas se diz de *nariz.* Outros exemplos deste tipo de colocações são: *manteiga rançosa, cabelo louro, vestido afogado, cavalo baio, cavalo acarneirado, cavalo pezenho, cavalo zarco, vinho abafado,* etc. (cf. Iriarte Sanromán, 2001: 188-191).

## 3.2.2 Herculano de Carvalho

A classificação a mencionar a seguir será a de J. G. Herculano de Carvalho. Não trata explicitamente das colocações ou idiomas, mas de conceitos semelhantes que possam ser (e, com efeito, de vez em quando são) confundidos com colocações e idiomas. Também são interessantes as observações sobre as diferenças entre os sintagmas livres e fixos, pois são critérios que nos possam ajudar a encontrar as características típicas das expressões não livres, inclusive as colocações e idiomas.

Herculano de Carvalho distingue três graus na gradação duma palavra a um sintagma:

> sintagma livre: *ficar de boca aberta, a boca do lobo*;
> sintagma fixo: *um ramo de bocas-de-lobo*;
> palavra composta: *ficar boquiaberto* (1979: 509).

O sintagma fixo não é de facto uma verdadeira palavra porque conserva algumas propriedades sintagmáticas. Pode ser definido como

> uma associação de palavras em sequência fixa, que constitui uma unidade sintáctica perfeita (funcionando como sintagmema mono-léxico) e também muitas vezes semântica (significando um conceito simples) e morfológica (1979: 522).

Os limites entre o sintagma livre e o sintagma fixo são claros. O sintagma fixo representa um conceito simples (*guarda-roupa*), do mesmo modo que a palavra, enquanto o sintagma livre significa um conceito complexo (*guardar a roupa na gaveta* = *guardar* + *roupa*). Consequentemente, a significação do sintagma fixo não resulta da combinação do significado dos seus termos (*guarda-chuva, bocas-de-lobo*), como acontece com o sintagma livre (*guardar a roupa, boca de lobo*).

### 3.2.3 Bernard Pottier

Bernard Pottier é um autor francês mas, bem como Mel'čuk, é frequentemente citado e adaptado pelos autores portugueses, portanto achámos importante mencionar aqui o seu (bem conocido e divulgado) conceito das lexias[56], unidades lexicais memorizadas (1972: 16). Na obra *Grammaire de l´Espagnol*, adaptada, posteriormente, para o português com o nome de *Estruturas Lingüísticas do Português*, Pottier marca distinção entre as lexias da seguinte forma:

> lexia simples: *árvore, saiu, entre, agora*;
> lexia composta: *primeiro-ministro, guarda-florestal, olho-de-sogra*;
> lexia complexa estável: *estado de sítio, cesta básica, uma estação espacial, Cidade Universitária*;
> lexia textual: *"quem tudo quer, tudo perde"*.

[56] O termo serviu principalmente para que não houvesse confusão entre os diferenets sentidos de *palavra* ou *vocábulo*. Veja-se o comentário em *O tratamento das lexias compostas e complexas*: «Ele foi muito feliz em cunhar este termo lingüístico, pois há muita confusão quando se usa vocábulo ou palavra. São vagos e imprecisos estes termos, de longa tradição na lexicografia tão pobre de ciência, conquanto, hoje, bem produtiva, em Língua Portuguesa. Embora estes termos, normalmente, identifiquem o plano das realizações discursivas, é de bom alvitre, em ciência, a precisão terminológica. Assim, lexia seria um bom termo para a manifestação do lexema já aceito como a unidade abstrata do léxico. A lexia, lexicalmente, seria, então, a manifestação discursiva do lexema.»

Pottier também diferencia entre lexia rígida, entendida como «uma sequência memorizada invariável» (1978: 270, em Iriarte Sanromán, 2001: 152), dando exemplo como *meter a mão, caso de honra, água pesada;* e lexia variável, que «se compõe de um quadro estável e de uma zona estável» (Ibidem.), com exemplos como *tudo leva a pensar/crer/supor que...*

## 3.3 Colocações na tradição linguística checa

Um dos primeiros linguistas checos que mencionam colocações foi Vilém Mathesius da Escola de Praga. Porém, quase não as menciona em relação ao checo senão ao inglês. No livro *A functional Analysis of Present Day English on a General Linguistic Basis* (1975) dedica-lhes quatro páginas. Contudo, não apresenta nenhuma classificação nem desrição sistemática, limitando-se a dizer que expressões como *evening paper* são colocações e não palavras compostas e que as palavras compostas tão abundantes por exemplo em alemão se tornaram, em inglês contemporâneo, em colocações (*Vergissmeinnicht – forget-me-not*). Ele também considera como colocações expressões de tipo *Oxford University Summer Vacation Course* (outra forma possível de *Summer Vacation Course of Oxford University*). A «definição» da colocação é simples: «uma combinação fechada e fixa»[57].

Um tratamento mais sistemático e vasto pode ser encontrado na obra de František Čermák, um dos linguistas checos mais importantes. Ele trata as colocações tanto do ponto de visto da linguística tradicional como do ponto de vista da linguística computacional. Quanto à primeira abordagem, esta é representada pelo livro *Frazeologie a idiomatika česká a obecná. Czech and General Phraseology.* O livro é composto por várias contribuições a conferências, anais e revistas linguísticas publicadas pelo autor nos passados 35 anos ou em inglês ou em checo; algumas são novas e publicadas pela primeira vez. Aqui Čermák aborda a problemática das colocações num contexto mais amplo, no âmbito da fraseologia**.** Em primeiro lugar, o autor reconhece que mesmo em 2007 a problemática das colocações está longe de ser resolvida definitivamente. Čermák também critica a linguística tradicional e a abordagem chomskiana: como ele salienta, a análise sintáctica (ou gramatical em geral) não é suficiente, pois as regras mais importantes são as

[57] a close, fixed combination

regras semânticas que sempre precedem todas as regras formais. Os nossos enunciados estão compostos por combinações feitas e por estereótipos que não podem ser analisados ao utilizarmos apenas abordagens tradicionais. Čermák critica que nenhuma das gramáticas aceitou este facto, todas fingindo que oferecem uma descrição exaustiva. Sob esta crítica, a gramática chomskiana está restringida apenas à sintaxe e os exemplos apresentados são hipotéticos e desligados da realidade. Pelo contrário, a linguística de corpus (que – contrariamente – usa exemplos reais, como já mencionámos) recusa esta abordagem generativa, sabendo que os textos e enunciados reais não podem ser definidos desta maneira; concentram-se na criação do significado através de combinações de formas e como estas combinações são formadas – segundo regras ou sem elas. Nas conclusões parciais dos estudos de grandes *corpora* vemos que pouco sabemos sobre o que é regular na língua e como é difícil distinguir entre *langue* e *parole*. Apesar disso, como o autor salienta, as gramáticas tradicionais outra vez fingem que são capazes de fazê-lo.

Segundo Čermák, as colocações ('kolokace', 'collocations') ainda estão a ser estudadas e falta muito até serem completamente entendidas e descritas. Os frasemas e idiomas formam um sub-grupo das colocações. Essas expressões são a única área da língua onde as regras semânticas e gramaticais convencionais nunca serão válidas inteiramente; trata-se sempre duma forma de anomalia que simplesmente não cabe sob uma fórmula ou um algoritmo generativista. Essas expressões devem ser estudadas sob um ângulo diferente. Daqui resulta a oposição da fraseologia – i. e. de algo anómalo e irregular – e da língua regular, i. e. de algo regido pelas regras semânticas e gramaticais. Para além disso, os estudos contemporâneos mostram que uma grande parte dos nossos textos e enunciados não é gerada como produtos originais de falantes: não somos tão criativos. Está a tornar-se evidente que uma parte significante dos textos é prefabricada e tira proveito das combinações já feitas, dos frasemas, etc.: frequentemente fala-se de estereótipos. Futuros estudos devem revelar o que na língua é formado e produzido por falantes e o que deve ser concebido como unidades pré-feitas. Uma ferramenta importante é os *corpora* que nos possibilitam um conhecimento sistemático da sintagmática da língua.

Nem o frasema nem o idioma podem ser delimitados por um (anómalo) traço, seja formal, semântico ou colocacional. Čermák critica o entendimento fixo da colocação como uma ligação de itens fixada e reproduzível cujo significado não é (parcialmente ou

inteiramente) dedutível do significado dos seus componentes, por este não convir a todos tipos (por exemplo aos frasemas com componente mono-colocável). Em fraseologia, as anomalias são percebidas no fundo das relações semânticas e formais da língua regular e são reconhecidas graças ao contraste com a maioria regular. Uma anomalia é um traço constitutivo. Enquanto as combinações regulares se baseiam nas regras semânticas e formais e possibilitam formar novas (e ilimitadas) combinações (*abrir um livro, pacote, olho*), as combinações fraseológicas anómalas são sempre únicas e nenhum dos seus componentes pode ser substituído por um outro com a mesma função ou significado (*cabeça aberta* vs. **costas abertas, *braço aberto*).

O quê então Čermák considera sendo colocações? Primeiro, sem defini-las explicitamente, o autor oferece a seguinte lista (2006: 9), arguindo que um falante nativo sabia criar pares (i. e. colocações) correctos sem nenhum problema.

a) {*malovat, psát, tvořit*} : {*dopis, obraz, symfonie*}
({*pintar, escrever, criar*} : {*carta, sinfonia*})

b) {*hluboce, vysoce*} : {*skličující, zajímavý*}
({*profundamente, altamente*} : {*deprimente, interessante*}

c) {*vůně (jídla), vůně (posekaného trávníku), zápach (čpavku)*} : {*linout se*}
(colocabilidade restringida ao cheiro bom/mau)

d) {*na, v*} : {*Morava, Čechy*}
(a preposição *em* é diferente ao preceder regiões diferentes; em português poderiamos encontrar exemplos semelhantes em (não-)emprego do artigo definido na denominação de alguns países)

e) {*všanc, najevo, na holičkách*}
(expressões mono-colocáveis que não encontramos noutras combinações)

Outro item da lista são, como já mencionámos, os idiomas e frasemas, «ligações únicas de dois itens no mínimo, dos quais algum (ou nenhum) não funciona da mesma maneira em outra locução (ou em mais locuções), ou encontra-se numa só expressão (ou poucas

expressões»[58]. Um traço característico – mas não obrigatório – do frasema/idioma é a figuratividade[59], i. e. os componentes frequentemente não são traduzíveis literalmente. A dicotomia entre os dois conceitos consiste na formalidade/semântica da sua interpretação: ao analisarmos uma formação combinatória do ponto de vista formal, falamos de um frasema. Pelo contrário, o idioma relaciona-se com a análise semântica. Para além disso, Čermák apresenta também os termos quasi-frasema ('kvazifrazém'; *dávat pozor – prestar atenção*) e termo ('termín'; um exemplo de termo em português seria *cal apagada*).

Čermák (2006: 12-13) também apresenta uma divisão das combinações lexicais baseada na distinção

(i) sistema vs. texto (langue vs. parole, fixo vs. não fixo);

(ii) regular vs. irregular (do ponto de vista formal e semântico);

(iii) colocação vs. não colocação;

(iv) nomes próprios vs. não próprios:

(i) Sistémicas (Systémové)

1. Regulares:

a. **colocações terminológicas**, termos pluriverbais ('**termínové kolokace**, víceslovné termíny')

*cestovní kancelář - agência de viagens*

*kyselina sírová – ácido sulfúrico*

b. **colocações próprias**, nomes próprios pluriverbais ('**propriální kolokace**, víceslovná propria')

*Kanárské ostrovy – Ilhas Canárias*

*Velká Británie – Grã-Bretanha*

2. Irregulares:

**colocações idiomáticas**, idiomas e frasemas ('**idiomatické kolokace**, idiomy a frazémy')

[58] Idiom a frazém je jedinečné spojení minimálně dvou prvků, z nichž některý (popř. žádný) nefunguje stejným způsobem v jiném spojení (resp. více spojeních), popř. se vyskytuje pouze ve výrazu jediném (resp. několika málo) (2006: 31).

[59] Příznačným, ne však obligatorním rysem takového frazému (idiomu) je i jeho častá přenesenost (Ibid.: 646).

*ležet ladem, údolí stínů, jen aby* – exemplos em português com uma função semelhante seriam *viver às moscas, pois é*, etc.

extensões e transições

*stará dobrá Anglie, černá díra – a boa velha Inglaterra, buraco negro*

(ii) Textuais (Textové)

3. Regulares

a. **colocações comuns**, combinações gram.-semânticas (‘**běžné kolokace**, gram.-sém. kombinace’)

*letní dovolená – férias de Verão*

*snadná odpověď – resposta fácil*

*dřevěná tužka – lápis de madeira*

b. **combinações analíticas das formas,** formas analíticas (‘**analytické kombinace tvarů**, analytické formy’)

*byl zapsán – foi inscrito*

*vzpomínající si – lembrando-se, recordando*

4. Irregulares

a. **colocações metafóricas individuais**, metáforas autorais (‘**individuální metaforické kolokace**, autorské metafory’)

*virové hrátky – brincadeiras virulentas*

b. **combinações circunvizinhas acidentes** (‘**náhodné kombinace sousední**’)

*(vývody) vzduchotechniky uvnitř (bytu) – (saídas) de ar condicionado dentro (dum apartamento)*

*že v – que em*

c. **outras combinações**, disparate (‘**jiné kombinace**, blábol’)

(iii) Textuais-sistémicas (‘Textové-systémové’)

5. **colocações usuais comuns** (**běžné kolokace uzální**)

*prát prádlo – lavar roupa*
*umýt si ruce – lavar as mãos*
*nastoupit do vlaku – embarcar no comboio*

Como na Grã-Bretanha, também na República Checa uma revolução nos estudos das colocações deu-se com a chegada da linguística computacional, sobretudo da linguística de corpus. Um verdadeiro centro dos estudos da linguística computacional formou-se na Universidade Carolina de Praga, na Faculdade de Letras (*Ústav teoretické a komputační lingvistiky, Instituto da linguística teórica e computacional*) e na Faculdade de Matemática e Física (ÚFAL – *Ústav formální a aplikované lingivistiky, Instituto da linguística formal e aplicada*). Em 2000 foi fundado o *Centro da linguística computacional* (*Centrum komputační lingvistiky*) como resultado da cooperação da Universidade Carolina, do Instituto da língua checa da Academia das ciências (Ústav pro jazyk český Akademie věd) e da Universidade de Pilsen (Západočeská univerzita Plzeň).

A linguística computacional trouxe novas abordagens baseadas em modelos matemáticos e estamos a chegar ao ponto de sermos capazes de extrair as colocações dos textos automaticamente. Alguns métodos serão apresentados mais em detalhe em 5.1. Os métodos de extracção automática de textos, em checo, foram desenvolvidos sobretudo por Pavel Pecina (2002, 2008). No caso de Praga, a extracção automática é facilitada por PDT, *Prague dependency treebanks*, um método da anotação de textos sofisticada. Anotam-se três níveis: morfológico, analítico (sintáctico) e textual-gramático (o significado linguístico). Assim, tendo em conta o nível de significado (omitido na maioria dos *corpora*), é possível reconhecer significados diferentes de uma palavra em combinações diferentes.

Quanto à classificação das colocações, Pecina e Holub (2002: 9-10) propõem a seguinte divisão. Contudo, os autores não a explicam em detalhe e é difícil distinguir entre os tipos propostos.

(i) **Colocações não-composicionais** (‘Kolokace dokonale nekompoziční’)

*natáhnout bačkory – bater as botas*
*Coca Cola*

O significado global é completamente diferente do signifidado dos componentes. A este grupo pertencem idiomas e também nomes próprios que não designam o género que denominam: não contêm nenhum nome genérico (de *Coca Cola* não se pode deduzir que se trata duma bebida).

(ii) **Colocações parcialmente não-composicionais** (‘Kolokace částečně nekompoziční’)

*Národní třída (em português e.g. Avenida de Liberdade)*

*Červený kříž - Cruz Vermelha*

*koruna stromu – topo da árvore*

O significado de pelo menos um componente **não** é incluído no significado global da colocação. Isto deve-se sobretudo ao facto de se tratar de palavras homónimas cujo significado apenas é determinado através da presença numa colocação. Frequentemente são também nomes próprios que incluem o género que denominam (avenida, largo etc.).

(iii) **Colocações minimamente não-composicionais** (‘Kolokace slabě nekompoziční’):

*řidičský průkaz – carta de condução*

*tisková konference - conferência de imprensa*

Os significados dos componentes **são** incluídos no sentido global mas sempre existe ainda um «valor acrescentado».

(iv) **Colocações composicionais – locuções fixas sem significado adicionado** (‘Kolokace kompoziční – ustálená slovní spojení bez přidaného významu’)

*nová verze – nova versão*

*nový rok – ano novo*

*velký objem - grande volume*

O seu significado é plenamente composicional, não contém nenhum «valor acrescentado»; são características pelo uso frequente e automatizado. A única regência é a sintática.

(v) **Colocações composicionais – palavras pertencentes ao mesmo ou próximo campo semântico** (‘Kolokace kompoziční - sémanticky příbuzná nebo blízká slova’)

*kapitán – loď (capitão – navio)*

*otázka – odpověď (pergunta – resposta)*

*lékař – pacient (médico – paciente)*

*starý – nový (velho – novo*)

O significado é plenamente composicional, mas não existe nenhuma regência sintática, apenas aparecem em contextos idênticos.

Pecina e Holub sugerem o seguinte teste para determinar se a colocação é composicional ou não: A alguém que conhece bem (perfeitamente) os significados dos componentes são apresentadas várias colocações. Se ele entender o significado global, trata-se de uma colocação composicional. Caso contrário, estamos perante uma colocação não-composicional.

## 3.4 Síntese das teorias e princípios da selecção de colocações para o dicionário de colocações checo-português

Como vimos, a problemática das colocações é muito complexa e dificilmente chegaríamos a um compromisso entre todas as classificações propostas. Nós próprios teremos que escolher o que entenderemos sob o termo 'colocação' na criação do dicionário de colocações checo-português. Há vários aspectos a levar em consideração e alguns autores baseiam sua classificação nos critérios que seguem abaixo. Contudo, consideramos esses critérios inconvenientes:

(i) **não-composicionalidade**

O significado global duma expressão não composicional não corresponde à soma dos significados dos componentes. Alguns autores, sobretudo os que consideram os idiomas sendo um sub-grupo das colocações (Čermák: *cabeça aberta* vs. **costas abertas, *braço aberto*) consideram a não-composicionalidade ser um traço típico das colocações. Não obstante, outros autores (Pecina) defendem que as colocações abrangem todos os graus no *continuum* da composicionalidade à não-composicionalidade.

As expressões que pretendemos incluir no dicionário são tanto composicionais (*lesní plod – fruto silvestre*) como não composicionais (*rodinný kruh – seio familiar*).

(ii) **possibilidade de substituição**

Alguns autores (Iriarte Sanromán) argúem, que não é possível substituir os elementos duma colocação (*dar um passeio* vs. **fazer um passeio*) mas há também evidência que às vezes uma substituição é possível (*fazer uma festa, dar uma festa*).

(iii) **possibilidade de modificação ou de inserção de outro elemento**

Há sugestões que as colocações não podem ser modificadas sem que o significado seja mudado (Mitchell: *ongoing* vs. *goings-on*) mas por outro lado há evidência que algumas transformações (*tomar uma decisão, uma decisão foi tomada, tomada de decisões*) bem como inserção de outros elementos (*dar frequentemente um passeio*) são possíveis.

(iv) **distância**

Sinclair limita a distância dos componentes duma colocação a quatro palavras, enquanto Mitchell e Halliday dão uma prova de a colocação poder ultrapassar os limites duma frase.

O único critério que parece agradar à maioria dos autores – e a nós também – é o critério da problematicidade da tradução para uma língua estrangeira. Contudo, como é um critério dificilmente mensurável, resolvemos enumerar os tipos de colocações que pretendemos incluir no dicionário duma forma explícita e taxativa:

(i) colocações-unidades lexicais
*daňový poplatník – contribuinte*

(ii) colocações do tipo B na classificação de Mel'čuk/Iriarte Sanromán
*železná vůle – vontade de ferro*

(iii) colocações do tipo C na classificação de Mel'čuk/Iriarte Sanromán
*padací most – ponte levadiça*

(iv) colocações do tipo D na classificação de Mel'čuk/Iriarte Sanromán
*žluklé máslo – manteiga rançosa*

(v) frasemas completos na classificação de Mel'čuk/Iriarte Sanromán
*černý trh – mercado negro*

(vi) quase-frasemas na classificação de Mel'čuk/Iriarte Sanromán
*nákupní centrum – centro comercial*

Além dos critérios acima apresentados, as colocações seleccionadas deverão ser

(i) colocações compostas por duas, raramente mais, palavras;

(ii) colocações nominais, i. e. de forma substantivo+adjectivo ou substantivo+substantivo;

(iii) colocações correntes no português europeu;

(iv) colocações correntes no português comum, i. e. não expressões técnicas e específicas.

# 4. Dicionários de colocações

## 4.1 O problema de metalexicografia

Quando começámos a nossa pesquisa, ficámos surpreendidos que as obras dedicadas à semântica prestem apenas mínima atenção ao problema da unidade semântica ou lexical, aplicável – entre outros – à criação de dicionários. O mesmo nota também Iriarte Sanromán (2001: 78):

> A linguística teórica (e a semântica em particular) não tem demonstrado interesse suficiente pelos problemas relativos à lexicografia e ao dicionário assim como pelos dados linguísticos que a prática lexicográfica e o dicionário podem revelar à linguística geral e que esta ignora por considerar irrelevantes do ponto de vista teórico.

Por outro lado, ele admite que muitos lexicógrafos não prestaram a devida atenção aos avanços das teorias linguísticas modernas, esforçando-se por aplicá-las aos seus dicionários (Ibidem):

> Tal situação não deverá melhorar até os lexicógrafos não aprenderem qualquer coisa sobre o que os teóricos dizem e até os teóricos não se familiarizarem mais com os dados pertinentes e reais relativos à língua e às necessidades dos utilizadores de dicionários.

Iriarte Sanromán também afirma que o dicionário como produto e como instrumento da investigação é subvalorizado e, consequentemente, o mesmo acontece com a própria investigação lexicográfica (Ibid.: 86).

> Linguística teórica nascida no século XIX teve de se afirmar como ciência e assim verificou-se a rejeição de actividades como a elaboração de dicionários que se consideravam menos ou pouco científicas dentro do paradigma estruturalista, pois poderiam pôr em dúvida a legitimidade da integração da linguística no grupo de tais "ciências".

Na Europa ocidental, estudos metalexicográficos faziam falta até aos anos 60 (Ibid.: 21) quando apareceram os primeiros estudos, *Problems in Lexicography*[60], actas do congresso realizado em Bloomington em 1960, e a tese de Bernard Quemada *Les Dictionnaires du français moderne, 1539-1863. Etude sur leur historie, leurs types et leurx méthodes*.

Quanto ao mundo lusófono, Barbosa[61] nota que a disciplina Lexicologia e Lexicografia existe na Universidade de São Paulo desde 1971, ou seja, desde então deve ter havido reflexões metalexicográficas no ensino superior brasileiro. Em Portugal, a lexicografia é representada sobretudo por Mário Vilela e acima mencionado Álvaro Iriarte Sanromán (2001).

## 4.2. Dicionários existentes de colocações e de combinações não livres

Segundo Iriarte Sanromán, o dicionário, nomeadamente os dicionários bilingues e os dicionários codificadores, deveria incorporar, em forma de unidades lexicográficas, não apenas palavras (*leite*) ou expressões idiomáticas (*leite e sangue, chorar sobre o leite derramado*), mas também combinações lexicais restritas (colocações) como *leite gordo*, e outras combinações de palavras que muitos poderiam considerar como sendo livres (*base de operações, década de sessenta, máximo comum divisor*) (cf. 2001: 147). Os mesmos princípios defende são defendidos por também František Čermák (1995) na tradição checa ou Sinclair (1991) na tradição anglo-saxónica.

Nas últimas décadas do século XIX, o desenvolvimento passou-se sobretudo nas línguas inglesa, francesa e alemã. Talvez graças à pragmática, que surge nos anos sessenta, relativamente muita atenção dos linguistas ingleses é prestada às colocações, ou seja, possibilidades não livres do lema. Assim surge *Oxford Collocations Dictionary for Students of English,* bem conhecido pelos estudantes de inglês do nível intermediário ou avançado. Quanto ao francês, mencionemos, por exemplo, o *Dictionnaire Explicatif et Combinatoire du Français Contemporain (*DEC*,* Mel'čuk *et al.* 1984-1999*)*, dicionário que «está pensado para fornecer toda a informação sobre "como dizer X" mais do que "o que

60 Householder, F. W. & Saporta: Problems in Lexicography. Bloomington, Indiana Universtity 1962.
61 Barbosa, M. A. O Grupo de Trabalho de Lexicologia, Lexicografia e Terminologia da ANPOLL: Formação e desenvolvimento. *Revista da ANPOLL*, v.1, p.53-60, 1995. p.55. Citado em *Breve histórico da metalexicografia no Brasil e dos dicionários gerais brasileiros.*

quer dizer X". [...] Preocupa-se com a descrição semântica exaustiva e rigorosa da cada lexema (*explicatif*) assim como com todas as possibilidades combinatórias não livres do lema (Ibid.: 26).» Até hoje elaboraram-se quatro volumes deste dicionário em francês. As funções lexicais[62] propostas por Mel'čuk e utilizadas no DEC foram também usadas no projecto em curso *Lexique actif du français* (LAF, Polguère 2000) bem como no DiCE, *Dicionario de colocaciones del español*. Existe também uma obra correspondente a DEC na língua russa, para além dos seus princípios lexicográficos terem sido verificados em línguas como o polaco, o inglês, o somali ou o alemão (Ibidem).

Na sua contribuição sobre o *Dicionario de colocaciones del español* (DiCE), Sanromán Vilas e Alonso Ramos (2007) afirmam que um dicionário de colocações deve satisfazer os aprendentes duma língua estrangeira bem como os falantes nativos. Os primeiros tiram proveito de tal dicionário dado que não é possível traduzir as colocações palavra por palavra (*to pay attention* ('pagar', ingl.), *prestar atención* ('prestar', esp.), *fare atenzione* ('fazer', it.), *kiinnitää humiota* ('fixar', fin.), *die Aufmerksamkeit schenken* ('dar (como uma prenda)', alemão) ou porque não há correspondências simétricas (*a piece of news* vs. *una noticia; to queue* vs. *hacer cola*). Os segundos, os falantes nativos (mas também os aprendentes duma língua estrangeira), podiam precisar de ter um acesso rápido à todas as maneiras de exprimir um certo sentido através de perspectivas diferentes dependentemente dos participantes na situação designada pela unidade lexical.

Quanto ao português, um dicionário de colocações (ou combinatórias como as chamam os autores) está disponível online[63]. Eis o comentário dos autores na página principal:

> Os resultados que se apresentam não têm como objectivo serem uma listagem de expressões fixas funcionalmente equivalentes a uma palavra, isto é, não foram aqui exclusivamente tratadas as locuções de várias categorias morfo-sintácticas, aforismos ou outras expressões fixas do português. Pelo contrário, interessou-nos recolher diversos

[62] As colocações são descritas por meio das funcões lexicais ('lexical functions', LF). A função lexical codifica a relação entre as duas unidades lexicais de maneira que uma deles, a base da colocação ('the base of the collocation'), controla a escolha lexical doutra ('the collocate'). Por exemplo, LF `Magn`[62] codifica a relação entre os seguintes pares adjectivo-substantivos: *medo mortal* 'great fear', *chuva torrencial* 'heavy rain' e *vontade de ferro* 'iron will'. Cada dos adjectivos é seleccionado pelo substantivo correspondente para exprimir o sentido 'intenso'.

[63] http://www.clul.ul.pt/sectores/linguistica_de_corpus/manual_combinatorias_online.php

tipos de expressões, com diversos tipos de composição e diversos graus de fixidez (desde a fixidez total encontrada, por exemplo, nas expressões aforísticas, até expressões semi-fixas e, ainda, expressões não composicionais e não lexicalizadas que apontam para associações lexicais preferenciais). Deste modo, o termo *combinatória* é, aqui, utilizado em sentido lato, na medida em que engloba diferentes tipos de associações de palavras.

O dicionário contém as seguintes expressões:

(i) expressões aforísticas que são geralmente consideradas como totalmente fixas, mas que no *corpus* podem ocorrer com grande variação (*no poupar é que está o ganho*; *no prejuízo é que está o ganho*; *no atacar é que está o ganho*; *no economizar é que está o ganho*; etc.);

(ii) expressões idiomáticas sintacticamente fixas, mas com possível variação flexional de um ou mais elementos (*esfregar as mãos de contentamento*; *esfregou as mãos de contentamento*; *esfregavam as mãos de contentamento*; *esfrega-se as mãos de contentamento*);

(iii) expressões idiomáticas ou composicionais que admitem alguma variação lexical (*onda/vaga/maré de assaltos*; *fazer/desenvolver/estabelecer/encetar contactos*);

(iv) casos de locuções prepositivas (*ao abrigo de*; *em consequência de*), conjuntivas (*a fim de que*; *da mesma forma que*), adverbiais (*por acaso*) ou, ainda, de regência verbal (*abdicar de*; *abater-se sobre*);

(v) associações lexicais preferenciais, i. e., semanticamente composicionais e sem fixidez sintáctica, mas com valores estatísticos combinatoriais muito significativos, bem como uma frequência muito alta, que podem indicar não só uma preferência associativa, mas também uma possível lexicalização progressiva do grupo de palavras (*instaurar um processo*; *cessar funções*; *erros e imprecisões*; *absolutamente indispensável*);

(vi) expressões sintagmáticas ou frásicas ao nível do discurso, com uso frequente (*francamente não sei*; *para ser franco*; *as consequências estão à vista*; *uma coisa garanto*) (Ibidem).

Os resultados são apresentados em vários ficheiros html, por ordem alfabética e cada linha do ficheiro contém informação sobre a colocação e o lema a partir do qual ela foi tratada. (Na linha, aparece em primeiro lugar o lema principal e depois o lema do grupo da colocação.) Ao cliquar sobre uma linha, abre-se uma janela com as concordâncias da colocação, isto é, com os contextos em que esta ocorreu no corpus.

Do ponto de vista informático (e, naturalmente, do ponto de vista de utilizador), este tratamento é inconveniente. O dicionário não consta duma base de dados explorável automaticamente. Com a vista de grande volume de dados, o dicionário está dividido em 17 partes (i. e. 17 páginas separadas), facto pelo que não é possível executar uma pesquisa rápida (Ctrl+F ou outros métodos). Algumas letras estão agrupadas em uma página (J-L, M-O etc.) mas as letras C, F e G estão divididas em várias partes sem que esteja escrito onde cada uma delas começa e acaba (por exemplo C1: cadeira – consentimento, C2: consequência – consumidor).

No que respeita ao checo, o projecto mais importante foi a criação do dicionário de fraseologismos, *Slovník české frazeologie a idiomatiky* (1983-2007). O dicionário tem quatro volumes:

(i) Parémias (Přirovnání) – expressões que contêm *como*.
(ii) Expressões não-verbais (Výrazy neslovesné) – frasemas nominais (sobretudo N-N, A-N), adverbiais e gramáticos (conjuncionais, preposicionais etc.).
(iii) Expressões verbais (Výrazy slovesné).
(iv) Expressões frasais (Výrazy větné) – proposições fixas ou seja frasemas em forma da frase já feita. Tradicionalmente trata-se se provérbios e frases-feitas, mas há expressões da vida quotidiana, citações populares etc.

Como salienta Čermák, não existe nenhum dicionário de frasemas proposicionais semelhante em nenhuma das línguas: este volume contém 9000-10 000 frasemas (todos os volumes contêm 25 000 frasemas).

Este dicionário, além da original metodologia e vasto material usados, contribuiu para o mundo linguístico com o descobrimento de que a nossa comunicação consiste em

mais prefabricados e clichés do que se propunha. Uma das peculiaridades características para este dicionário foi o facto de que como fonte serviu não apenas a linguagem escrita (dicionários e literatura) mas também a linguagem oral. Esta metodologia conduz a uma mais fácil identificação do frasema e define a frequência do seu uso activo, oferecendo exemplos baseados no seu uso real.

Além dos dicionários mencionados realce-se a importância do dicionário técnico checo-português e português-checo que envolve áreas de construção, electrotécnica, metalurgia, química e ciências naturais. Dado que se trata de material terminológico, há um grande número de colocações.

Quanto aos existentes dicionários bilingues checo-português, notámos que no maior dicionário checo-português (LEDA 1997) as colocações verbais são bem tratadas (ao termos em conta que se trata dum dicionário bilingue não especializado em colocações) mas as colocações nominais aparecem esporadicamente, a razão pela qual decidimos criar um dicionário de colocações nominais.

# 5. Criação de um dicionário de colocações checo-português

## 5.1 Métodos de detecção e extracção de colocações existentes

Existem vários métodos de detecção e extracção de colocações e não é o objectivo deste trabalho enumerá-los duma forma exaustiva; mencionaremos apenas os métodos mais importantes e relevantes. Em primeiro lugar é necessário definir se estamos à procura de colocações num texto monolingue ou se procuramos pares de colocações em duas línguas.

### 5.1.1 Colocações monolingues

**O método manual** é o método mais antigo e pouco eficiente mas mesmo assim utilizado até ao dia de hoje. Gradualmente está a ser substituído pela extracção automática mas apesar de existirem programas cada vez mais eficientes, o elemento humano melhora os resultados duma forma significante: frequentemente, as colocações são extraídas automaticamente e depois controladas manualmente.

**O método automático.** Há imensos métodos de extracção automática mas como nos interessa a tradução das colocações em checo para o português, mencionaremos primeiro o método mais eficiente utilizado para o checo. Este método baseia-se num *corpus* sintacticamente anotado (já mencionado em respeito a *Prague Dependency Treebanks*), neste particular caso um *corpus* jornalístico. Em primeiro lugar é preciso detectar os «candidatos»; aqui empregam-se as relações sintácticas. De seguida, as verdadeiras colocações são seleccionadas com a ajuda de vários métodos matemáticos e estatísticos. Os mais conhecidos são a frequência, *t test*, *Z score*, $x^2$ *test* e a proporção de probabilidades; no entanto Pecina (2006) utiliza mais de oitenta medidas e os melhores resultados atingem 80% de colocações reconhecidas.

Também Milena Uzeda Garrão e Maria Carmelita Dias (2002) usaram métodos automáticos com a finalidade de identificar as colocações verbais no português do Brasil. Tal como o nome (*The corpus never lies: a statistical approach for the identification of verbal collocations*) sugere e contrariamente à abordagem que combina a intuição humana

com a evidência adquirida num *corpus*, as autoras apresentam uma perspectiva baseada puramente em dados recuperados dum *corpus*, nomeadamente do ponto de vista estatístico. A necessidade da intuição humana vê-se como indesejada e as autoras também defendem que uma análise de um *corpus* fornece mais detalhes sobre o fenómeno.[64]

A metodologia, pela primeira vez testada nas frases verbais do português do Brasil, consiste em

(i) o emprego dum corpus robusto;

(ii) a aplicação de filtros Java;

(iii) a aplicação dum testo de probabilidade, nomeadamente o escore Log Likelihood[65].

O objectivo é distinguir as colocações das combinações não restringidas (*tomar parte* ou *tomar conta* vs. *tomar* + *substantivo* no sentido de 'apanhar'). Os resultados foram satisfatórios: detectaram-se 87.2 % de colocações correctamente formadas.

### 5.1.2 Pares bilingues

A detecção e extracção de pares bilingues de colocações causa maiores dificuldades devido ao facto de os métodos de extracção bilingue ainda não serem bem elaborados. Será necessário tirar proveito de ferramentas «tradicionais» aliás menos especializadas. Apresentaremos alguns métodos e ferramentas que nos facilitarão a procura duma colocação numa língua estrangeira e ferramentas que possibilitarão guardar pares de colocações para futuro uso. Concentrar-nos-emos na procura de tradução checo-português.

**Dicionários**. Excepto pelos casos de procura das colocações plenamente composicionais e das colocações-unidades lexicais, existe sempre uma probabilidade de encontrar a colocação inteira através da palavra-chave, ou seja através da base da colocação, ao consultarmos um dicionário monolingue português. Contudo, a questão da validade dos resultados obtidos é um problema grave. A segunda possibilidade é um dicionário bilingue: como já dissemos, as colocações verbais são quase bem tratadas no

[64] By choosing the alternative path, we not only avoid the time-consuming and controversial human intuitions for their assessment but also we get much richer lexicographic information on the phenomenon, such as its statistics, and its usual textual environment (2002: 354).

[65] Manning-Schutze: 1999.

recente bilingue dicionário checo-português mas é possível encontrar também algumas colocações nominais (*dopravní značka – sinal de trânsito*).

**Pesquisa em textos escritos em português, incluso os *corpora* monolingues**[66] traz problemas idênticos: não temos certeza que a colocação encontrada tem o sentido intencionado. Apesar de ser um bom instrumento de apoio à tradução, as limitações do texto fonte monolingue permanecem.

Uma ferramenta eficaz é um ***corpus* paralelo**[67]. Um *corpus* checo-português está a ser criado na Universidade Carolina no âmbito do projecto Intercorp mas contém apenas 16 livros literários. Os textos literários são mais adequados para os estudos da tradução literária do que para os fins lexicográficos, pois a tradução literária é um processo muito específico e notavelmente distinto da tradução técnica, ou seja não-literária. Além disso, sempre há problemas com os direitos de autor e não é possível ver o *corpus* inteiro; consultam-se apenas pequenos traços como resultados da pesquisa.

Pouco se sabe da possibilidade de criar um *corpus* paralelo dos materiais da DGT, Direcção-Geral da Tradução[68]. É possível fazer download da memória de tradução (TM – *translation memory*)[69], escolher o par de línguas desejado e alinhar o *corpus* frase a frase. O resultado é um ficheiro .tmx[70], i. e. uma TM que pode ser utilizada em Trados/SDXL ou outro software de apoio à tradução. Não obstante, este corpus é fechado e não possibilita alinhar pares de frases de textos tirados de outras fontes. Por isso resolvemos criar as

---

[66] Por exemplo http://www.corpusdoportugues.org/ para o português europeu ou ftp://ftp.liv.ac.uk/pub/linguistics/, *Corpus of Brazilian Media Portuguese.* Alternativamente, **BootCaT** (CaT ou CAT é abreviatura utilizada para designar as ferramentas de aspoio à tradução - *computer assisted translation)* possibilita criar o nosso próprio corpus que contém palavras inseridas na pesquisa. Uma versão online, i. e. sem instalação, **WebBootCaT**, foi criada na Faculdade de Informática de Universidade Masaryk em Brno.

[67] Um *corpus* paralelo é uma ferramenta útil para tradução em geral e para os estudos de tradução. Veja-se o comentário na página inicial de COMPARA, um corpus paralelo bidireccional de português e inglês: «O COMPARA é uma ferramenta que permite estudar a tradução humana e contrastar o português e o inglês através de pesquisas automáticas. Por exemplo, se inserirmos uma palavra em português, podemos ver como essa palavra foi traduzida para inglês em diferentes contextos.»

[68] Veja-se mais em http://langtech.jrc.it/DGT-TM.html#DGT-TM: "This extraction of aligned sentences can be used to produce a parallel multilingual corpus of the legislative documents (*Acquis Communautaire*) of the European Union in 22 EU languages. The aligned sentences ("*translation units*") have been provided by the Directorate-General for Translation of the European Commission by extraction from one of its large shared translation memories in *Euramis* (*European advanced multilingual information system*). This memory contains most, although not all, of the documents of the Acquis Communautaire, as well as some other documents which are not part of the Acquis."

[69] http://wt.jrc.it/lt/Acquis/DGT_TU_1.0/data/

[70] É também possível modificar o sufixo .tmx obtendo um arquivo texto .txt ou .doc.

nossas próprias ferramentas e o nosso próprio *corpu*s paralelo, mencionado em detalhe em 5.2.

O **2lingual**[71] é uma ferramenta recente, lançada em 2008. O 2lingual baseia-se no Google (o Google próprio é um *corpus* sui generi) e no Google Translate e oferece pesquisa simultânea em duas línguas (de 42 línguas em total). Devido aos volumes de textos em inglês disponíveis, os melhores resultados são obtidos se uma das línguas consultadas é o inglês. O programa toma em conta sequências de palavras, não apenas palavras únicas. Vejamos como o 2lingual lida por exemplo com *natural language processing* em tradução para o português e checo; é interessante observar os resultados parciais e o resultado final:

*natural – natural - přirozený*

*natural langugage – linguagem natural – přirozený jazyk*

*natural language processing – processamento da linguagem natural – zpracování přirozeného jazyka*

Como já dissemos, os resultados da pesquisa entre o checo e o português nunca serão tão satisfatórios como se uma das línguas fosse o inglês mas mesmo assim é possível encontrar várias colocações (*daňový poplatník – contribuinte, slovní hříčka - trocadilho*). Para além disso, apesar de os resultados correctos não serem exibidos imediatamente (*státní dluh* – apenas aparece *dívida*), às vezes é possível encontrar os resultados nos textos encontrados (*dívida pública*).

Dado que o programa tira a sequência inteira de textos disponíveis (da mesma maneira funciona também o Google Translate), podemos afirmar que é uma ferramenta eficaz para a tradução de colocações. Infelizmente, nem sempre os resultados são satisfatórios (*státní rozpočet – Membro orçamento* em vez de *Orçamento do Estado*). Para além disso, é necessário ter em conta a validade relativamente baixa dos resultados: ainda que a pesquisa se restrinja ao domínio .pt (para eliminar ocorrências das formas do português do Brasil), sempre será necessário verificar os resultados; 2lingual tem apenas uma função orientadora.

[71] www.2lingual.com

O último método a mencionar será a **cooperação com outros tradutores**, sobretudo na área de terminologia técnica. Há vários fóruns de discussão ou **glossários** disponíveis (por exemplo no servidor de tradutores e traduções ProZ[72]).

## 5.2 Métodos utilizados na criação do dicionário de colocações checo-português

É necessário salientar que o objectivo do presente trabalho não foi a criação de um dicionário de colocações mas sim recolhimento da base teórica e metodológica para tal dicionário que se espera vir a ser criado. Não obstante, decidimos criar um pequeno dicionário-pioneiro para apoiar a demonstração da importância das colocações e para encontrar e descrever problemas que a criação dum tal dicionário traz. O dicionário contém 150 pares de colocações e os métodos utilizados na criação seguem:

**O método manual**, conveniente se pretendermos extrair relativamente poucos pares. As colocações incluídas no dicionário foram tiradas de jornais, revistas e artigos online e verificadas noutros contextos ou dicionários monolingues portugueses para assegurar que a tradução esteja correcta.

Algumas colocações foram também extraídas do ***corpus* paralelo** checo-português que criámos com a ajuda de Ondřej Bojar de ÚFAL*, Instituto da linguística formal e aplicada*. Como base serviram os *corpora* tirados de *Acquis* da União Europeia[73]. Primeiro, era preciso segmentar, *tokenizar*[74] e *etiquetar*[75] o *corpus* checo bem como o português: há programas disponíveis online mas era necessário fazer algumas alterações e utilizar os nossos próprios dados de treinamento para que os programas funcionassem correctamente com o tipo de documentos que tivemos à nossa disposição. A seguir, alinhámos os dois *corpora* frase por frase, criando assim um *corpus* paralelo. O *corpus* estará disponível online em http://ufal.mff.cuni.cz/umc/ (UMC - ÚFAL Multilingual Corpora) onde serão publicados também os detalhes sobre a sua criação e sobre os experimentos com a extracção automática.

[72] www.proz.com
[73] http://wt.jrc.it/lt/Acquis/
[74] *Tokenização* é uma segmentação mais refinada.
[75] *Etiquetador* (*tagger* em inglês) é uma ferramenta que coloca etiquetas morfossintáticas em cada palavra de um *corpus*. O etiquetador usado foi o *Etiquetador Tree-Tagger* disponível em http://www2.lael.pucsp.br/corpora/etiquetagem/index.html.

Tentámos extrair as colocações automaticamente (por meios de alinhamento das palavras, não apenas das frases), mas os resultados não foram satisfatórios; o elemento humano ainda é necessário. As colocações foram então extraídas manualmente. O método da extracção de um corpus paralelo é eficiente mas mesmo assim encarámos certas dificuldades. Em primeiro lugar, apesar de a tradução técnica ser mais fiel do que a tradução literária e de ser do dever dos tradutores não literários traduzir conceitos e termos sem alterações no significado, nem sempre é possível traduzir uma expressão exactamente. Este problema é ainda ampliado pelo facto de os textos incluídos no *corpus* não serem traduções directas entre o checo e o português; a maioria deles são traduções de inglês ou francês para o checo e o português. Em teoria, este facto não devia ser relevante mas acreditamos que algumas finezas do significado se podem «perder» no processo da tradução, especialmente se não houver equivalentes exactos em inglês num lado e checo ou português noutro lado.

Assim, não incluímos traduções disputáveis como as nos seguintes exemplos:

4. Z účasti na základním kapitálu nevyplývá ani hlasovací právo, ani právo na dividendy nebo na úroky. Právo na úhradu jmenovité hodnoty splaceného podílu na základním kapitálu vzniká pouze v případě likvidace Agentury.

4. A participação no capital não confere nem direito de voto nem direito a dividendos ou a um juro. Dá direito ao reembolso do montante nominal das fracções de capital transferidas, unicamente no caso de dissolução da Agência.

2. Agentura požívá v každém členském státě nejširší právní způsobilost přiznávanou právnickým osobám daným vnitrostátním právem. Může zejména nabývat a zcizovat movitý i nemovitý majetek, uzavírat smlouvy, poskytovat věcné nebo osobní záruky, jednat jako zprostředkovatel, oprávněný zástupce nebo jednatel, vystupovat před soudem, účastnit se rozhodčích řízení, jakož i provádět obchodní transakce a přijímat nařízení nezbytná pro plnění svých úkolů.

2. A Agência goza, em todos os Estados-membros, da capacidade jurídica mais lata reconhecida às pessoas colectivas de direito público e privado. Pode, nomeadamente, acquirir ou alienar bens móveis e imóveis, concluir todos os contratos, dar o consentimento a todas as garantias reais ou pessoais, agir como corretor, mandatário ou comissário, intentar acções judiciais, obrigar-se, transigir e proceder a todos os actos comerciais bem como a todas as regulamentações necessárias para o comprimento das suas missões.

## 5.3 Versão electrónica do dicionário de colocações checo-português

O dicionário-pioneiro, além de ser incluído em anexo deste trabalho, tem também uma versão electrónica, baseada em PHP+MySQL e disponível online[76]. As expressões são guardadas numa base de dados e podem ser recuperadas através da pesquisa duma colocação em checo ou em português. Neste momento é possível pesquisar colocações no sentido checo-português bem como português-checo mas é preciso ter em conta que as expressões incluídas são apenas aquelas que podem ser designadas como colocações em checo (podem então ter um equivalente português monoverbal). Os visitantes podem também inserir próprios pares de colocações. Estes guardam-se numa base de dados separada e depois de serem verificados serão adicionados à base de dados activa.

A forma básica das expressões é a forma singular (com a excepção de poucas expressões que ocorrem apenas ou frequentemente em plural). As variantes ortográficas (*směnný kurz/směnný kurs – taxa de câmbio*) foram tomadas em conta: cada variante tem a sua própria entrada; caso contrário não seria possível pesquisar bidirecionalmente.

Contudo, a pesquisa está longe de ser óptima e ainda há muito por melhorar. Neste momento apenas é possível pesquisar formas exactas: convinha possibilitar a pesquisa de expressões parecidas e também a pesquisa sem diacríticos (*smenny kurz* ou *taxa de cambio* neste momento não seriam reconhecidos). Outro melhoramento por considerar seria a explicitação que é possível inserir apenas uma expressão que queremos traduzir mas que não está na base de dados. A lista dessas expressões estaria disponível: existe sempre a possibilidade de que outro visitante conheça a tradução e possa inseri-la.

Quando o novo acordo ortográfico entrar em vigor, será também necessário implementá-lo no dicionário.

É possível que algumas das colocações incluídas no dicionário possam ser encontradas também no dicionário checo-português mas seria inconveniente incluir apenas aquelas que o dicionário não contém.

[76] http://kolokace.wz.cz/kolokace.php

## 5.4 Trabalho futuro

Naturalmente, o objectivo do nosso esforço é criar um dicionário de colocações checo-português. Para atingi-lo, será necessário elaborar os métodos da extracção automática de corpora bilingues, seja totalmente automática seja parcialmente automática – e.g. a extracção automática duma colocação em checo e a procura manual da tradução para o português num *corpus* paralelo. As ferramentas inventadas serão úteis – talvez após algumas modificações – também para outros pares de línguas. Neste momento estamos a verificar as possibilidades de colaboração com o *Centro do processamento da linguagem natural* na Faculdade de Informática da Universidade Masaryk em Brno.

No que respeita a Letras, consideramos conveniente que os estudantes aprendam usar as ferramentas mais úteis, pois elas são indispensáveis não apenas na tradução de colocações mas na tradução em geral. Hoje em dia é difícil produzir traduções efectivas sem utilizar ferramentas avançadas e ferramentas de apoio à tradução (Trados/SDXL, Wordfast, StarTransit etc.). Como as faculdades de Letras frequentemente servem como a única possibilidade de obter qualificação em línguas estrangeiras e os finalistas supostamente devem ser capazes de trabalhar como tradutores, convinha que o conhecimento dos avanços técnicos seja mais divulgado.

É possível disponibilizar a lista das ferramentas para os estudantes (e.g. em ELF, sistema de e-learning) ou online na página do dicionário de colocações. Os próprios estudantes também podem participar na criação do dicionário, inserindo expressões que eles encontraram. Se no futuro houver um seminário de tradução técnica, não literária[77], termos altamente úteis poderiam ser inseridos no dicionário e guardados para futuro uso.

[77] Há pouca literatura lusófona por traduzir para o checo, então consideramos mais útil ensinar a tradução não literária. Dela tirarão proveito não apenas futuros tradutores mas todos os que trabalharão para uma empresa lusófona (ou multinacional) e terão que produzir textos em português.

# 6. Conclusão

As colocações são um fenómeno linguístico que, apesar de ter capturado a atenção de numerosos autores, permanece sem ser descrito de uma forma exaustiva e satisfatória. Vários autores propõem várias classificações, algumas sendo até contraditórias.

Não obstante, resolvemos dedicar-nos a esta problemática, sabendo que as colocações têm um papel importante na aquisição e processamento de linguagem, ensino de línguas, tradução, linguística computacional e lexicografia. Devido ao facto de as colocações causarem problemas sobretudo no momento de um falante não nativo produzir texto numa língua estrangeira, é necessário incluir as colocações em dicionários bilingues. Como o dicionário existente de checo-português não contém muitas colocações nominais, decidimos criar um dicionário que as contenha.

Além da demonstração da importância das colocações, o presente trabalho teve mais dois objectivos: criar a base teórica para o dicionário de colocações checo-português e esboçar a metodologia de criação de tal dicionário.

No que respeita à base teórica, examinámos vários conceitos e classificações apresentados por autores anglo-saxónicos, portugueses e checos. Basicamente, há três abordagens diferentes: a abordagem fraseológica considera as colocações como um tipo particular de unidade fraseológica. A abordagem baseada em frequências trata as colocações como um fenómeno evidenciado pela estatística, definindo-as como uma combinação frequente e relativamente fixa de palavras. Esta abordagem é apoiada pela linguística de corpus. A abordagem semântico-sintáctica define o fenómeno da colocação como uma relação abstracta entre as palavras, destacando as possibilidades combinatórias existentes. Esta abordagem abrange também as colocações puramente gramaticais, às vezes designadas como coligações.

A divulgação do termo ‘colocação’ (ou ‘combinatória’) nas linguísticas portuguesa e checa começou com atraso comparando com a linguística anglo-saxónica, a razão pela qual tivemos que recorrer também aos estudos fraseológicos.

As conclusões da parte teórica deveram servir-nos como a base para a criação do dicionário de colocações checo-português, i. e. a teoria deveria ajudar-nos a definir o que é uma colocação e quais os tipos de colocações que serão incluídos no dicionário. Devido a

uma certa discordância nas classificações decidimos enumerar os tipos duma forma taxativa.

A seguir, procedemos à criação do dicionário. Na procura da metodologia conveniente encontrámos três problemas principais: primeiro, a inexistência de um *corpus* paralelo checo-português que contenha textos de várias áreas (jornalística, científica), não apenas textos provenientes dos materiais da União Europeia. Resolvemos criar um *corpus* paralelo e apesar de este incluir – neste momento – apenas materiais da União Europeia (por a UE oferecer *corpora* monolingues já feitos), as ferramentas criadas durante o processo poderão ser utilizadas no futuro para adicionar outros tipos de textos.

O segundo problema enfrentado era a precisão insuficiente da extracção automática de colocações, sendo esta imprescindível para a criação dum dicionário de maior volume. O desenvolvimento de tais métodos e a elaboração de tais ferramentas será um desafio para a linguística computacional. Não obstante, acreditamos que devido ao progresso rápido nesta área será possível extrair automaticamente pares bilingues dentro de poucos anos. Devido à impossibilidade de extrair as colocações automaticamente, usámos o método manual, extraindo as colocações de textos escritos em português e do *corpus* paralelo checo-português.

O terceiro problema era a improbidade dos resultados. Apesar de termos verificado o significado das colocações encontradas em textos escritos em português, sempre pode haver uma interpretação incorrecta. Semelhantemente, as traduções dos materiais da União Europeia, apesar de terem sido produzidas por tradutores profissionais, também podem conter erros.

O último maior problema a mencionar será a impossibilidade de enumerar todas as possibilidades de tradução. Incluímos apenas as traduções encontradas nos textos e verificadas mas um falante não nativo não tem a competência para decidir quais todas as formas possíveis de uma colocação. Para que o dicionário possa ser vinculativo, será necessário colaborar com falantes nativos.

# Bibliografia

Alm-Arvius, Ch. *Fixed, flexible or fragmentary? Types of idiom variation. In Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes*. Joensuu: Joensuu University Press, 2007, pp. 14-26.

Alonso Ramos, M. *Las Funciones Léxicas en el Modelo Lexicográfico de I. Mel'chuk*. Madrid: UNED, 1993.

Antunović, G. *Croatian translators' take on Swedish collocations and idioms*. *In Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes*. Joensuu: Joensuu University Press, 2007, pp. 27-40.

Bacelar do Nascimento, M. F., L. A. S. Pereira. *Dicionário de Combinatórias do Português: Associações lexicais frequentes observadas num corpus de Português contemporâneo*. *In* Actas do XI Encontro Nacional da Associação Portuguesa de Linguística, Vol. II: Dicionários, 1996, pp. 43-54.

Bäcklund, Ulf. *Restrictive Adjective-Noun Collocations in English*. Umea: Acta Universitatis Umensis, 1981.

Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P.: *WebBootCaT: instant domain-specific corpora to support human translators*. Proceedings of EAMT 2006, Oslo, 2006, pp. 247-252.

Bartsch, S. *Structural and functional properties of collocations in English: a corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Tübingen: Gunter Narr Verlag, 2004.

Becker, J. *The phrasal lexicon*. Proceedings of the 1975 workshop on Theoretical issues in natural language processing, Cambridge, pp. 60-63. Em Nattinger-DeCarrico, 1992.

Carvalho, J. Herculano de. *Teoria da Linguagem. Natureza do Fenómeno Linguístico e a Análise das Línguas*. Coimbra: Atlântida, 1979.

Čermák, F., M. Šulc. *Kolokace*. Praha: Nakladatelství Lidové noviny, 2006.

Čermák, F. *Frazeologie a idiomatika česká a obecná. Czech and General Phraseology*. Praha: Karolinum, 2007.

Čermák, F., R. Blatná. *Manuál lexikografie.* Praha: Nakladatelství H&H, 1995.

Chomsky, N. *Aspect of the theory of syntax.* Cambridge: Mass, 1965. Em Vilela, 1994.

Coseriu, 1979. *Teoria da Linguagem e Lingüística Geral. Cinco Estudos*. Rio de Janeiro: Presença/Editora da Universidade de São Paulo, 1979. Em Iriarte Sanromán, 2001.

Eskildsen, S.W., T. Cadierno. *Are recurring multi-word expressions really syntactic freezes? Second language acquisition from the perspective of Usage-Based Linguistics. In Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes*. Joensuu: Joensuu University Press, 2007, pp. 86-99.

Filipec, J., Čermák, F. *Česká lexikologie*. Praha: Academia, 1985.

Firth, J. R. *Papers in linguistics*, 1934-1951. London : Oxford University Press, 1957.

Greenbaum, S. *Verb-Intensifier Collocations in English. An Experimental Approach*. The Hague: Mouton, 1970.

Gledhill, C. *Collocations in Science Writing*. Tübingen: Narr, 2000.

Hakuta, K. *Prefabricated patterns and the emergence of structure in second language aquisition.* Language Learning 24: 287-97, 1974. Em Nattinger-DeCarrico, 1992.

Haensch, G., L.Wolf , S.Ettinger, R.Werner. *La lexicografía*. Madrid: Editorial Gredos, 1982.

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

Krashen, S. D. and R. Scarcella: *On Routines and Patterns in language acquisition and performance*. In Language Learning 28:283-300, 1978. Em Nattinger-DeCarrico, 2002.

Lewis, M. *The Lexical Approach*. The state of ELT and a Way forward. London: Language Teaching Publications, 1993.

Lyons, J. *Linguistic Semantic.* Cambridge: Cambridge University Press, 1995.

Malmkjær, K. *The linguistics encyclopedia*. London: Routledge, 2002.

Manning, Christopher D. and Hinrich Schütze. *Foundations of statistical natural language processing.* Cambridge, Mass: MIT Press, 1999.

Mathesius, V. *A functional Analysis of Present Day English on a General Linguistic Basis*. Prague: Academia, 1975.

Mel'čuk 1995. *Phrasemes in Language and Phraseology in Linguistics*. In Everaert (ed.) (1995), pp. 167-232. Em Iriarte Sanromán, 2001.

Mitchell, T.F. *Principles of Firthian linguistics*. London: Longman, 1975.

Nattinger, J. R, J. S. DeCarrico. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press, 1992.

Nesselhauf, N. *Collocations in a learner corpus*. Amsterdam: John Benjamins, 2005. Em Eskildsen-Cadierno, 2002.

Nesselhauf, N. 2004. *What are collocations*? In D. J. Allerton et allii (eds.). Phraseological Units: basic concepts and their application. International Cooper Series in English Language and Literature. vol 8. Schwabe Verlag Basel. Switzerland. pp 1-21. Em Antunes, 2007.

Palmer, E. H. *A Grammar of English Words: One thousand English words and their pronunciation, together with information concerning the several meanings of each word, its inflections and derivatives, and the collocations and phrases into which it enters*. London: Longmans, Green, 1938. Em Bartsch, 2002.

Pecina, P., M. Holub. *Sémanticky signifikantní kolokace. Automatická detekce kolokací v českém textovém korpusu*. ÚFAL/CKL Technical Report TR-2002-13.

Pecina, P., P. Schlesinger. *Combining Association Measures for Collocational Extraction*. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, July 2006, pp. 651–658.

Pottier, B., A. Audubert, C. T. Pais: *Estruturas lingüísticas do português*. São Paulo: Difel, 1975.

Pottier, B. *Grammaire de lespagnol*. Paris: Presses Universitaires de France, 1972. Em *O tratamento das lexias compostas e complexas*.

---. *Linguística geral. Teoria e descrição*. Rio de Janeiro: Presença, 1978. Em Iriarte Sanromán, 2001.

Przywara, Č. *Metody extrakce víceslovných výrazů z textu*. Bakalářská práce, Univerzita Karlova v Praze, 2008.

Sanromán Vilas, B., M. Alonso Ramos. *Collocation dictionary as an elaborate pedagogical tool for Spanish as a foreign language*. *Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes*. Joensuu: Joensuu University Press, 2007, pp. 282-296.

Stubbs, M. *Notes on the history of corpus linguistics and empirial semantics*. *In Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes*. Joensuu: Joensuu University Press, 2007, pp. 317-329.

Tomasello, Michael. *Constructing a language*. Cambridge: Cambridge University Press, 2000. Em *Collocations and Idioms 1*, 2002.

Uzeda-Garrão, M., M. Carmelita Dias. *The corpus never lies: a statistical approach for the identification of verbal collocations. In Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes*. Joensuu: Joensuu University Press, 2007, pp. 354-362.

Vilela, M. *Estudos de Lexicologia do Português*. Coimbra: Livraria Almedina, 1994.

-- -. *Léxico e gramática*. Coimbra: Livraria Almedina, 1995.

Wilensky, R., Y. Arens, D. Chin. *Talking to UNIX in English: an overview of UC*. Communications of the ACM 27:574-93, 1984. Em Nattinger-DeCarrico, 1992.

Wood, M. *A Definition of Idiom*. Manchester: Centre for Computational Linguistics, University of Manchester. Reprinted by the Indiana University Linguistics Club, 1986. Em Nattinger-DeCarrico, 1992.

Wong-Fillmore, L. *The Second Time Around: Cognitive and Social Strategies in Second Language Acquisition*. Unpublished doctoral dissertation, Stanford University, 1976. Em Nattinger-DeCarrico, 1992.

Zernick, U., M. Dyer. *The self-extending phrasal lexicon*. Computational Linguistics 13:308-27, 1987. Em Nattinger-DeCarrico, 1992.

## Internet

Aarts, B., G. Alderman. Sidney Greenbaum [online]. 2008 [consultado 16/3/2009]. <https://www.ucl.ac.uk/english-usage/about/greenbaum.htm>

An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications [online]. 2002 [consultado 13/3/2009]. <http://www.clul.ul.pt/equipa/lpereira/euralex_2002_pereira_mendes.pdf>

Antunes, S. Combinatórias do Português: uma abordagem corpus-driven com fins lexicográficos [online]. 2007 [consultado 13/3/2009]. <http://www.clul.ul.pt/artigos/antunes_sandra.pdf>

Breve histórico da metalexicografia no Brasil e dos dicionários gerais brasileiros [online]. 2005 [consultado 15/2/2009]. <http://www.unb.br/il/let/welker/metalex_Matraga>

Collocation [online]. [consultado 15/2/2009]. <http://esl.fis.edu/grammar/easy/colloc.htm>

COMBINA-PT: A Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions [online]. [consultado 13/3/2009]. <http://www.clul.ul.pt/equipa/lpereira/combina_lrec2006.pdf>

COMBINA-PT - Combinatórias Lexicais do Português [online]. 2004 [consultado 13/3/2009].
<http://www.clul.ul.pt/sectores/linguistica_de_corpus/projecto_combina.php>

COMBINATÓRIAS LEXICAIS DO PORTUGUÊS. Manual do Utilizador [online]. [consultado 13/3/2009].
<http://www.clul.ul.pt/sectores/linguistica_de_corpus/manual_combinatorias_online.php>

COMPARA [online]. 2008 [consultado 13/3/2009].
<http://www.linguateca.pt/COMPARA/compara_pt.html>

Etiquetador Tree-Tagger [online]. 2007 [consultado 15/4/2009].
<http://www2.lael.pucsp.br/corpora/etiquetagem/index.html>

John McHardy Sinclair [online]. 2009 [consultado 17/3/2009].
<http://en.wikipedia.org/wiki/John_McHardy_Sinclair>

Martins, E. S. O tratamento das lexias compostas e complexas [online].
[consultado 28/3/2009].
<http://www.gelne.ufc.br/revista_ano4_no2_15.pdf>

Michael Halliday [online]. 2008 [consultado 16/3/2009].
<http://en.wikipedia.org/wiki/Michael_Halliday>

Mineiro, A. Que tipo de combinatória lexical no discurso da Náutica? Um estudo baseado num corpus linguístico [online]. 2005 [consultado 13/3/2009].
<http://www.iltec.pt/pdf/wpapers/2005-amineiro-maritima.pdf>

Moraeus, H. F. R. Um estudo Contrastivo de colocações adverbiais (inglês português) sob o enfoque da lingüística de corpus [online]. 2005 [consultado 13/3/2009]. <http://www.fflch.usp.br/dlm/comet/artigos/Helmara_SILEL2006.pdf>

Pecina, P. A Machine Learning Approach to Multiword Expression Extraction [online]. 2008 [consultado 13/4/2009]. <http://ufal.mff.cuni.cz/~pecina/publications/mwe-2008-shared-task.pdf>

Pecina, P. Reference Data for Czech Collocation Extraction [online]. 2008 [consultado 13/4/2009]. <http://ufal.mff.cuni.cz/~pecina/publications/mwe-2008-resource.pdf>

Pereira, L.A., M. F. Bacelar do Nascimento. Dicionário de Combinatórias do Português do Centro de Linguística da Universidade de Lisboa [online]. [consultado 13/3/2009]. <http://www.clul.ul.pt/equipa/fbacelar/berlim_2000_nascimento_pereira.pdf>

Ranchhod, Elisabete Marques. O lugar das expressões fixas na gramática do português [online]. 2002 [consultado 13/3/2009]. <http://label.ist.utl.pt/publications/docs/LEFnGP.pdf>

T. S. Mitchell [online]. [consultado 13/3/2009]. <http://www.yek.me.uk/mitchell.html>

The DGT Multilingual Translation Memory of the Acquis Communautaire: DGT-TM [online]. 2009 [consultado 13/4/2009]. <http://langtech.jrc.it/DGT-TM.html>

The Prague Dependency Treebank [online]. [consultado 3/4/2009]. <http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/Doc/whatis.html>

Valente, R. Diferenças e similaridades colocacionais entre o português brasileiro e o português europeu. Estudo baseado na noção de função lexical da teoria sentido texto [online]. [consultado 25/3/2009].
<http://www.filologia.org.br/anais/anais%20iv/civ07_9.htm>

Valente, R. Por que razão a combinatória lexical não pode ser um princípio para se considerar verbos e adjetivos como sendo especializados? [online]. 2005 [consultado 13/3/2009].
<http://www.iltec.pt/pdf/wpapers/2005-rvalente-maritima.pdf>

Valente, R. S., I. O. Tchobánova, M. S. Ferreira. Problemas de tratamento e sistematicidade na compilação das combinatórias lexicais no Novo Aurélio Século XXI [online]. 2004 [consultado 13/3/2009].
<http://www.iltec.pt/pdf/wpapers/2004-rvalente-combinatorias.pdf>

# Apêndice 1: Lista das colocações

| | |
|---|---|
| akustická kytara | guitarra acústica |
| adoptivní rodina | família adoptiva |
| atomová bomba | bomba atómica |
| autorská práva | direitos de autor |
| autorské právo | direito de autor |
| bankovní účet | conta bancária |
| bílá magie | magia branca |
| bílé maso | carne branca |
| bílé víno | vinho branco |
| bílé víno | uva branca |
| bleší trh | feira de ladra |
| bojová umění | artes marciais |
| bojové umění | arte marcial |
| cela předběžného zadržení | prisão preventiva |
| celní unie | união aduaneira |
| centrální plánování | planificação central |
| cukrová vata | algodão doce |
| čaj o páté | chá das cinco |
| černá díra | burraco negro |
| černá listina | lista negra |
| černá magie | magia negra |
| černý trh | mercado negro |
| červené víno | uva preta |
| červené víno | vinho tinto |
| členský stát | estado-membro |
| dálkové světlo | máxima |
| daňový poplatník | contribuinte |
| diskusní fórum | fórum de discussão |
| dlažební kostka | paralelepípedo |
| domácí mazlíček | animal de estimação |
| domácí násilí | violência doméstica |
| doporučený dopis | carta registrada |
| dopravní značka | placa de sinalização |
| dopravní značka | sinal de trânsito |
| drahá polovička | cara-metade |

| | |
|---|---|
| duševní zdraví | saúde mental |
| elektrická kytara | guitarra elétrica |
| emisní banka | banco-emissor |
| evropský komisař | comissário europeu |
| finanční poradce | assessor pessoal |
| finanční trh | mercado financeiro |
| fyzické vyčerpání | esgotamento físico |
| generální ředitel | Director-Geral |
| generální tajemník | secretário-geral |
| geneticky modifikovaný | geneticamente modificado |
| geneticky modifikovaný organismus | transgénico |
| geneticky upravený | geneticamente modificado |
| geneticky upravený organismus | transgénico |
| globální oteplování | aquecimento global |
| havarijní plán | plano de emergência |
| hlasovací právo | direito de voto |
| horní hranice lesa | linha das árvores |
| hostitelská rodina | família de acolhimento |
| hromadná doprava | transportes públicos |
| humanitní disciplína | disciplina humanística |
| invalidní důchod | pensão de deficiência |
| jaderná hlavice | ogiva nuclear |
| jaderná válka | guerra nuclear |
| jaderná zkouška | teste nuclear |
| jednosměrná ulice | estrada de sentido único |
| jednotný trh | unicidade de mercado |
| kabelová televize | televisão de cabo |
| kandované ovoce | fruta de calda |
| kapesní nožík | canivete |
| kapesní nůž | canivete |
| kapitálový trh | mercado financeiro |
| klimatické změny | alterações climáticas |
| kolečkové křeslo | cadeira de rodas |
| koncentrační tábor | campo de concentração |
| krevní tlak | pressão arterial |
| krutý útok | feroz ataque |
| kupní síla | poder de compra |
| kvalifikovaná většina | maioria qualificada |

| | |
|---|---|
| lesní plod | fruto silvestre |
| lesní požár | fogo florestal |
| magnetické pole | campo magnético |
| míra inflace | taxa de inflação |
| míra nezaměstnanosti | taxa de desemprego |
| movitý majetek | bens móveis |
| mrtvý jazyk | língua morta |
| nemovitý majetek | bens imóveis |
| neochvějná jistota | certeza inabalável |
| nízkotučné mléko | leite magro |
| Nobelova cena | prémio Nobel |
| období dešťů | estação das chuvas |
| obecní úřad | junta de freguesia |
| odtahová služba | reboque |
| odtučněné mléko | leite desnatado |
| ochrana spotřebitele | defesa de consumidor |
| olej na smažení | óleo de fritura |
| olivový olej | azeite |
| opční právo | direito de opção |
| operační systém | sistema operativo |
| osobní záruka | garantia pessoal |
| ostnatý drát | arame farpado |
| ostrá kritika | dura crítica |
| otisk prstu | impressão digital |
| ozonová díra | buraco do ozono |
| ozónová díra | buraco do ozono |
| ozonová vrstva | camada de ozono |
| ozónová vrstva | camada de ozono |
| padací most | ponte levadiça |
| paměťová karta | cartão de memória |
| peněžní trh | mercado monetário |
| plná čára | linha contínua |
| plná zaměstnanost | pleno emprego |
| podmíněný reflex | condicionamento clássico |
| podmíněný reflex | condicionamento pavloviano |
| podmíněný reflex | condicionamento respondente |
| podpora v nezaměstnanosti | subsídio de desemprego |
| pojízdné lůžko | maca |

| | |
|---|---|
| policejní stanice | esquadra |
| polotučné mléko | leite meio gordo |
| poradní výbor | comité consultivo |
| poruchy příjmu potravy | distúrbios alimentares |
| poruchy zažívání | distúrbios digestivos |
| poslední pomazání | último abencerragem |
| poslední večeře | Última Ceia |
| povinná školní docházka | escolaridade obrigatória |
| pracovní trh | mercado de trabalho |
| právní předpis | disposição legal |
| právní způsobilost | capacidade jurídica |
| prezidentské volby | eleição para a presidência |
| prezidentské volby | eleição presidencial |
| předčasný důchod | pensão antecipada |
| přednost v jízdě | prioridade de passagem |
| přenesený význam | sentido figurativo |
| přerušovaná čára | linha descontínua |
| přímý příbuzný | familiar directo |
| příruční zavazadlo | bagagem de mão |
| ptačí chřipka | gripe das aves |
| pupeční šňůra | cordão |
| Rada Evropy | Conselho Europeu |
| rizikové chování | comportamento de risco |
| rodinný krb | lar conjugal |
| rodinný kruh | seio da família |
| rodinný kruh | seio familiar |
| rodné  město | cidade natal |
| rodný list | certidão de nascimento |
| rostlinný olej | óleo vegetal |
| rovné zacházení | igualdade de tratamento |
| rozený vůdce | líder nato |
| roztroušená skleróza | esclerose múltipla |
| rudá tvář | pele-vermelha |
| růžové víno | vinho rosé |
| skleníkový efekt | efeito da estufa |
| slovní hříčka | trocadilho |
| služební cesta | viagem de negócios |
| směnný kurs | taxa de câmbio |

| | |
|---|---|
| směnný kurz | taxa de câmbio |
| snubní prsten | aliança |
| sociální demokracie | social-democracia |
| sociální zabezpečení | segurança social |
| spínací špendlík | alfinete |
| státní deficit | défice público |
| státní dluh | dívida pública |
| státní hymna | hino nacional |
| státní rozpočet | Orçamento de Estado |
| státní rozpočet | Orçamento do Estado |
| stopy krve | rasto de sangue |
| střed zájmu | foco das atenções |
| studená válka | Guerra fria |
| svalová hmota | massa muscular |
| svěcená voda | água benta |
| Svědci Jehovovi | Testemunhas de Jeová |
| světelný rok | ano-luz |
| světelný signál | sinal luminoso |
| telefonní odposlech | escuta telefónica |
| tepelná elektrárna | central termoeléctrica |
| textový soubor | arquivo texto |
| trávicí ústrojí | aparelho digestivo |
| trest smrti | pena capital |
| tučné mléko | leite gordo |
| tuková buňka | célula adiposa |
| ukazatel směru | marca de direcção |
| úroková sazba | taxa de juro |
| válečný průmysl | indústria bélica |
| válečný zločin | crime de guerra |
| včelí královna | abelha-mestra |
| věcná záruka | garantia real |
| věková kategorie | faixa etária |
| veřejná zakázka | contrato público |
| veřejné  výdaje | despesa pública |
| veřejný ochránce práv | provedor je Justiça |
| vesmírná loď | nave espacial |
| větrná elektrárna | instalação de energia eólica |
| vězeňský tábor | campo de prisoneiros |
| vodní elektrárna | central hidroeléctrica |

| | |
|---|---|
| volný čas | lazer |
| výkonný orgán | órgão executivo |
| zahraniční dluh | dívida externa |
| záchvat pláče | crise de choro |
| zapsané zavazadlo | bagagem de porão |
| zavírací nožík | canivete |
| zavírací nůž | canivete |
| zavírací špendlík | alfinete |
| zdrojový kód | código fonte |
| znaková řeč | língua gestual |
| zúčtovací jednotka | unidade de conta |
| životní prostředí | meio ambiente |

# Apêndice 2: Código fonte da página web do dicionário de colocações online

```
<?
$title = "Česko-portugalské kolokace / As colocações checo-
português";
?>

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html>
<head>
<meta content="text/html; charset=UTF-8" http-equiv=Content-Type>
<meta content="cs" http-equiv=content-language>
<meta content="kolokace, colocações, čeština, portugalština"
name=keywords>
<meta content="Online slovník česko-portugalských kolokací"
name=description>
<meta content="Alena Vašíčková; arnostka@gmail.com" name=author>

<title><?=$title?></title>

<link rel="stylesheet" type="text/css" href="kolokace.css">
<link rel="stylesheet" type="text/css" href="print.css"
media="print">

</head>
<body>

<div id="page">
<div id="top">
 
</div>
<div id="bg">


<!-- = = = = = = = = = = = = M A I N = = = = == = = = = = = = -->


<div id="main">

<?
include("AdoDb/adodb.inc.php");
$CONN = NewADOConnection('mysql');
```

```
$CONN->PConnect("mysql.webzdarma.cz", "kolokace", "*********",
"******");


// vymazani hodnot z formulare pri zmacknuti tlacitka
// preloz do PT/CS/odesli

if ($_POST['SUBMIT1']) {
  $pt = "";
} else {
  $pt = $INTOPT;
}

if ($_POST['SUBMIT2']) {
  $cs = "";
} else {
  $cs = $INTOCS;
}

if ($_POST['SUBMIT']) {
  $newCS = "";
  $newPT = "";
} else {
  $newCS = $CS;
  $newPT = $PT;
}

?>


<h2>Vyhledávání překladu kolokace</h2>

<FORM METHOD="POST" ACTION="kolokace.php">
<span class="form">cesky</span><INPUT TYPE=TEXT NAME=INTOPT
VALUE="<?=$pt?>">
  <INPUT TYPE="SUBMIT" VALUE="preloz do PT" NAME="SUBMIT1">
     </FORM>



<FORM METHOD="POST" ACTION="kolokace.php">
<span class="form">portugalsky</span><INPUT TYPE=TEXT NAME=INTOCS
VALUE="<?=$cs?>">
  <INPUT TYPE="SUBMIT" VALUE="preloz do CS" NAME="SUBMIT2">
     </FORM>


<?

if ($_POST['SUBMIT1']) {
```

```
    $sql = "SELECT * FROM slovnicek WHERE cs='$INTOPT'";
    $r = &$CONN->Execute($sql);


    if ($CONN->Execute($sql) === false) {
      echo('<span class="b">Nastal nejaky problem s vyberem z
databaze.</span><br>');
      echo("Problem spociva v: ".$CONN->ErrorMsg()."<br>");
      echo("Napis to <a
href=\"mailto:arnostka@gmail.com\">webmasterovi</a>.<br>");
    } else {
        if ($r->EOF) {
          echo ("Hledany vyraz neni v databazi.");   }
          else  {

       while (!$r->EOF) {
         echo('<br><b>'.$r->fields[1].'</b>: '.$r->fields[2]);
         $r->MoveNext();
           }

       }

       }

       }


if ($_POST['SUBMIT2']) {


    $sql = "SELECT * FROM slovnicek WHERE pt='$INTOCS'";
    $r = &$CONN->Execute($sql);


    if ($CONN->Execute($sql) === false) {
      echo('<span class="b">Nastal nejaky problem s vyberem z
databaze.</span><br>');
      echo("Problem spociva v: ".$CONN->ErrorMsg()."<br>");
      echo("Napis to <a
href=\"mailto:arnostka@gmail.com\">webmasterovi</a>.<br>");
    } else {
        if ($r->EOF) {
          echo ("Hledany vyraz neni v databazi.");   }
          else  {

      while (!$r->EOF) {
         echo('<br><b>'.$r->fields[2].'</b>: '.$r->fields[1]);
         $r->MoveNext();
           }
       }

         }
```

```
        }

?>

<h2>Přidejte svůj výraz</h2>


<FORM METHOD="POST" ACTION="kolokace.php">
<span class="form">cesky</span><INPUT TYPE=TEXT NAME=CS
VALUE="<?=$newCS?>"><br>
<span class="form">portugalsky</span><INPUT TYPE=TEXT NAME=PT
VALUE="<?=$newPT?>"><br>
  <INPUT TYPE="SUBMIT" VALUE="odeslat" NAME="SUBMIT">
     </FORM>

<?

if ($_POST['SUBMIT']) {


      $sql = "INSERT INTO slovnicek (cs, pt) VALUES ".
         " ( '$CS', '$PT' ) ";

    if ($CONN->Execute($sql) === false) {
      echo('<span class="b">Nastal nejaky problem s pridanim
zaznamu do databaze.</span><br>');
      echo("Problem spociva v: ".$CONN->ErrorMsg()."<br>");
      echo("Napis to <a
href=\"mailto:arnostka@gmail.com\">webmasterovi</a>.<br>");
    } else {
        echo('<span class="b">Vyrazy pridany do
databaze.</span>');
             }

         }


?>

 <!-- = = = = = = = = = = F O O T E R = = = = = = = = = = = = -->



</div><!-- bg ends -->
</div><!-- page ends -->
</body></html>

<body></html>
```