

As colocações em corpora de PL2:
identificação, classificação e análise de erros.

Ângela Maria Pereira da Costa

Tese de Doutoramento em Linguística

2017

DECLARAÇÕES

Declaro que esta Tese é o resultado da minha investigação pessoal e independente. O seu conteúdo é original e todas as fontes consultadas estão devidamente mencionadas no texto, nas notas e na bibliografia.

A candidata,

Ângela Maria Pereira de Costa

Lisboa, 13. de Julho de 2017

Declaro que a Tese se encontra em condições de ser apreciada pelo júri a designar.

A orientadora,

Maria Tereza Rijo da Fonseca Lima

Lisboa, 13. de Julho de 2017

A co-orientadora,

Luísa Col

Lisboa, 13. de Julho de 2017

Tese apresentada para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Linguística.

Especialidade em Lexicologia, Lexicografia e Terminologia.

Realizada sob a orientação científica da

Professora Doutora Maria Teresa Rijo da Fonseca Lino

e a co-orientação da

Professora Doutora Maria Luísa Torres Ribeiro Marques da Silva Coheur.

Este trabalho teve o apoio financeiro da Fundação para a Ciência e Tecnologia através da Bolsa de Doutoramento ref. SFRH/BD/85737/2012.

Aos meus pais

AGRADECIMENTOS

A realização deste trabalho não teria sido possível sem o contributo precioso de algumas pessoas a quem gostaria de expressar o meu afecto e agradecimento:

Às Professoras Doutoras Teresa Lino e Luísa Coheur, pela orientação, apoio, amizade e total disponibilidade sempre demonstrados ao longo deste trabalho;

A todos os meus ex-alunos que contribuíram com as suas produções escritas para a análise realizada nesta dissertação;

A todos meus colegas engenheiros e linguistas do L2F agradeço o enorme apoio e companheirismo ao longo destes anos;

À Professora Francisca Athayde por ter despertado em mim o interesse pelas idiossincrasias da língua;

Aos meus pais, irmã, amigos e Miguel pela confiança que depositaram em mim e pelas palavras de incentivo em momentos de desânimo.

Agradeço a todos.

**AS COLOCAÇÕES EM CORPORA DE PORTUGUÊS L2:
IDENTIFICAÇÃO, CLASSIFICAÇÃO E ANÁLISE**

ÂNGELA MARIA PEREIRA DA COSTA

RESUMO

Apesar de não existir uma definição clara, não controversa e abrangente de colocação e de se tratar de um conceito de contornos ainda em discussão, as colocações são, sem dúvida, um fenómeno de grande potencial de investigação. Talvez por estas razões se considerem estruturas cruciais para qualquer teoria e prática lexicográfica, tal como para o ensino/aprendizagem das línguas estrangeiras (Fontenelle, 1994: 47).

Não obstante seja do conhecimento geral que as colocações são tão indispensáveis como problemáticas para os aprendentes de uma língua estrangeira e que, por isso, devem assumir um papel importante nas aulas de língua, as dificuldades dos alunos na produção de colocações não foi ainda alvo de uma investigação exaustiva. Tendo em mente esta lacuna, propomo-nos, através da análise de textos produzidos por alunos de Português como Língua Segunda (PL2), elaborar um estudo aprofundado sobre a utilização de colocações por parte de alunos de três níveis de PL2 para, seguidamente, fazermos uma descrição detalhada dos erros encontrados. Temos ainda como objetivo estabelecer comparações entre as produções dos alunos estrangeiros e de nativos de português. Finalmente, faremos ainda uma comparação com erros de colocações produzidos no contexto da tradução automática, uma vez que estes, apesar de não serem erros humanos, podem assumir certas semelhanças com erros produzidos pelos alunos estrangeiros.

PALAVRAS-CHAVE: colocações, português L2, erros

**THE COLLOCATIONS ON PORTGUESE L2 CORPORA:
IDENTIFICATION, CLASSIFICATION AND ANALYSIS**

ÂNGELA MARIA PEREIRA DA COSTA

ABSTRACT

Although there is no clear, uncontroversial and encompassing definition of collocation, as they are still subject of discussion between specialists, collocations are undoubtedly a phenomenon of great research potential. Maybe for these reasons they are crucial structures for any lexicographic theory and practice, as for teaching/learning foreign languages (Fontenelle, 1994: 47).

Despite being common knowledge that the collocations are as indispensable as problematic to learners of a foreign language and, therefore, should play an important role in foreign language classes, students' difficulties in the production of collocations have still not been target of a thorough investigation. Bearing in mind this gap, we propose, through the analysis of texts produced by students of Portuguese as a Second Language, to draw up a detailed study on the use of collocations by students of three levels of Portuguese L2 to, then, do a detailed description of the errors found. We also have as a goal to draw comparisons between the productions of foreign students and native Portuguese. Finally, we will also make a comparison between errors produced in the context of machine translation, since these, although not being human errors, may present certain similarities to errors produced by foreign students.

KEYWORDS: collocations, Portuguese L2, errors

Introdução	1
Motivação e contexto	1
Principais objetivos	4
Organização da tese	8
 Capítulo 1. As Fraseologias	 10
1.1. Caracterização geral do fenómeno fraseológico	10
1.2. Propriedades das fraseologias	13
1.2.1. A fixidez	13
1.2.2. A idiomaticidade	16
1.2.3. A frequência de utilização de fraseologias	17
1.3. Questões tipológicas	18
 Capítulo 2. As colocações	 24
2.1. Definição	24
2.2. Características das colocações	26
2.2.1. Construções de fronteira: entre o fixo e o livre	27
2.2.2. Transparência semântica	28
2.2.3. Imprevisíveis – unidades da língua, não do discurso ..	32
2.2.4. O papel da recorrência	34
2.2.5. Estatuto dos componentes	36
2.2.6. Estrutura sintática	37
2.3. Tipologia de colocações	39
2.4. Funções lexicais	44
2.5. Relações paradigmáticas e sintagmáticas	46

Capítulo 3. As colocações em *corpora* de aprendizagem de L2 49

3.1. O papel das colocações no ensino de L2	50
3.1.1. Devem as colocações ser ensinadas?	53
3.1.2. Materiais de ensino de colocações	56
3.2. Colocações em <i>corpora</i> de L2	59
3.2.1. Tratamento automático dos erros	59
3.2.2. Comparação com as produções de nativos	60
3.2.3. Estudos com produções em L2	61
3.2.3.1. <i>Corpora</i>	62
3.2.3.2. Utilização de uma palavra ou construção sintática ...	62
3.2.3.3. Exercícios	63
3.2.4. Estudos em <i>corpora</i> de PL2	63
3.3. Definição de colocação em estudos com <i>corpora</i> de L2	64
3.3.1. Definição contextual vs definição estatística	65
3.3.2. Definição baseada na LEC	66
3.3.3. Definição abrangente vs definição restrita	67
3.3.4. Definição baseada no uso	68
3.3.5. Definição baseada em padrões	68
3.3.6. Definição baseada na noção de <i>continuum</i>	69
3.4. Identificação de colocações em <i>corpora</i> de L2	70
3.4.1. Processo de anotação das colocações	70
3.4.2. Taxonomia de análise erros	72
3.4.2.1. Análise a três níveis	72
3.4.2.2. Análise a dois níveis	74
3.4.2.3. Análise a um nível	74
3.4.2.4. Análise de aceitabilidade	75

3.4.3.	Dificuldades na anotação de colocações	76
3.4.3.1.	Identificação das colocações	76
3.4.3.2.	CrITÉrios de correção	77
3.4.3.3.	Interpretação do erro	78
Capítulo 4.	<i>Corpora</i> e sistemas de extração	79
4.1.	Recolha de <i>corpora</i>	79
4.1.1.	<i>Corpus</i> de Produções Escritas de Aprendentes de Português Língua Segunda	79
4.1.2.	Recolha de Dados de Aprendizagem de Português Língua Estrangeira	83
4.1.3.	Produções dos Alunos do Curso Semestral de Língua Portuguesa para Estrangeiros	87
4.1.4.	Textos dos alunos do Curso Online da Ciberescola da Língua Portuguesa	89
4.1.5.	Grupo de controlo	91
4.2.	Sistemas de extração para Português	92
4.2.1.	Parâmetros de avaliação	93
4.2.2.	Ferramentas	94
4.2.2.1.	Wortschatz	94
4.2.2.2.	DeepDict	96
4.2.2.3.	CRPC	97
4.2.2.4.	Sketch Engine	98
4.2.3.	Comparação entre os sistemas	102
4.2.3.1.	Extração de colocações	102
4.2.3.2.	Recolha de <i>corpora</i> e estatísticas	103

Capítulo 5. Processo de anotação das colocações 107

5.1. Orientações gerais	107
5.2. Noções necessárias para a definição de colocação	109
5.2.1. Transparência (não fixidez semântica)	109
5.2.2. Arbitrariedade (não previsibilidade).....	110
5.2.3. Co-ocorrência restrita	111
5.2.4. Frequência	112
5.2.5. Construções de fronteira	114
5.2.5.1. Diferença entre combinatória livre e colocação	114
5.2.5.2. Diferença entre locução e colocação	117
5.2.6. Carácter binário e assimétrico	119
5.2.7. Classificação sintática das colocações	120
5.3. Critérios usados para a definição de colocação.....	122
5.4. Anotação manual de colocações	124
5.4.1. Identificação de colocações.....	124
5.4.2. Taxonomia de erros de colocações.....	125
5.4.2.1. Localização	125
5.4.2.2. Descrição	127
5.4.2.3. Explicação	134
5.4.3. Desafios na anotação de colocações	138
5.4.3.1. Problemas na identificação de colocações	138
5.4.3.2. Problemas no julgamento de correção.....	139
5.4.3.3. Problemas na interpretação dos erros	140

Capítulo 6. Análise das colocações em PL2.....143

6.1. Colocações no <i>corpus</i> de PL2	143
---	-----

6.1.1. Descrição geral da produção de colocações	144
6.1.2. Padrões sintáticos	146
6.1.3. Verbos e substantivos	151
6.1.4. Colocações mais frequentes	157
6.2. Tipos de erros identificados nos <i>corpora</i>	160
6.2.1. Localização dos erros	160
6.2.2. Descrição dos erros	164
6.2.3. Explicação dos erros	169
6.3. Medidas de avaliação da competência colocacional	176
6.3.1. Proliferação	177
6.3.2. Precisão	178
6.4. Competência colocacional nativa e não-nativa	178
6.4.1. Proliferação nativa e não-nativa	179
6.4.2. Precisão nativa e não-nativa	181
6.5. Impacto das variáveis individuais	181
6.5.1. Nível	182
6.5.2. Língua Materna	183
6.5.3. Anos de estudo da L2	184
6.5.4. Imersão	186
6.5.5. Uso da língua alvo	187
6.5.6. Número de L2	188
6.5.7. Idade	189
6.5.8. Formação académica	190
6.6. Erros colocacionais em PL2 e em MT	191
6.6.1. Localização dos erros colocacionais	193
6.6.2. Descrição dos erros colocacionais	195

Conclusão	198
Resumo do trabalho realizado	198
Principais conclusões	200
Contribuições	207
Limitações e sugestões para trabalho futuro.....	209
 Bibliografia	 212
 Listas	 226

Introdução

Motivação e contexto

Apesar de existir um interesse crescente sobre o ensino-aprendizagem do léxico por parte de investigadores e didatas das línguas, a verdade é que, no que especificamente concerne o Português L2 (PL2), são ainda escassos os trabalhos de investigação. O estudo nuclear de Leiria (2006), debruçando-se sobre a aquisição e ensino do léxico em português língua não materna, constitui, por este motivo, uma referência fundamental no âmbito dos estudos desta natureza. A investigação levada a cabo pela autora avalia o vocabulário usado por alunos de Português L2 em material escrito, aprecia quais os itens lexicais, e quais as suas componentes e características que maior resistência oferecem à aquisição, apontando, ainda, sugestões para a aplicação dos resultados obtidos ao ensino e à produção de materiais didáticos. Pela minha parte, centrar-me-ei em questões que se prendem com a aprendizagem dum sector específico do léxico – o fraseoléxico – por parte de alunos de português língua não materna.

Pawley e Syder (1983) argumentam que atingir a proficiência numa L2 que se aproxime da de um falante nativo requer não só o conhecimento de um sistema de regras que gere um número infindável de enunciados, mas também o conhecimento de “sequências memorizadas” e de “frases lexicais” (Ellis, 1994: 85). No entanto, esta competência fraseológica, ou falta dela, recebeu ainda pouca atenção (Howarth, 1998).

Muitas das dificuldades dos alunos de L2 correspondem a áreas tradicionais da língua: gramática ou vocabulário, sendo estes problemas resolvidos com uma gramática ou dicionário. A falta de competência fraseológica é mais complexa, uma vez que os alunos muitas vezes não têm consciência dessa falha e os professores nem sempre a sabem descrever. Este tipo de problema pode não ser considerado muito grave, uma vez que a inteligibilidade pode ser pouco afetada, mas pode ter um impacto grande na qualidade de um texto e o seu acumular pode resultar numa falta de precisão. Uma construção gramaticalmente correta, como **reach findings* em vez de *arrive at findings* (ing.), não quebra as regras gerais da língua, mas demonstra problemas de restrição, neste caso na seleção do verbo, que não são fáceis de explicar a um aluno (Howarth, 2001 [1998]: 162). Os professores de L2, em geral, reconhecem este tipo de erros mas desconhecem uma metodologia de análise do mesmo. Não conseguem especificar o que está errado e, portanto, não conseguem dar assistência.

So often the patient language learner is told by the native speaker that a particular sentence is perfectly good English... but that native speakers would never use it. How are we to explain such a state of affairs? (Allerton, 1984: 39 apud Howarth, 1998b: 13)

Para que o aluno aprenda/memorize estas estruturas fraseológicas tem, necessariamente, e num primeiro momento, de conseguir reconhecê-las enquanto tais e, para tal, é necessária uma descrição do fenómeno. Um dos cenários em que esta informação pode ser veiculada aos aprendentes é a sala de aula, mas o processo institucional nem sempre confere a devida centralidade à dimensão sintagmática das línguas, passando o ensino de estruturas complexas, fixas ou semifixas, para segundo plano ou sendo simplesmente apresentado como uma lista de expressões.

Do vasto conjunto de unidades plurilexémicas que constituem o fraseoléxico, esta dissertação debruçar-se-á sobre as chamadas colocações. Esta escolha recaiu sobre estas estruturas porque se trata de um fenómeno que assume particular relevância no quadro da combinatória lexical estável, em virtude da elevada frequência com que ocorre nas línguas. Citando Prieto *et al.*, (2009: 368), o fenómeno colocacional surge tanto na linguagem quotidiana (*tomar um café*), como

em âmbitos mais específicos¹ (*clonar um gene*), por isto, o seu ensino deve afrontar-se desde o início do processo de aprendizagem, já que as colocações permitem designar inúmeras atividades do cotidiano, como *dar um passeio, tomar banho, levantar dinheiro* ou *fazer uma pergunta*.

Não obstante a sua notória prevalência no discurso e sendo entidades complexas semanticamente transparentes, uma vez que o seu significado pode, muitas vezes, ser deduzido do significado das partes que a compõem, bem como da forma como se combinam, nem sempre lhes é reconhecida fixidez por parte dos alunos. Assim, não é óbvio para o aprendente que as colocações devam ser memorizadas de forma holística. Frequentemente, é apenas aquando da realização de tarefas de tradução que o aluno toma consciência da coesão interna que as caracteriza. Assim, quando o aprendente se depara com a necessidade de encontrar um equivalente português para uma colocação da sua língua materna (LM), apercebe-se, não raras vezes, de que uma tradução literal nem sempre será a solução mais adequada. O exemplo de Sanromán (2006) ilustra bem esta questão complexa. Segundo o autor, muitas línguas dispõem de uma palavra que significa *pergunta*, mas em cada língua esta combina-se com um verbo diferente: em espanhol *hacemos una pregunta*, em inglês *pedimos uma pergunta* (*to ask a question*) e em francês *colocamos uma pergunta* (*poser un question*). Por estas razões as colocações causam especiais dificuldades de tradução e levantam problemas no ensino-aprendizagem de línguas estrangeiras.

A aprendizagem de colocações tem uma importância reconhecida pela comunidade no ensino de L2 (Granger, 1998b; Howarth, 1998; Nesselhauf, 2003, 2005; Alonso Ramos, 2010; Lewis, 2000). Para estes autores dominar uma língua estrangeira e atingir um nível semelhante ao de um nativo implica não só conhecer as suas regras gramaticais e fazer escolhas lexicais, mas também selecionar a colocação apropriada. Estas não são uma opção estilística e ornamental na superfície textual; são essenciais para a comunicação efetiva. A sua aprendizagem em contexto facilita a sua recuperação como blocos lexicais e, ao mesmo tempo, melhora a fluidez e a correção do aluno (Higuera Garcia, 2004: 15).

¹ Cf., a este propósito, o trabalho de Deschamps (2013) sobre o ensino das colocações terminológicas, no âmbito jurídico.

A grande presença de colocações nas línguas e a importância crescente que têm vindo a adquirir na área da tradução e do ensino de L2 fez com que, nos últimos anos, tenham surgido vários estudos sobre este tema. No entanto, são ainda escassos os recursos colocacionais existentes, nomeadamente dicionários, mas são também necessários estudos sobre as colocações presentes em *corpora* de aprendentes, sendo que o primeiro está estritamente relacionado com o segundo. O estudo de dados reais permite-nos, através da linguística de *corpus*, fazer uma análise significativa de textos escritos por aprendentes estrangeiros e compreender as dificuldades reais dos alunos para, assim, serem criados recursos lexicais que verdadeiramente ajudem o aprendente de línguas.

Para colmatar estas necessidades propomo-nos fazer uma avaliação exaustiva da utilização das colocações em *corpora* de aprendizagem de Português L2. Para a aprendizagem das colocações ser significativa, temos que descobrir os erros cometidos com maior frequência pelos estudantes, de modo a que possamos identificá-los, classificá-los e analisá-los. Só com um trabalho prévio desta natureza será possível criar material didático específico orientado ao ensino das colocações.

Principais objetivos

Tendo em mente a necessidade de conhecer as verdadeiras dificuldades experienciadas pelos alunos na produção de colocações e de perceber quais são efetivamente as colocações que os alunos usam e quais são corretas e incorretas, para assim se oferecer uma ajuda eficiente, o principal objetivo deste trabalho é descrever a produção colocacional livre de alunos de PL2.

Para cumprir este objetivo é necessária uma análise linguística fina dos erros em colocações e a sua anotação em *corpora* de aprendizagem. Assim, a primeira tarefa a realizar é a recolha de produções de alunos de PL2. Vamos recolher dados produzidos em tarefas escolares de escrita por alunos de três níveis de proficiência, de modo a percebermos as variações na produção de colocações ao longo dos anos de aprendizagem, e de cinco línguas maternas diferentes: Espanhol, Italiano, Francês, Inglês e Alemão. Seleccionámos estas línguas porque são aquelas que

dominamos, sendo este conhecimento fundamental para percebermos a origem dos possíveis erros, nomeadamente a influência da LM.

Antes de começarmos a análise do *corpus*, e dado o problema de definição e delimitação do conceito de colocação, estabeleceremos critérios definitórios desta.

Estabelecido o nosso objetivo principal e o âmbito da nossa investigação, propomo-nos:

- **Avaliar motores de extração**

Realizamos uma avaliação de motores de extração a funcionar para o português. Esta avaliação permitiu-nos responder a questões, como:

Qual o motor que extrai mais colocações corretamente?

Quais as métricas utilizadas para extração das colocações?

Apesar de termos feito esta pequena avaliação das ferramentas de extração, uma vez que os resultados obtidos não serviam os nossos propósitos, as colocações presentes nos *corpora* recolhidos serão identificadas manualmente com a validação de um segundo anotador. As colocações encontradas serão ainda anotadas como certas ou erradas. A observação destes dados vai-nos permitir dispor de informação não só sobre os erros colocacionais dos alunos, mas também sobre a sua destreza na hora de produzir colocações e das carências detetadas no seu uso.

- **Caracterizar a produção de colocações dos alunos de PL2**

Depois de terem sido identificadas as colocações no *corpus* de aprendizagem dos alunos de PL2, começamos a caracterização da produção colocacional dos alunos, respondendo às seguintes questões:

Quais os padrões sintáticos das colocações utilizados pelos alunos?

Quais são os verbos e substantivos mais usados na construção de colocações?

Quais as colocações mais frequentes?

- **Estabelecer uma taxonomia de erros em colocações**

Em relação às colocações que foram consideradas erradas e para descrevermos os tipos de erros encontrados, propomos uma taxonomia de erros em colocações, tendo em consideração a localização, a descrição e a explicação do erro. Existem trabalhos que comparam quantitativamente e qualitativamente o uso de colocações, mas nós vamos focar-nos na análise detalhada dos erros em colocações e a sua anotação em *corpora* de aprendizagem. Esta sistematização vai-nos permitir responder a questões, como:

Será possível sistematizar os erros em colocações?

Será que os erros colocacionais apenas se localizam no colocativo?

Será que a substituição de um dos constituintes da colocação é o erro mais comum?

Será que os alunos criam colocações quando deveriam usar uma palavra?

Qual o papel da língua materna na produção das colocações?

Que mecanismos de apreensão são utilizados pelos alunos estrangeiros na aprendizagem de colocações?

- **Definir métricas colocacionais e estabelecer o seu impacto em variáveis individuais dos alunos**

Para descrevermos a produção de colocações dos aprendentes com uma maior precisão, vamos medir a competência colocacional dos alunos. Para tal, propomos duas métricas: proliferação e precisão. Estas métricas ajudam-nos a evidenciar a relação entre o número total de colocações produzidas em relação ao número de palavras (proliferação), o número de colocações corretas em relação ao número total de colocações produzidas (precisão). Depois de medirmos a competência colocacional, o próximo objetivo ao qual nos propomos é compreender como variáveis individuais dos alunos se relacionam com a proliferação e a precisão. Tentamos então responder a questões, como:

Será que alunos com mais anos de estudo de português são mais precisos na produção de colocações?

Será que a imersão pode ajudar no aumento da proficiência colocacional?

Alunos que usam a L2 fora do contexto escolar são mais prolíficos?

Será que saber um maior número de línguas estrangeiras ajuda na produção de colocações?

Como é que a idade do aluno se relaciona com a proficiência colocacional?

Será que formação académica do aluno influencia o conhecimento colocacional dos alunos?

- **Comparar a competência colocacional nativa e não-nativa**

Depois de termos caracterizado a utilização de colocações por parte dos aprendentes, levamos a cabo uma comparação entre estes e os nativos de português, usando dados de um grupo de controlo. Tentaremos então dar resposta à pergunta:

Quais são as diferenças entre a produção de colocações em aprendentes de PL2 e nativos do português?

Para responder a esta questão, tivemos em consideração aspetos como a proliferação e precisão, o tipo de construção sintática, os verbos e os substantivos mais usados e as colocações mais usadas.

- **Comparar entre os erros colocacionais em PL2 e em MT**

Por último, realizamos uma comparação entre os erros produzidos em colocações por alunos e os erros produzidos por motores de tradução também em colocações. O nosso objetivo será perceber se:

Os erros em colocações produzidas por sistemas de tradução automática demonstram semelhanças com os erros produzidos pelos alunos de PL2 e se poderão ser usados para prever erros humanos?

Organização da tese

A presente dissertação está organizada em seis capítulos. No primeiro capítulo, esclarecemos alguns conceitos teóricos indispensáveis à leitura dos dados da investigação empírica. Começarei, assim, por esclarecer o que se entende por “léxico”, “fraseologia” e “fraseoléxico” para, posteriormente, me centrar na definição de “colocação”, situando-a no vastíssimo e muito heterogêneo leque de estruturas subsumidas no conceito genérico de “fraseologia” (cap. 2).

No terceiro capítulo, discutimos o papel das colocações no ensino de L2 e apresentamos trabalho relacionado na área da análise da produção de colocações em contexto de aprendizagem de uma língua estrangeira, explicitando as definições que usam e as taxonomias de erros que seguem.

No quarto capítulo, descrevemos a recolha do *corpus* de produções de alunos de português língua estrangeira, apresentamos dados estatísticos relativos ao mesmo e descrevemos ferramentas de extração automático de potenciais colocações em Português Europeu. Terminamos o capítulo, com uma comparação entre esses sistemas para que possamos perceber quais as suas vantagens e limitações.

No quinto capítulo, apresentamos o processo de anotação manual do *corpus* recolhido. Começamos por descrever as orientações gerais e os critérios definitórios de colocação que usaremos. Identificamos alguns problemas encontrados no processo de identificação das colocações e, por último, descrevemos a taxonomia de erros seguida.

No sexto capítulo, apresentamos a análise realizada sobre as colocações produzidas pelos alunos de PL2: observamos dos seus padrões sintáticos, os verbos e nomes usados e as colocações mais frequentes, comparando-os com produções colocacionais de nativos. Seguidamente, fazemos a análise dos erros, seguindo a taxonomia previamente apresentada. Na terceira parte deste capítulo, propomos métricas para a avaliação da competência colocacional. Seguidamente, usamos estas mesmas métricas para comparar a competência colocacional nativa e não-nativa. Relacionamos também estas métricas com algumas variáveis individuais dos alunos,

como os anos de estudo da L2, área da formação académica ou a idade do aluno. Finalmente, vamos ainda estabelecer uma comparação entre os erros dos alunos e os erros em colocações gerados no contexto da tradução automática.

Por fim, no último capítulo, serão apresentadas as conclusões finais e o trabalho futuro.

Capítulo 1. As Fraseologias

1.1. Caracterização geral do fenómeno fraseológico

Uma vez que este trabalho se centra num tipo particular de estruturas de natureza fraseológica – as colocações –, é necessário esclarecer e delimitar previamente alguns conceitos que com elas estão intimamente associados, designadamente, os conceitos de “léxico” e “Fraseologia”/ “fraseologia”.

Começando por definir o conceito mais genérico, neste estudo entender-se-á ‘léxico’ como “uma componente das línguas que tem por função produzir, armazenar, processar e transmitir signos que os falantes usam como matéria-prima na elaboração de raciocínios e na construção de enunciados verbais” (Rio-Torto 2006: 12)². São muito diversificadas as unidades que têm um estatuto lexical: unidades univerbais (simples, derivadas ou compostas), como nomes, verbos pleno³,

² Segundo a nova Terminologia Linguística para os ensinos Básicos e Secundário (TELBS), ‘léxico’ é entendido como o “(...) conjunto de todas as palavras ou constituintes morfológicos portadores de significado possíveis numa língua, independentemente da sua atualização em registos específicos. O léxico de uma língua inclui não apenas o conjunto de palavras efetivamente atestada num determinado contexto, mas também as que já não são usadas, as neológicas e todas as que os processos de construção de palavras da língua permitem criar” (<http://tlebs.dgicd.min-edu.pt/>). A forte ligação entre as estruturas lexicais e a gramática não poderá, igualmente, ser ignorada. ‘Léxico’ pode também ser entendido como “léxico mental” – o conjunto de unidades lexicais armazenadas na memória de longo prazo às quais o indivíduo recorre no processo de receção e produção do discurso. Por outro lado, o termo ‘léxico’ pode ainda ser usado na aceção de ‘dicionário’ ou ainda de “léxico computadorizado” – componente de um sistema de processamento automático das línguas naturais. Cf. Athayde (2007: 20-21).

³ Por ‘verbo pleno’ entende-se um verbo que é portador de significado lexical, tendo, por isso, ao invés do que sucede com os verbos auxiliares, a capacidade de funcionar, sintacticamente, como núcleo de uma expressão predicativa.

adjetivos e alguns advérbios; unidades funcionais – preposições, determinantes, pronomes, conectores, entre outros; entidades abaixo do nível da palavra – i.e., os formativos de palavras complexas⁴ – e entidades acima do nível da palavra – as unidades pluriverbais estáveis (fraseologias) (Athayde, 2007: 21).

Burger (2003: 16-23), um nome de referência no domínio da investigação fraseológica, tem uma concepção lata e dinâmica de fraseologia/Fraseologia⁵. Para este autor, as fraseologias têm que respeitar alguns critérios, nomeadamente (i) terem uma natureza pluriverbal, (ii) revelarem fixidez, assumindo esta duas dimensões: (ii.i) a fixidez psicolinguística (o que pressupõe que correspondem a representações holísticas, armazenadas no léxico mental dos falantes de uma comunidade) e (ii.ii) a fixidez estrutural (morfofossintáctica e lexical). Em relação à idiomaticidade, entendida como não-composicionalidade semântica⁶, sublinhe-se que esta é uma condição suficiente, mas não imprescindível para a inclusão de uma combinatória no domínio fraseológico. Estes critérios permitem que, sob a alçada “ampla” do termo “fraseologia” se incluam pares fraseológicos, colocações, construções com verbo-suporte, fraseologias de valor comunicativo (fórmulas comunicativas ou fraseologias pragmáticas), combinações com estrutura frásica ou mesmo com estrutura textual, como provérbios, máximas, sentenças, aforismos, fórmulas de juramento, anúncios necrológicos ou convites. Burger (2003: 48) acrescenta ainda a este elenco as fraseologias que pertencem a línguas de especialidade.

As fraseologias, tal como os processos de criação e de formação de novas palavras⁷ estão ao serviço do enriquecimento do léxico das línguas (Fleischer, 1997: 1). Aliás, as unidades plurilexémicas fixas ou semifixas não têm uma função acessória

⁴ A classificação das entidades abaixo do nível de palavra como itens lexicais é discutível, uma vez que não existe ainda um consenso na comunidade científica sobre se devem, ou não, ser incluídas no léxico. A este respeito, cf. Schindler (2002: 38).

⁵ O termo “Fraseologia” aplica-se igualmente à subdisciplina da Lexicologia que se ocupa do estudo das combinações, mais ou menos estáveis, de unidades lexicais constituídas por mais de duas palavras gráficas.

⁶ O significado de uma expressão idiomática não resulta, conseqüentemente, dos significados parciais dos elementos que a compõem, nem da sua forma de combinação, ou seja, a interpretação do significado de uma expressão idiomática não se pode basear apenas numa leitura literal dos seus constituintes, mas implica sempre uma leitura fraseológica.

⁷ Tal como a criação de novas palavras (neologismos), entrada de palavras estrangeiras no léxico, mudança de significado de uma palavra, derivação ou conversão.

em relação às entidades monolexémicas: a relação entre ambas é de complementaridade. As fraseologias não são apenas “apêndices de luxo” da comunicação, mas peças fundamentais, quer dos processos cognitivos, quer da interação verbal (Athayde, 2007: 21). Na realidade, estas combinatórias constituem uma mais-valia no âmbito do léxico das línguas: o facto de serem compostas por vários elementos com diferentes traços semânticos faz com que tenham uma enorme potencialidade semiótica⁸. Burger e Fleischer são dois dos autores que, nos seus trabalhos, refletem sobre a mais-valia expressiva das fraseologias (Burger, 2003) e sobre o papel pragmático-retórico que desempenham na construção do texto (Fleischer, 1997).

Pelo que ficou dito e tal como demonstra o seguinte extracto retirado de Altenberg (2001 [1998]: 101), a Fraseologia é uma disciplina cujo objecto de estudo é bastante vasto e, assim sendo, de difícil delimitação. Esta dificuldade de demarcação verifica-se em duas vertentes: a nível do que é ou não é fixo ou semifixo (fronteira entre o léxico e a sintaxe, dependendo dos diferentes graus de fixidez e lexicalização), mas também ao nível das diferentes classes de fraseologias. Mas é, justamente, esta dificuldade de análise que a torna uma área de especial interesse para tradutores, lexicógrafos e didatas das línguas (maternas e estrangeiras):

Phraseology is a fuzzy part of language. Although most of us would agree that it embraces the conventional rather than the productive or rule-governed side of language, involving various kinds of composite units and ‘pre-patterned’ expressions such as idioms, fixed phrases, and collocations, we find it difficult to delimit the area and classify the different types involved. Indeed, (...), the existence of a large number of more or less prefabricated expressions in language blurs the distinction between lexicon and grammar and strongly suggests that ‘lexicalization and productivity are matters of degree’ rather than a clear-cut-dichotomy. This state of affairs creates problems of description for both the empirical and the theoretical linguist, at the same time as it provides a challenge to anyone who wants to get a better understanding of language and language use.

Segundo Zuluaga (2002: 58), a procura de fronteiras claras e de um critério básico que sirva para delimitar todos os tipos de unidades fraseológicas é um erro na

⁸ No entanto, existem fraseologias que não têm função referencial, antes assumem um valor primordialmente comunicativo, como, por exemplo, fraseologias pragmáticas do tipo *Bom dia*.

investigação fraseológica. A única solução, segundo este autor, passa pela adopção dos conceitos de **gradualidade** e de **escalaridade**. No entanto, como também refere Zuluaga, as distinções conceptuais necessárias para a análise destas unidades não têm que ser imprecisas e pouco claras. Apesar de as fronteiras entre subtipos de fraseologias não serem nítidas, os conceitos, definições e critérios de delimitação não têm de sê-lo igualmente. Recuperando a imagem utilizada por este investigador (Zuluaga, 2002: 58), não poderemos estabelecer fronteiras claras entre o dia e a noite, por exemplo, mas podemos conhecer os conceitos aos quais correspondem.

1.2. Propriedades das fraseologias

Prosseguimos esta exposição com o inventário e explicitação de algumas propriedades das fraseologias, sendo elas a fixidez, a idiomaticidade e a frequência.

1.2.1. A fixidez

As fraseologias são, como atrás referi, estruturas plurilexémicas fixas ou semifixas, armazenadas, de forma mais ou menos holística, dependendo do grau de fixidez, no léxico mental do(s) falante(s) (tal como as unidades lexicais uniberbais), ficando assim disponíveis, enquanto blocos pré-formados, para momentos de produção e recepção linguísticas (Athayde, 2007: 10-20). São muitas as designações propostas para este tipo de construções: Saussure (1915 *apud* Zuluaga, 1980: 37) chama-lhes “locutions toutes faites”; Jakobson (1956 *apud* Zuluaga, 1980: 59) “stereotyped utterances”, “coded wordgroups” e “idioms”; Burger (²2003) chama-lhes fraseologias (“Phraseologismen”) e mais tarde frasemas (“Phraseme”) (Burger, 2004), ou seja, combinatórias fixas de palavras (“feste Wortverbindungen”) e combinatórias fraseológicas de palavras (“phraseologische Wortverbindungen”), por oposição às combinatórias livres (“freie Wortverbindungen”).

Sinclair (1991: 109-115), um dos linguistas mais atentos à dimensão sintagmática do léxico, explica o modo como textos produzidos por falantes veiculam significados através da alternância entre dois diferentes princípios, o “open-choice principle” (“At each point where a unit is completed (a word or a

phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness.” (*idem*: 109)) e o “idiom principle” (“a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable as segments” (*idem*: 110)). Tal como Sinclair, também Altenberg e Mel’čuk reconhecem que o segundo princípio referido será o dominante (Athayde, 2007: 24):

(...) it is obvious that even a small sample of recurrent word-combinations can tell us a great deal about the phraseology of speech. What is perhaps the most striking impression that emerges from the material is the pervasive and varied character of conventionalized language in spoken discourse. The use of routinized and more or less prefabricated expressions is evident at all levels of linguistic organization and affects all kinds of structures, from entire utterances operating at discourse level to smaller units acting as single words and phrases. (Altenberg, 2001 [1998]: 120)

People speak in set phrases, rather than in separate words (...). A phraseme is a lexical unit; and, more crucially, it is the numerically predominant lexical unit: in any language – i.e. in its lexicon – phrasemes outnumber words roughly ten to one. (Mel’čuk, 2001 [1998]: 24)

Porque não se pode falar de combinatórias estáveis sem falar do seu oposto, parece-me importante, neste momento da exposição, mencionar os factores que aproximam e afastam as combinatórias lexicais estáveis das combinatórias livres de palavras.

Do ponto de vista da estrutura sintáctica, os dois tipos de sequências demonstram semelhanças, mas diferenciam-se pelo facto de as fraseologias serem **reproduzidas** num enunciado e não **produzidas**, de acordo com as regras morfossintáticas e semânticas do sistema (Athayde, 2007: 22). Por outras palavras, as expressões fixas são fruto da constância do emprego de uma combinação no seio de uma comunidade de falantes, i. e., são produto de processos de repetição na diacronia das línguas (Zuluaga, 1980: 16). Dito isto, compreende-se facilmente que existam estados intermédios de fixidez e lexicalização, isto é, a “fixidez” e a “lexicalização” são propriedades de natureza escalar. Enquanto num extremo do *continuum* da fixidez e da lexicalização encontramos as expressões de cunho idiomático, semântica e sintacticamente não-decomponíveis e que representam uma dada unidade semântica e psicolinguística (são armazenadas holisticamente no

léxico mental dos falantes), noutro extremo encontram-se combinações menos estáveis e semanticamente composicionais (não-idiomáticas), como os pares fraseológicos (*são e salvo*), as colocações (*fumador inveterado*) e as construções com verbo-suporte (*dar um abraço*). Estes dois últimos tipos de construções estão “a meio caminho” entre o “livre” e o “fraseológico”, já que os seus constituintes mantêm, em regra, o significado de que são portadores fora da combinatória e continuam a gozar de alguma autonomia sintáctica, constituindo, assim, um fenómeno de transição. Por outras palavras, não podendo ainda ser consideradas expressões lexicalizadas e completamente fixas, estas combinatórias situam-se numa zona de fronteira entre léxico e gramática (Sintaxe, Morfologia), encontrando-se em vias de fixação⁹.

A fixidez é uma característica que assume diferentes dimensões. A já mencionada fixidez psicolinguística, que permite que as expressões sejam armazenadas no léxico mental dos falantes como estruturas pré-fabricadas, tal como uma palavra. Podemos também referir a dimensão estrutural da fixidez. Ou seja, a nível semântico, as expressões estão sujeitas a restrições que não permitem a comutação de componentes, como: *A cavalo/*burro dado, não se olha o dente*. Também existem limitações a nível gramatical, por exemplo a variação no número dos substantivos (*Em casa de ferreiro espeto *espetos de pau*) ou transformações como a interrogativa, passiva, negativa ou alteração da ordem dos componentes (*de alma e coração/ *de coração e alma*). Nestes casos, destruimos o significado fraseológico e passamos a ter combinatórias livres.

Todavia, em certos casos e dentro de certos limites, a fixidez não é absoluta e podem ser admitidas variações, sem perda do significado fraseológico: *meter a pata/pé na poça*. É possível também encontrar expressões modificadas para efeitos de ironia, para causar estranhamento, provocar riso ou para efeitos semântico-estilísticos. Como por exemplo: *O pior cego é aquele que... se recusa a ter cão!* ou *Gato escaldado... morre, naturalmente!*

⁹ Se determinado tipo de fraseologias com menor grau de estabilidade se aproximam da sintaxe livre (*pôr a mesa, fumador inveterado*), outros tipos de sintagmas fixos estão próximos das palavras compostas (*mercado negro, guerra fria, centro comercial*). Outras subclasses de fraseologias são semelhantes a palavra, em termos pragmático-funcionais (*Bom dia! = Olá!*). A este respeito, cf. Athayde (2007: 22).

Por último, a fixidez pode também ter uma dimensão pragmática. Trata-se de fraseologias que têm uma função pragmática, ou seja, fórmulas de rotina convencionadas a determinadas situações, como: *Bom dia*. Mas também fórmulas que regulam a comunicação, como: *Estás a perceber?*

1.2.2. A idiomaticidade

Apesar da idiomaticidade não ser uma condição necessária para a definição das fraseologias, é uma característica partilhada por algumas destas estruturas.

Por idiomaticidade entende-se o facto de o significado de uma expressão não resultar dos significados dos elementos que a compõem, nem da sua forma de combinação. Ou seja, a interpretação do significado de uma expressão idiomática não se pode só basear numa leitura literal dos elementos que a compõem, existe sempre uma leitura fraseológica.

Fleischer (²1997:33) diz que se devem distinguir graus de idiomaticidade, uma vez que, em algumas expressões um ou mais componentes conservam o sentido literal, sendo mais fácil interpretá-lo. Dou alguns exemplos de expressões parcialmente idiomáticas (“teilidiomatische Ausdrücke”) em português: *armado até aos dentes*, *tocar de ouvido*. Nestas expressões um dos elementos mantém o seu significado literal. Por outro lado, temos as totalmente idiomáticas (“vollidiomatische Ausdrücke”), como: *perder a cabeça*, *esticar o pernil*, *bater no fundo* ou *cantar vitória*. Nestes casos o significado não pode ser inferido através da interpretação dos constituintes, uma vez que se trata de significados figurados. Quanto maior for o grau de idiomaticidade mais coesão estrutural existe entre os componentes, mais compactas e mais fixas são e também mais difíceis são de interpretar.

O termo “idiomático” remete para o que é próprio de um idioma. De facto, as expressões idiomáticas refletem, em muitos casos, especificidades históricas e sócio-culturais de uma comunidade e daí advêm as dificuldades de encontrar os equivalentes diretos entre línguas. Estas idiossincrasias das línguas representam, sem dúvida, uma dificuldade acrescida para os tradutores, para os aprendentes estrangeiros, mas também para nativos, tratando-se de expressões marcadas por

regionalismos.

Embora as expressões idiomáticas estejam muito marcadas pelas particularidade de cada língua e cada cultura, muitas vezes podem também mostrar-nos o carácter universal do ser humano e da linguagem. Desde já é possível encontrar expressões idiomáticas em todas as línguas, obviamente nem todas com uma relação de equivalência total. Concluo citando Weinreich (1966: 69): “Idiomacity is important for this reason, if for no other, that there is so much of it in every language” (*apud* Zuluaga, 1980: 20).

1.2.3. A frequência de utilização de fraseologias

Análises teóricas do fenómeno fraseológico associadas a resultados trazidos pela Linguística Computacional e Linguística do *Corpus*, bem como por estudos de natureza psicolinguística, vieram chamar a atenção para a surpreendente prevalência das sequências discursivas (de tipo funcional e lexical) convencionalizadas no discurso oral e escrito dos falantes/escreventes (Athayde, 2007: 23).

A constatação da frequência de uso de determinadas combinatórias fixas ou semifixas de palavras pôs, pois, em causa a questão da liberdade de que, teoricamente, gozam os falantes na produção do seu discurso. Uma parte considerável das combinatórias de palavras não é inteiramente livre; os falantes têm à sua disposição uma grande quantidade de blocos *pré-fabricados*, que, no entanto – sobretudo quando não são de natureza idiomática –, não são reconhecidos como tal pelos falantes nativos (Corpas Pastor, 1996: 15). Como escreveu André Clas (1994: 576), “Nous sommes donc d’une certaine façon en “liberté de parole surveillée!”. Também Jakobson tinha já chamado a atenção para esse facto: “this freedom is relative, and the pressure of current clichés upon our choice of combination is considerable” (Jakobson, 1956: 58).

Apesar da presença de unidades fraseológicas não ser condição para a existência de uma língua, Jakobson e Ágel defendem que esta presença é uma característica geral que pode ser observada empiricamente em todas as línguas:

In any language, there exist also coded word-groups called phrase-words. (Jakobson, 1956: 59 apud Zuluaga, 1980: 20)

Mehr oder weniger feste Wortverbindungen stellen keine Sonder-, sondern vielmehr Normalfälle sprachlicher Zeichenbildung. (Ágel, 2004: 65)

[As combinações, mais ou menos fixas, de palavras não são uma exceção, mas o caso mais normal de construção de significados nas línguas.]

(...) phraseologische Sprachzeichen [repräsentieren] nicht Sonderfälle, sondern den sprachlichen Normalfall... (idem: 70)

[Os signos linguísticos de natureza fraseológica não representam casos excepcionais, mas o fenómeno linguístico mais recorrente.]

Tendo igualmente observado a elevada frequência de emprego das fraseologias por parte dos falantes, Aisenstadt é mesmo da opinião de que a co-ocorrência restrita de palavras pode ser considerada um universal linguístico (Aisenstadt, 1979 *apud* Corpas Pastor, 1996: 80). Apesar de existirem poucos estudos que definam fraseologismos tendo por base um critério quantitativo de ocorrências, alguns estudiosos, especialmente da linha de pensamento de Sinclair¹⁰, argumentam que a frequência observada da ocorrência de fraseologias deve exceder a frequência esperada (cf. Gries, 2008).

1.3. Questões tipológicas

A primeira classificação de combinação de palavras que descreveremos é a de Hausmann (1984). Este autor diferencia inicialmente dois grandes grupos: estruturas fixas e estruturas não-fixas (cf. Figura 1). Dentro das estruturas fixas, encontramos as palavras compostas e frases feitas, como *caixa-forte* e *fazer a cabeça a alguém*. Dentro das não-fixas, encontramos combinações, que se dividem em *Konter-Kreation* (counter-creation), *Kollokation* (collocation) e *Ko-Kreation* (co-creation). A Ko-Kreation é uma combinatória livre, que respeita apenas as regras do sistema da língua, *uma casa bonita* ou *uma aula interessante*. As *kollokation*, como já foi

¹⁰ Esta perspetiva foi utilizada para a construção da *Collins Birmingham University International Language Database*, mas conhecida como Cobuild.

explanado nos pontos anteriores, são produtos “semi-prontos” da língua, o falante não as pode alterar criativamente, vai buscá-las à memória como um todo e o ouvinte reconhece-as. Como exemplo, podemos referir *dar um passeio* ou *pôr a mesa*. Por último, as Konter-Kreation são combinações nas quais um elemento é selecionado, com base no seu significado, porém é aplicável a um número restrito de objetos extralinguísticos e, assim, admite um número muito restrito de colocativos, podendo concluir-se que têm uma capacidade combinatória limitada por motivos semânticos. Exemplos deste fenómeno seriam combinações do género de *um rebanho de ovelhas* ou *um nariz aquilino* (cf. Hausmann, 1984: 398-399).

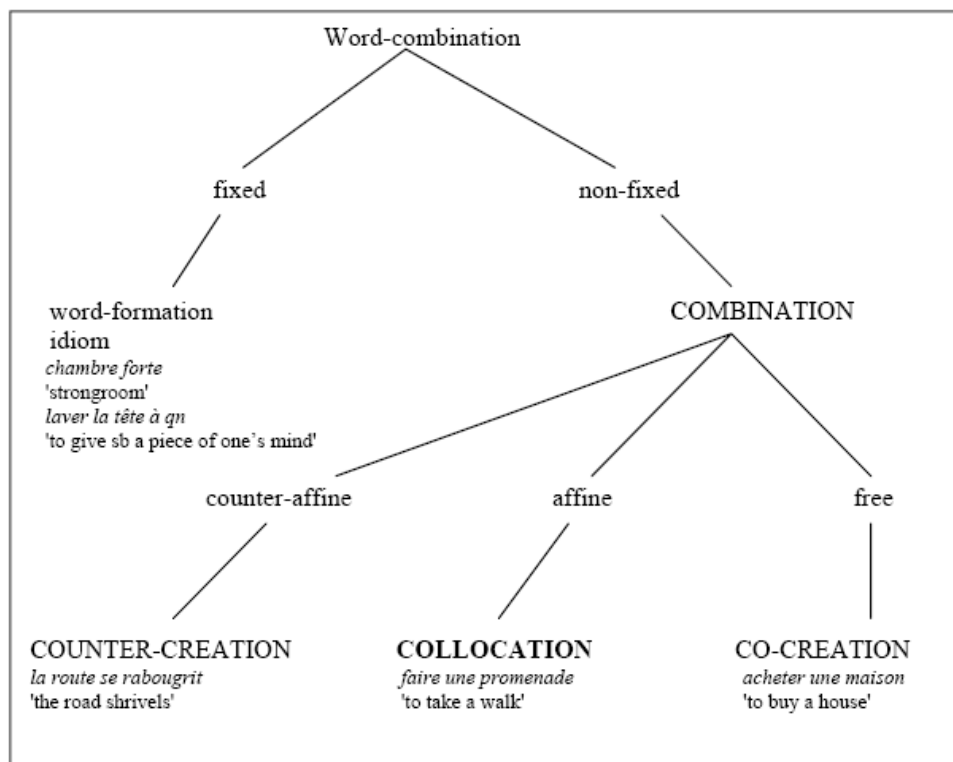


Figura 1 Classificação de combinação de palavras de Hausmann (1984).

Mel’cuk divide a combinatória lexical restrita em frasesmas semânticos e frasesmas pragmáticos ou pragmatemas (cf. Figura 2). “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 $/A \oplus B/$, mas cujo significado é diferente do da soma dos significados dos lexemas constituintes.” (Mel’cuk, 1995: 181 *apud* Sanromán, 2001: 182). Os pragmatemas são fórmulas cristalizadas cuja representação sintática pode ser considerada irregular, assim como a formação do

significado do todo a partir das partes. O seu uso é situacional por isso incluem marcadores conversacionais, fórmulas de rotina, provérbios, citações, *slogans*, etc. Exemplos de pragmatemas são *Muitos parabéns*, *Proibido estacionar*.

Por sua vez, os frasemas semânticos dividem-se em frasemas completos (expressões idiomáticas), semi-frasemas (colocações) e quase-frasemas.

Os frasemas completos¹¹, sem composicionalidade, têm o estatuto de unidade linguística¹². Na expressão *baixar os braços*, o significado não resulta da soma dos dois elementos *baixar* (A) e *braços* (B), mas cria-se um significado diferente “C” (desistir) que não inclui nem A, nem B (Mel’cuk, 1998). Alonso Ramos (1993) também sublinha esta especificidade, caracterizando os frasemas como sendo não semanticamente composicionais, ou seja, a soma do sentido dos constituintes não é igual ao seu sentido global, perdendo cada um dos elementos a sua função referencial. A metáfora, a metonímia ou outros processos estilísticos podem explicar a formação destas expressões¹³. Trata-se pois de expressões coesas, que assumem pouca variação formal e que, no plano semântico, se distinguem claramente das colocações, uma vez que, tipicamente, na construção das colocações, pelo menos um dos elementos mantém o seu sentido habitual.

Os quase-frasemas são combinatórias que mantêm o significado individual de cada um dos lexemas que as compõem, mas é-lhes acrescentado um novo sentido não dedutível através da soma das partes constituintes. Atentemos no exemplo *cinturão negro*. Nesta combinação está expresso o sentido de *cinto* (A) e de *negro* (B), porém existe um terceiro sentido (C) que surge da combinação em si (“grau de conhecimento nas artes marciais”). Mais próximas das expressões idiomáticas do que as colocações, podemos mesmo considerá-las unidades lexicais e não uma combinação de várias unidades (Sanromán, 2001: 167).

Finalmente, uma colocação, ou semi-frase, AB é uma combinação de dois ou mais lexemas A e B, cujo significante é a soma regular dos significantes dos

¹¹ Estas expressões beneficiam de múltiplos nomes sem que haja um consenso (“expression figées”, “locutions”, “unités phraséologiques” (Bally, 1909), “phrasèmes complets” (Mel’cuk *et al.*, 1995), “unités polylexématiques” (Corbin, 1997).

¹² Este título serve também para as palavras simples, mas não para as colocações que são codificadas na base. Se bem que no caso das colocações com colocativo tipificante, estas possam também ser consideradas unidades referenciais.

¹³ Estes processos podem também ocorrer na formação das colocações opacas.

lexemas constituintes $/A \oplus B/$, e cujo significado 'X' inclui o significado do lexema A mais um significado 'C' ('X' = 'A $\oplus$ C'), de tal maneira que o lexema B que exprime 'C' não é selecionado livremente (Sanromán, 2001: 170). A noção de colocação de Mel'cuk será mais detalhadamente explicada no capítulo seguinte.

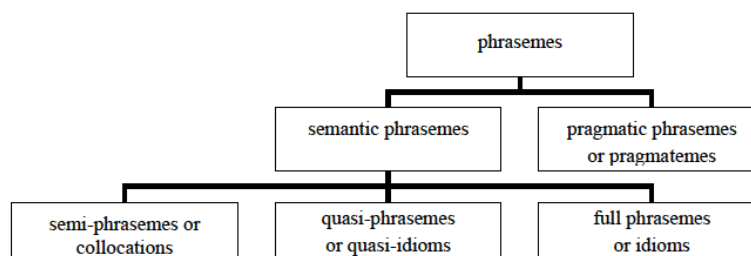


Figura 2 Classificação de combinação de palavras de Mel'cuk (1998)

Cowie (2001[1998]) separa as combinações de palavras entre dois grandes tipos: *composites* e *formulae* (cf. Figura 3). *Formulae* são combinações com uma função essencialmente pragmática, do tipo *Bom fim-de-semana* ou *Muitos parabéns*. Os *composites* têm essencialmente uma função sintática. As estruturas que pertencem a este grupo distingue-se tendo por base dois critérios: o critério da transparência e da comutabilidade. A transparência refere-se ao facto dos elementos da cominatória terem um significado literal ou não. A comutabilidade refere-se ao grau de comutação dos elementos dentro da combinatória. Cowie distingue três tipos de combinatórias:

- **colocações restritas:** alguma substituição é possível, mas existem limitações arbitrárias na substituição e um dos elementos tem um sentido não-literal e outro mantém o seu sentido literal. A combinação é transparente (ex.: *desempenhar um cargo*).
- **Idiomas figurativos:** a substituição dos elementos muitas vezes é possível e a combinação tem um sentido figurativo, mas preserva uma interpretação literal (ex.: *fazer uma inversão à U*).
- **Idiomas puros:** a substituição dos elementos é impossível e a combinatória tem um significado figurativo que não admite uma interpretação literal (ex.: *abrir os olhos a alguém*).

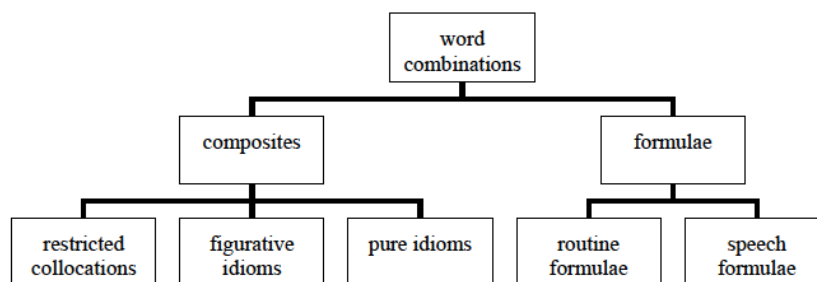


Figura 3 Classificação de combinação de palavras de Cowie (2001[1998]).

O critério que norteará a organização tipológica das fraseologias e o consequente enquadramento das construções que analisarei, nesta dissertação, será o da sua função semiótica. **Adoptarei, assim, a tipologia que se generalizou no âmbito da Germanística, e que foi esboçada por Burger (2003), mas que é aplicável a outras línguas, nomeadamente ao Português.** A classificação inicial assenta num critério funcional, tendo-se, assim, em consideração a função que as fraseologias desempenham na comunicação. Com base neste critério, estabelece-se uma distinção entre a) **fraseologias de valor referencial** (“referentielle Phraseologismen”), b) **fraseologias estruturais** (“strukturelle Phraseologismen”) e c) **fraseologias de valor comunicativo** (“kommunikative Phraseologismen”). As fraseologias de valor referencial (a) têm a capacidade de denotar objetos ou “estados de coisas” da realidade extralinguística. Como exemplo, podemos aduzir expressões como *Em casa de ferreiro, espeto de pau, deitar achas para a fogueira, sorriso amarelo* (provérbios, idiomatismos ou colocações, respectivamente). Em b), encontramos as fraseologias que assumem uma função essencialmente gramatical – as locuções – como *em relação a, assim como*; em c) encontram-se as expressões com função pragmática, como *Tanto quanto sei..., Com os meus melhores cumprimentos..., Era uma vez..., Feliz Natal!*. Para uma classificação mais fina desta última classe, Burger (2003: 38) propõe, num segundo momento, a aplicação de um critério semântico. Assim, as fraseologias de valor referencial seriam passíveis de uma subclassificação em **fraseologias proposicionais**, com capacidade para **realizar asserções** sobre objetos e “estados de coisas” (*Em casa de ferreiro espeto de pau, Até o diabo se ria*) e **fraseologias denominativas**, que **referem/denominam** objetos

e “estados de coisas” (*sorriso amarelo, esticar o pernil*)¹⁴. Ainda segundo Burger (2003: 39 e ss.), é necessária uma subdivisão das fraseologias proposicionais através de critérios sintáticos e textuais¹⁵: as “frases fixas” (que constituem réplicas e, consequentemente, evidenciam menor autonomia sintática, de que são exemplos *Já cá faltava!, Só faltava esta agora!*) e as “fórmulas tópicas” (enunciados genéricos, como os provérbios (*Quem vai ao ar perde o lugar, Cão que ladra não morde*) e os lugares-comuns (*Ninguém é perfeito, A vida é curta*), com maior independência sintática). Para as fraseologias denominativas, Burger (2003: 37-38) propõe uma subdivisão semântica, de acordo com o grau de idiomatidade (não-composicionalidade semântica) que estas possam revelar: fraseologias idiomáticas (“Idiome”), como *ficar fora de si*, parcialmente idiomáticas (“Teil-Idiome”), como *jurar a pés juntos* e não-idiomáticas ou pouco idiomáticas (“nicht-idiomatisch” ou “schwach-idiomatisch”), como *lavar os dentes*. Estas últimas são incluídas por Feilke (1998) e Burger (2003) no conceito genérico de ‘colocação’ (“Kollokation”¹⁶). A Figura 4 resume a classificação apresentada.

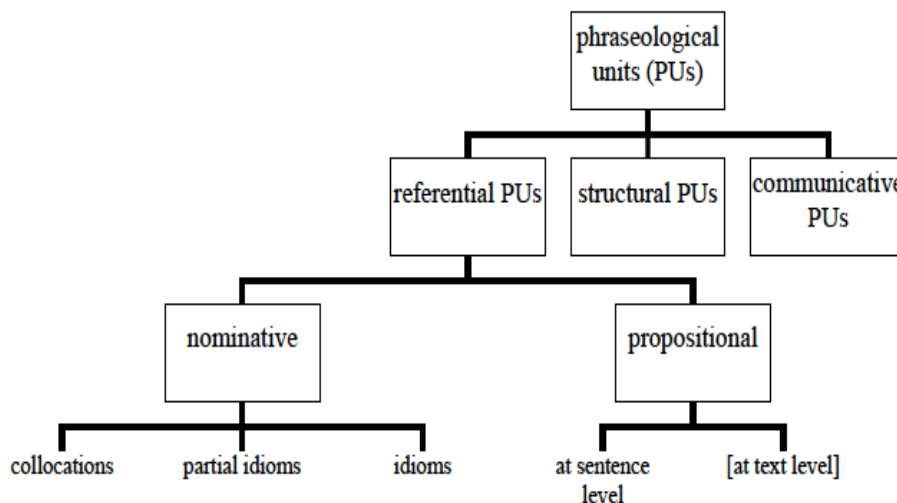


Figura 4 Classificação de combinação de palavras de Burger (2003).

¹⁴ Esta subdivisão assenta, também, em critérios sintáticos. As fraseologias denominativas ocorrem sob a forma de unidades abaixo do nível da frase, enquanto que as fraseologias proposicionais apresentam a forma de frase ou mesmo de texto.

¹⁵ De acordo com Burger (2003: 38-39), às fraseologias proposicionais seria contraproducente aplicar um critério semântico, pois tal operação criaria uma cisão na classe dos provérbios, uma vez que estes apresentam diferentes graus de idiomatidade.

¹⁶ Um resumo da tipologia proposta por Burger (2003) pode encontrar-se em Athayde (2008).

Capítulo 2. As colocações

Es cierto que no podemos establecer fronteras rígidas entre el día y la noche ni, menos, entre el crepúsculo y la noche, pero ello no quiere decir que no podamos tener un concepto claro, o una intuición precisa, de lo que son el día, la noche y el crepúsculo. (Zuluaga, 2002: 58)

Recuperando a imagem utilizada por Zuluaga (2002: 58), que compara as colocações com o crepúsculo, é certo que não podemos estabelecer fronteiras claras entre o dia e a noite, mas podemos conhecer os conceitos aos quais correspondem. Tendo esta ideia em mente, neste capítulo, vamos elencar algumas das definições de colocação mais utilizadas pela literatura. Descrevemos algumas das suas características definitórias e apresentamos as tipologias mais conhecidas para a classificação deste fenómeno fraseológico. Por último, abordamos as noções de função lexical e de relação sintagmática e paradigmática.

2.1. Definição

Enquanto fenómeno linguístico, as colocações têm sido objecto de inúmeros trabalhos de investigação e têm-se revelado uma temática extremamente profícua de estudo.

Já no início do século passado, Bally (1909:139) tinha assinalado o fenómeno da co-ocorrência usual de lexemas e define-a como uma combinação que tem para o falante nativo o carácter de já formada e conhecida (“déjà vu”), mas cujos elementos constituintes conservam a sua autonomia semântico-sintáctica, apesar de

contribuírem para o significado global da expressão. Bally chama a estas combinatórias “groupements usuels” (*apud* Zuluaga, 2002: 56-57).

O termo “colocação”¹⁷ foi primeiramente cunhado por Firth (1951 *apud* Zuluaga, 2002: 56-57): “I propose to bring forward a technical term, meaning by “collocation””. Todavia, Firth entende o fenómeno colocativo como uma combinação de palavras, abrangendo associações lexicais paradigmáticas (*médico, hospital*), associações lexicais sintagmáticas (*argumento de peso*) e associações que incluem palavras gramaticais (*lembrar-se de*) (cf. Grossmann e Tutin, 2002: 8). Esta concepção lata, introduzida por Firth nos anos 50 é recuperada pelos funcionalistas ingleses. Como tal, para Halliday e Hasan (1976), as colocações têm como papel funcional contribuir para a coesão textual¹⁸.

Aisenstadt (1979) chama-lhes *colocações restritas*, ou seja combinações cujos constituintes têm uma comutabilidade restrita. Ao contrário das expressões fixas, as colocações restritas não formam uma unidade semântica e dispõe de um certo grau de variabilidade. Ao contrário das frases totalmente livres, os elementos não são livremente comutáveis. Isto explica porque não podemos dizer que o *leite estava podre* e que o *ovo estava rançoso ou amargo* (*apud* Fontenelle, 1994: 42).

John Sinclair (1991, 70) adota a mesma definição de Firth, mas essencialmente defini-la através de critérios do tipo estatístico. Esta abordagem abrangente é ainda partilhada por alguns investigadores do Tratamento Automático da Língua (por exemplo, Smadja e McKeown, 1991) e por investigadores que se enquadram na linha de pensamento dos contextualistas ingleses (Williams, 1998). Contudo, alguns anos depois, como se observa em Athayde (2009), “Uma concepção mais restrita de ‘colocação’, ancorada essencialmente nos textos fundadores – no âmbito da Lexicologia, Didática das Línguas e Lexicografia – de Alan Cruse (1986) e (2000), Mel’čuk (2001 [1998]) e, sobretudo, de Hausmann (1984), (1985), (1989), foi-se, todavia, impondo”. **Neste trabalho, tal como em Athayde (2009), definir-se-á**

¹⁷ O termo “colocação” deriva do Latim “collocare” que significa “juntar, reunir, montar”.

¹⁸ “laugh... joke, blade... sharp, ill... doctor (...) The cohesive effect of such pairs depends not so much on any systematic relationship as on their tendency to share the same lexical environment, to occur on COLLOCATION with one another. In general, any two lexical items having similar patterns of collocation – that is, tending to appear in similar contexts – will generate a cohesive force if they occur in adjacent sentences” (Halliday & Hasan, 1976, 285 – 286).

‘colocação’, de acordo com Grossmann e Tutin (2002: 3), como “a co-ocorrência lexical privilegiada de dois (ou mais) elementos linguísticos que entre si estabelecem uma relação sintáctica”.

Este fenómeno linguístico de preferência de um item lexical por um outro, em vez de um seu sinónimo, deve-se a constrangimentos que não se justificam ao nível sintáctico ou ao nível do significado conceptual, mas através do uso (Van Roey, 1990: 45 *apud* Granger, 2001: 146-147). Como explica Fontenelle, os adjectivos *bad*, *rotten*, *sour* ou *rancid* podem todos ser combinados com nomes de alimentos, mas não são comutáveis entre eles. Isto significa que algumas palavras são mais prováveis de se combinarem com itens específicos para formarem combinações que “soam” mais naturais, enquanto que outro tipo de combinações simplesmente não ocorrem, apesar de serem possíveis e compreensíveis, pelo menos em teoria (Fontenelle, 1994: 42).

Como podemos concluir, as colocações são um fenómeno de fronteira: trata-se de unidades pré-fabricadas, que se encontram entre as locuções e as combinações livres. Talvez este seja um dos motivos que faz com que estas construções sejam tão difíceis de definir, mas que sejam também um exemplo bastante claro da natureza gradual e dinâmica das línguas (Zuluaga, 2002: 56). Apesar da dificuldade em encontrar uma definição que seja comumente aceite entre os especialistas e de, na literatura, serem apresentadas variadíssimas definições, existem algumas características das colocações que parecem ser apontadas pela maioria dos autores.

2.2. Características das colocações

Nas seguintes secções, vamos abordar algumas das características definitórias de colocação. Iniciamos por descrever a sua posição de fronteira entre o fixo e o livre. No ponto 2.2.2., descrevemos como o seu significado se constitui. Abordamos a sua imprevisibilidade na secção 2.2.3. A sua recorrência e as estruturas sintáticas que assumem serão discutidas nos pontos 2.2.4. e 2.2.5., respetivamente.

2.2.1. Construções de fronteira: entre o fixo e o livre

Não é nova a ideia de que existem estados intermédios de fixidez entre as expressões memorizadas e as combinações livres. Esta ideia surge em Bally, no *Traité de Stylistique Française* de 1909. Num dos extremos, encontramos as expressões fixas, como provérbios ou frases feitas. Nestes casos, o significado total não se reflete no significado dos componentes, ou seja, a interpretação semântica não se faz de uma maneira composicional. Do outro lado do *continuum*, encontramos as combinações livres, do tipo *argumento interessante*. Nestes casos, o sentido é analisável e previsível, não apresentam coesão e são desconhecidas antes do ato de fala que as produz. Entre estes dois extremos, encontram-se grupos intermédios. Bally chama-lhes *séries fraseológicas*; outros autores, como Zuluaga (1980: 23) e Bahns (1993: 57) chamam-lhes colocações. Os seus elementos conservam a sua autonomia e, apesar da evidente afinidade entre eles, estão a meio caminho entre o livre e o fraseológico. Não se identificam propriamente com nenhum deles. São um fenómeno de intercepção, apresentando características comuns aos dois (Zuluaga, 2002: 59). Segundo Bally, *gravemente ferido*, *amar perdidamente* e *desejar ardentemente* constituem-se como exemplos protótipos de colocação, uma vez que os advérbios intensificadores não são comutáveis entre si. Causaria alguma estranheza, se alguém dissesse *ardentemente ferido*, *desejar perdidamente* ou *amar gravemente* (cf. Grossmann e Tutin, 2002: 8).

No *Manual de Fraseologia Española* de Gloria Corpas Pastor (1996), colocação é definida como unidade fraseológica que, do ponto de vista do sistema da língua, não se distingue dos sintagmas livres, gerados a partir de regras criativas da língua. É ao nível da norma que se observam diferenças, visto que as colocações demonstram preferências de combinação e restrições impostas pelo uso¹⁹ (fixação interna) (*apud* Liang, 1991-1992: 154). Por outras palavras, a co-ocorrência lexical livre é uma questão semântica; a co-ocorrência lexical restrita, fenómeno em que se enquadram as colocações, é uma questão meramente lexical, determinada pelo uso

¹⁹ Sendo assim, não é de estranhar que as colocações não apresentem as artificialidades formais (recursos fónicos, disposições rítmicas, esquemas sintácticos, etc) nem os rasgos arcaicos próprios de um considerável número de unidades fraseológicas pertencentes a outras esferas, como as locuções ou as fórmulas de rotina (Corpas Pastor, 1996: 76).

(Sanromán, 2001: 170). Assim *comprar um chapéu* não é uma colocação, porque o sentido de *comprar* ou *chapéu* não é restrito ou limitado. Não existe qualquer afinidade entre as duas palavras, uma vez que podemos comprar inúmeras coisas (Clas, 1994: 577). Já uma expressão do género de *café fraco*, é uma colocação, pois existe restrição na escolha do adjetivo *fraco*, uma vez que não se pode dizer *café débil*, apesar de *débil* ser um sinónimo de *fraco*. Como refere Clas, a partir do momento em que se escolhe uma palavra, a seguinte é, de uma certa maneira, imposta (Clas, 1994: 576). Estas combinatórias estão disponíveis na mente como uma estrutura *pré-fabricada*, são psicolinguisticamente salientes, ao contrário das combinações livres, e são já conhecidas antes do ato de fala que as reproduz.

As colocações distinguem-se também das expressões fixas (provérbios, frases feitas, por exemplo *bater a bota/ de mãos a abanar*) através da não-fixidez e da transparência. A fixidez é um conceito gradual em vários aspetos: no fim da escala encontramos as expressões fixas e, no outro extremo, encontram-se expressões com poucos sinais de fixidez, as colocações²⁰ (Burger, 2004: 21).

O significado das colocações advém do significado dos constituintes, ao contrário de muitas expressões fixas. Quando falamos em *bater a bota*, não estamos efetivamente em falar em botas, mas de um significado não-literal que advém dessa expressão. Quando falamos em *leite azedo*, trata-se realmente de leite que está estragado pois, neste caso, não é permitida uma leitura que não seja literal²¹.

Sendo uma construção de fronteira entre a pré-construção e as construções livres, as colocações constituem um problema real para os linguistas e para a linguística aplicada (Grossman e Tutin, 2002: 1).

2.2.2. Transparência semântica

Para Grossman e Tutin (2002), transparência semântica, ou seja, o não-sentido figurado das expressões em questão, é um aspeto específico do fenómeno

²⁰ Mas mesmo ao nível das colocações, existem diferentes graus de coesão entre os seus constituintes, isto é, com diferentes níveis de estabilidade lexical. (Athayde, 1994:35). Howarth (2001 [1998]: 169) propõe 5 níveis de categorização das colocações: do nível 1 (mais livres) ao nível 5 (mais restritas).

²¹ No entanto, existem áreas de transição entre as colocações e os *idioms* (Cruse, 86).

colocativo. Cruse (2000) e Hausmann (1989) sublinham que as colocações podem ser interpretadas pelo locutor, mas a sua forma não deixa, por isto, de ser imprevisível. Segundo Hausmann, é exatamente a capacidade de serem transparentes que diferencia as colocações das locuções²². Cruse (1986), na sua definição de colocação, chama atenção para esse aspeto: “O termo colocação será usado para referir as sequências de itens lexicais, que habitualmente co-ocorrem, mas que não deixam de ser completamente transparentes no sentido de que, cada constituinte lexical, é também um constituinte semântico “.

Mel’čuk (1997), Hausmann (1984, 2004) e Cruse (2000) também entendem as colocações como um fenómeno de codificação. Sublinham que as colocações podem ser interpretados pelos locutores, mesmo que a sua forma não seja previsível, ou seja, são transparentes, do ponto de vista da recepção, porém para uma locutor não nativo podem causar problemas na sua produção. Os exemplos de Hausmann (*fumeurs invétérés* e *feuilleter un livre*) serão facilmente compreendidos por um locutor não-nativo de francês, mas este seria incapaz de os produzir, se não os tivesse já memorizado como tal. Como chamam atenção Grossman e Tutin (2002:1), as colocações são transparentes na recepção, os problemas surgem quando o aluno estrangeiro tem que as produzir/codificar.

Apesar das colocações serem classificadas como estruturas de fácil decodificação, devido à sua transparência semântica, algumas combinações podem causar particular estranheza ao aluno estrangeiro. Tratam-se de combinações onde existe um elemento que é usado num sentido metafórico ou “figurado”, convencionado, como *lever* em *lever dinheiro*. Nestes casos, os lexemas são ampliados ao seu sentido físico ou primário, um sentido figurado (por isso também se fala em linguagem figurada). Isto faz-nos compreender que, em cada combinação, realiza-se somente uma aceção, uma das possibilidades de sentido dos componentes. Assim *sentido* em *sentido de humor*, não significa direção, nem significado, nem função de percepção feita por um órgão, mas capacidade ou aptidão. Isto explica-se por estas combinações serem atos de fala (produtos

²² Mel’čuk chama atenção para alguns casos de colocações onde não existe transparência (ex.: *peur bleue*). Assim, a transparência semântica, segundo este autor, não pode ser considerada uma característica definitiva.

instantâneos do discurso) e, na fala, orientam-se os signos abstractos, virtuais, polissémicos para o particular e circunstancial. Nas colocações só se realizam algumas das possibilidades que cada um dos seus componentes em separado, no sistema de signos, tem para oferecer (Zuluaga, 2002: 61-62).

É de primordial importância salientar que a transparência não impede a colocação de ser imprevisível. O aprendente estrangeiro consegue compreender a colocação (se conhecer as palavras), mas não é automaticamente capaz de a reproduzir. Ele tem que as aprender, pois as línguas, apesar de todas as combinações logicamente possíveis, fazem escolhas idiossincráticas.

Hausmann (2004: 309), a fim de dilucidar as dificuldades do aprendente estrangeiro, narra a história de Momo, um aluno muito aplicado que conhece todas as línguas do mundo, excepto o Alemão. Esta pequena história põe em relevo as dificuldades que este tipo de combinatória coloca aos aprendentes de línguas estrangeiras e os obstáculos que estes encontram ao consultarem dicionários. Passo a destacar o seguinte extracto:

Kennen Sie eigentlich Momo? Momo lernt Deutsch als Fremdsprache. Er kann schon alle Sprachen der Erde, nur Deutsch lernt er gerade. Heute morgen ist er aufgestanden und hat sein Bett gebaut. Nein, nicht gebaut, gebaut ist militärisch, er hat sein Bett gemacht. Dann hat er sich die Zähne gewaschen. Nein, nicht gewaschen, gewaschen ist Französisch (se laver les dents), er hat sich die Zähne geputzt. Dann hat er eine Dusche gemacht. Nein, nicht gemacht, machen ist Italienisch (fare la doccia). Er ist unter die Dusche gegangen oder hat geduscht. Kann man sagen: er hat eine Dusche genommen? Momo wird sich erkundigen. Dann muss er für das Frühstück den Tisch legen. Nein, nicht legen, legen ist Englisch (to lay the table). Er muss den Tisch decken. Und nach dem Frühstück muss er den Tisch freimachen. Nein, nicht freimachen, das ist Englisch (to clear the table), er wird den Tisch abräumen. Und

*so geht das endlos weiter*²³.

Como podemos concluir da leitura deste trecho, apesar de Momo conhecer todas as línguas do mundo, esse conhecimento não o ajuda a produzir corretamente as colocações em língua alemã.

Será, portanto, fundamental distinguir, no domínio das colocações e para a dilucidação da natureza destas, a “codificação” da “descodificação”. Na terminologia de Makkai (1972), retomada por Cruse (2000: 76), as colocações são consideradas “idioms of encoding”, mas não, simultaneamente, tal como as expressões idiomáticas, “idioms of decoding” (Athayde, 2009). As colocações não causam, pois, especiais dificuldades no processo de compreensão, visto serem estruturas relativamente transparentes (não-idiomáticas) e sintacticamente regulares; o problema reside na codificação, pois a relação entre base e colocativo é, na maior parte dos casos, arbitrária, no sentido em que se apresenta como uma combinação imprevisível, sem uma motivação semântica particularmente evidente que a explique²⁴. Só a comparação de diferentes línguas mostra, com clareza, como a escolha do colocativo não é livre, mas, antes, imposta (Sanromán, 2001: 167). Assim, a maioria dos falantes não-nativos de uma dada língua será capaz de compreender uma colocação; as dificuldades, porém, residem na produção. Tal circunstância faz com que, para o aluno estrangeiro, não exista nenhuma razão, *a priori*, que o leve a

²³ [Será que conhecem o Momo? Momo está a aprender Alemão como língua estrangeira. Ele já sabe todas as línguas do mundo, mas só agora é que está a aprender Alemão. Hoje de manhã levantou-se e construiu a cama [er hat sein Bett gebaut]. Não, sein Bett bauen [construir a cama] é linguagem militar, ele fez a cama [er hat sein Bett gemacht]. Depois lavou os dentes [er hat sich die Zähne gewaschen]. Não, não é waschen, waschen é Francês (se laver les dents), os alemães dizem que ele limpou os dentes [er hat sich die Zähne geputzt]. A seguir fez um duche [er hat eine Dusche gemacht]. Não, não é gemacht, machen é Italiano [fare la doccia]. Foi para debaixo do chuveiro [er ist unter die Dusche gegangen] ou tomou um duche [er hat geduscht]. Pode-se dizer: ele tomou um duche [er hat eine Dusche genommen]? Momo vai informar-se. Depois precisa de deitar ou assentar a mesa [den Tisch legen] para o pequeno-almoço. Não, não é legen, legen é Inglês (to lay the table). Ele tem de pôr a mesa [den Tisch decken]. E a seguir ao pequeno-almoço tem de libertar a mesa [den Tisch freimachen]. Não, não é freimachen, isso é Inglês (to clear the table), ele tem de levantar a mesa [den Tisch abräumen]. E por aí adiante, sem fim à vista. (Hausmann, 2004: 309) (tradução de Athayde (2009)).

²⁴ É indispensável, todavia, sublinhar que nem sempre a preferência combinatória é semanticamente imotivada, como demonstra Feilke (1998: 74-76), a propósito de colocações do tipo *die Zähne putzen/*waschen, die Haare waschen/*putzen, die Nase putzen/*reinigen, das Haus putzen/*waschen, das Auto waschen/*putzen*. De facto, será vantajoso referir que há traços de *waschen* – [+Wasser] ([+água]) e de *putzen* – [+Gerät] ([+instrumento/+utensílio]) – que justificam a combinatória preferencial de certos Ns com determinados Vs, apesar de, também nestes casos, não estar ausente o factor ‘arbitrariedade’/‘convencionalidade’ (Athayde, 2009).

construir a colocação *café fraco*²⁵ em vez de **café débil* (Mackin, 1978 *apud* Herbst, 1996: 386). É a exposição reiterada à colocação e a constância do emprego da combinatória no seio da comunidade de falantes que, de facto, conduz à fixação da expressão numa determinada forma, ficando excluídas pela norma linguística social outras formas que, segundo o sistema, também seriam possíveis (Zuluaga, 1980: 95). Deverá, pois, reiterar-se a ideia de que a co-ocorrência lexical livre é uma questão essencialmente semântico-sintáctica e de que a co-ocorrência lexical restrita, fenómeno em que se enquadram as colocações, é uma questão fundamentalmente lexical determinada pelo uso (Athayde, 2007: 112).

2.2.3. Imprevisíveis – unidades da língua, não do discurso

Um outro aspeto característico das colocações, já mencionado nos pontos anteriores, é a arbitrariedade das associações lexicais dos itens que a compõem. As colocações não são previsíveis. Como Hausmann refere: a transparência não impede que a colocação seja imprevisível. Não existe uma motivação semântica que possa explicar o facto de *torrencial* poder aparecer em co-ocorrência com *chuva*, enquanto que *precipitação torrencial* nos parece, claramente, uma combinatória menos natural (Grossman e Tutin, 2002: 3). Cada combinatória implica um processo de seleção de componentes, em detrimento de outros, que poderiam perfeitamente cumprir a função de comunicar o mesmo conteúdo: *chá forte/ *chá poderoso*. Se considerarmos as palavras *fraco* e *débil* sinónimos, então o facto de *café fraco* ser aceitável e *café débil* não o ser, não pode ser explicado semanticamente. *Café fraco* não é previsível e deve ser visto como um fenómeno típico do uso, ou usando a terminologia de Coseriu (1977: 143-161), da norma linguística (*apud* Herbst, 1996: 386). Estas restrições combinatórias devem-se ao uso repetido, não existem regras semânticas que as expliquem (Zuluaga, 2002: 63). A repetição conduz à fixação da expressão numa determinada forma, ficando excluídas pela norma linguística social outras formas que, segundo o sistema, também seriam possíveis (Zuluaga, 1980: 95).

As colocações são estruturas idiossincráticas que envolvem conhecimento sobre a palavra individual, que dificilmente é descrito em regras gerais (Herbst,

²⁵ Exemplo do texto de Mackin (1978 *apud* Herbst, 1996: 386) adaptado ao português.

1996: 391). Assim, faz-se uma distinção entre as combinações nas quais uma possível restrição na substituição dos elementos é devida a propriedades semânticas, estas são chamadas *combinatórias livres*, segundo Cowie (1991), e as combinações cujas restrições são, até um certo ponto, arbitrárias – as colocações. Por exemplo, a combinação *ler o jornal* não pode ser alterada para **beber o jornal* ou **ler água*, uma vez que *beber* requer um nome com a propriedade semântica “líquido” e *ler* requer um nome que tenha a propriedade semântica “conter linguagem escrita”. Na combinação *chegar a uma conclusão*, por outro lado, pode ser substituído por nomes que denotem “um objectivo particular”, como *conclusão* ou *compromisso*, mas não *alvo*. Esta restrição não resulta de propriedades semânticas dos elementos em questão, mas de uma convenção da língua. A “afinidade” demonstrada pela associação das duas palavras faz com que as possibilidades combinatórias sejam limitadas. Esta afinidade elimina outras combinações plausíveis e restringe os sentidos possíveis. Embora aqui seja bastante clara a diferença entre restrições semanticamente motivadas e restrições arbitrárias, não é sempre tão claro estabelecer limites (Nesselhauf, 2003: 225).

Nesselhauf desenvolveu a noção de “restricted sense” e estabeleceu dois critérios, aplicáveis a construções de V + N, para perceber se o verbo é usado num sentido restrito. O sentido do verbo é restrito se corresponder a um (ou ambos) dos seguintes critérios:

critério 1 – o verbo (nome) é tão específico que se combina só com um número reduzido de nomes (verbos).

critério 2 – o verbo (nome) não pode ser usado no mesmo sentido com todos os nomes (verbos) sintáctica e semanticamente possíveis.

O verbo *want* não é restrito, uma vez que pode ser combinado com um grande número de nomes e não tem restrições de comutabilidade. O verbo *dial* já tem um sentido restrito uma vez que ocorre com um número limitado de nomes. Trata-se de uma colocação, se o critério 1 ou 2 ou ambos se aplicarem ao verbo da combinação, de uma combinação livre, se o verbo e o nome não forem restritos, e de um *idiom*, se o verbo e nome tiverem um sentido restrito (baseia-se na ideia de que os elementos de uma colocação não têm o mesmo estatuto). Esta classificação tem limitações, uma vez que as combinações de palavras têm uma natureza escalar

e, muitas vezes, uma delimitação exata é impossível. O problema de distinguir arbitrariedade e restrição semanticamente motivada não é resolvido com esta classificação, é apenas facilitado (Nesselhauf, 2003: 226).

Por serem típicas, específicas, características e não haver, aparentemente, uma explicação semântica que justifique os enlaces que compõem as colocações, estas podem ser classificadas como unidades da *langue*, no sentido saussuriano²⁶. Este é um aspeto que também partilham com os frasemas, mas, ao contrário destes, a colocação é sempre transparente: o sentido global é apreendido através dos componentes, um caracteriza o outro (Clas, 1994: 577). Enquanto que *folhear um livro* é uma colocação, pois é uma construção que apresenta fixidez e algum grau de especialidade semântica e, como tal, pertence à *langue*. Já *comprar um livro* trata-se de uma construção não específica e, como tal, pertence à *parole* (Hausmann, 1985: 118).

2.2.4. O papel da recorrência

Segundo Irsula Peña (1994: 32), cada combinação ou co-ocorrência lexical é uma potencial colocação e, cada colocação, teve origem numa combinação, mas nem todas as estruturas de palavras se tornam colocações. As colocações são combinações lexicais hipotéticas, que se tornaram estáveis através do uso repetido (Zuluaga, 2002: 56).

Na definição de colocação, vários factores têm de ser tidos em consideração, desde factores puramente linguísticos, a factores estatísticos (Fontenelle, 1994: 47). Na tradição contextualista inglesa, as colocações, entendidas como palavras que tendem a aparecer juntas, são indissociáveis da noção de *corpus*, uma vez que é a recorrência das associações lexicais nos textos que lhes dá o estatuto de colocação (Sinclair: 1991).

Existem programas de computador que têm como função identificar

²⁶ Do ponto de vista Saussuriano, à *langue* (língua) pertencem os princípios sistemáticos abstractos de uma língua, sem os quais não seria possível produzir um enunciado com significado. Estes princípios resultam de um trabalho colectivo que é interiorizado pelos membros de uma comunidade linguística, mas que vai além da vontade individual. *Parole* (discurso) diz respeito ao uso concreto da língua, ou seja, à produção de enunciados. Trata-se do uso do sistema linguístico, mas não o sistema linguístico.

colocações, em extensos *corpora*, e calcular as suas frequências, através de cálculos de ocorrências. Para muitos lexicógrafos que trabalham com este tipo de programas, as colocações são um fenómeno estatístico. O potencial colocativo é encontrado através da análise e contagem do *span* das palavras que aparecem à direita e esquerda de um determinado item (*node*), chamado *collocational span*²⁷. Se um conjunto de palavras ocorre mais vezes junto que cada uma isoladamente, ou seja, a frequência de co-ocorrência é maior do que seria de prever, formam uma colocação, sendo esta a diferença entre uma combinatória casual e uma colocação significativa.

Mas, usando as palavras de Hausmann (1985: 124): “Não é a frequência que faz a diferença entre uma combinação de *parole* e uma combinação da *langue*” (*apud* Irsula Peña, 1994: 35). Muitas colocações não são usadas com muita frequência e, no entanto, estão disponíveis como unidades pré-fabricadas (Hausmann, 1985: 124).

Através do método estatístico, encontraram-se alguns pares de palavras que não são considerados colocações por alguns linguistas. Pares como *doutor* e *enfermeira* foram encontrados em textos com bastante frequência, mas será que são colocações? Ou são simplesmente palavras que ocorrem num mesmo contexto, porque pertencem ao mesmo campo semântico e têm significados semelhantes? (Fontenelle, 1994: 47).

Segundo Herbst (1996: 382), a frequência da co-ocorrência não é, por si só, um bom critério para o significado de uma combinação. Berry-Rogghe (1972 *apud* Herbst, 1996: 382), tendo como base a análise de um *corpus* de 72.000 palavras, chega à conclusão de que as palavras mais frequentes que aparecem com *house*, são os determinantes *the* e *this* e o verbo *sell*. Isto não é particularmente surpreendente, nem particularmente interessante.

A utilização de cálculos de ocorrências, bem como uma perspectiva mais linguística podem ser utilizadas para a identificação de colocações, embora tenham diferentes implicações, nomeadamente na elaboração de um dicionário de colocações. Uma abordagem que tenha em consideração o número de ocorrências mas que também se baseie numa definição linguística pode ser uma abordagem

²⁷ De acordo com Sinclair (1991, 1970) “the usual measure of proximity is a maximum of four words intervening”.

mais completa.

2.2.5. Estatuto dos componentes

É Hausmann (1984: 401) que esclarece que o estatuto dos constituintes da colocação não é semelhante, existindo entre eles uma relação hipotáctica²⁸ (ou de dependência). Hausmann, o responsável por esta observação, chama *base* (“Basis”) à palavra que determina a escolha do outro elemento co-ocorrente (o substantivo de *a* a *c*, o verbo em *d*, o adjetivo em *e* e o substantivo em *f* na tabela 1) e *colocativo* (“Kollokator”) ao constituinte determinado. Na colocação *fumador inveterado*, o constituinte *fumador* é autónomo no plano semântico. O outro constituinte (*inveterado*) acrescenta uma característica que não modifica a identidade do caracterizado mas que, apesar de transparente, é dificilmente previsível (Hausmann, 1989: 1010). Esta visão de colocação foi adoptada por Evert (2005: 15-25), Heid (1998) e na criação do *Oxford Collocations Dictionary for Students of English* (Gouws e Heid, 2006).

a)	<i>Confirmed bachelor</i>	Eingefleischter Junggeselle	célibataire endurci
b)	his anger falls	Zorn verraucht	la colère s’apaise
c)	to <i>withdraw</i> money	Geld abheben	retirer de l’argent
d)	it is raining heavily	Es regnet in Strömen	il pleut à verse
e)	<i>seriously</i> injured	Schwer verletzt	grièvement blessé
f)	<i>a gust of</i> anger	Wutanfall	une bouffée de colère

Tabela 1 Exemplos de colocações em inglês e as suas traduções em alemão e francês.

Segundo Hausmann (1985: 119), no processo de produção linguística de um falante fica bastante clara a hierarquia base/colocativo. Adoptando a teoria de Hausmann à língua portuguesa, diria que se alguém quer produzir (oralmente ou por

²⁸ No entanto, Feilke (1998: 74) inclui os pares fixos – (pt.) *sal e pimenta*, por exemplo – no conceito de colocação. Neste caso, não estamos perante uma relação hipotáctica entre os constituintes, mas perante uma relação de coordenação (paratáctica).

escrito) um enunciado com o significado “fumador que fuma muito”, não tem dificuldade em usar a palavra “fumador”, mas sim em encontrar o adjetivo “inveterado”, caso não seja falante nativo. Esta é uma prova psicolinguística que demonstra que o falante parte sempre da base para o colocativo (Pöll, 1996: 37).

A dissimetria entre os componentes das colocações é bastante visível quando um aluno tem a necessidade de procurar informação colocacional no dicionário, devendo começar por procurar a base, palavra que não lhe causa problemas, pois tem um correspondente direto na sua língua. Mas esta não é uma tarefa fácil, uma vez que, no caso das colocações, a entrada do dicionário não é, em regra, a base, como deveria ser, mas o colocativo, “muitas vezes distinto de língua para língua e, assim sendo, imprevisível” (Athayde, 2009). Um dicionário de codificação deveria dar-nos a informação de que o nome *fumador* pode ser usado com o adjetivo *inveterado*, visto que é a base que determina a presença do adjetivo e não o contrário. Para propósitos de codificação, faz sentido registar o nome como lema, uma vez que é natural que um aluno estrangeiro se questione sobre como pode exprimir a ideia de alguém que “fuma muito” ou que é um “grande fumador” (Fontenelle, 1994: 43).

2.2.6. Estrutura sintática

Apesar de, como já se acentuou, as colocações se caracterizarem por um certo grau de arbitrariedade na associação dos seus constituintes, é relevante dizer que, do ponto de vista formal, e devido à sua parcial fixidez, as colocações seguem determinados padrões sintáticos regulares²⁹. Por exemplo, as colocações construídas em torno de um Nome incluem Adjetivos e Verbos, mas provavelmente não Advérbios e Conjunções (Grossmann e Tutin, 2002: 3), ocorrendo essencialmente como estruturas binárias (constituídas por dois lexemas ou por dois constituintes – Lexema + Sintagma Preposicional (SP)). Como foi notado por Carter e McCarthy (Carter e MacCarthy, 1988: 35 *apud* Fontenelle: 1994: 2), o conceito de

²⁹ Contudo alguns investigadores consideram que as colocações podem assumir virtualmente qualquer configuração sintática. “The term collocation refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure” (Fontenelle, 1992: 222).

colocação é independente da categoria gramatical da base: a relação semântica que se verifica entre o V *argue* e o Adv *strongly* é a mesma que existe entre o N *argument* e o A *strong*. Surgem-nos, assim, combinações de V + N (sujeito) – *vento/tempestade + amainar, problema + residir*; V + N (objecto) – *assumir responsabilidades, dar um abraço, desempenhar um cargo*; N + A – *fumador inveterado; sono profundo*; Adv + A/Particípio – *profundamente agradecido, firmemente convencido; estreitamente associado*; V + Adv – *opor-se categoricamente, amar perdidamente*; A + SP [prep + N] – (pt.) *louco de ciúmes*; V + SP [prep + N] – *estar em crise, entrar em funcionamento, ficar em pânico*³⁰ (Athayde, 2009).

Neste âmbito, Hausmann (2004: 316) tem também em consideração as estruturas tripartidas (“Tripelkollokationen”), do tipo de *fazer um sorriso amarelo, sofrer uma derrota estrondosa, adoptar uma medida drástica*. Trata-se de construções que resultam da fusão de duas colocações – *sofrer uma derrota + derrota estrondosa, adoptar uma medida + medida drástica*, por exemplo –, associação essa que se tornou usual, isto é, que se convencionalizou (Athayde, 2009).

Por último, chamo a atenção para o facto das colocações, muitas vezes, se construírem de maneira atípica (por vezes arcaica) em relação às regras *standard* da língua. Um exemplo desta atipicidade é a ausência de determinante nas construções, do género de: *ter fome, ter sede*. No entanto, este desvio da norma não é sistemático, uma vez que se diz *perpetrar um delito*.

As colocações podem também demonstrar menos flexibilidade ao nível da distribuição sintáctica. É o caso de colocativos epítetos, que têm uma função predicativa. Estes não podem aparecer como atributos, mas nas construções livres, este contexto é praticamente sempre possível (*Jean a eu une peur bleue. Jean a ressenti une peur immense. *La peur de Jean a été bleue. La peur de Jean a été*

³⁰ Estas últimas expressões são subcategorizáveis como “construções com verbo-suporte” ((al.) *Funktionsverbgefüge*), fraseologias cujos constituintes nominais exprimem um conteúdo semântico-lexical, integrando a construção numa dada área temática (um campo semântico-lexical), e cujos componentes verbais (os verbos-suporte), para além de serem portadores de categorias gramaticais próprias da classe ‘verbo’, veiculam valores de *Aktionsart*, causatividade/não-causatividade e incluem a construção numa determinada classe de representação de estados de coisas (estados, processos, acções, etc.) (cf. Athayde, 1999; 2000; 2001).

immense (fr.)). Mas existem colocações verbais que admitem uma certa alternância sintáctica. Por exemplo, as construções de verbo suporte que descreveremos em mais detalhe na próxima secção.

Mais uma vez, podemos concluir que a fixidez sintáctica não pode ser critério de distinção das combinatórias livres (Zuluaga, 2002: 61).

2.3. Tipologia de colocações

Tendo por base o que foi dito em capítulos anteriores e usando a tipologia de Hausmann, podemos dizer que existem palavras que têm uma grande capacidade de combinabilidade, outras cujas restrições de seleção podem ser explicadas através de características semânticas e, finalmente, outras que ocorrem apenas com certas palavras. O carácter gradativo das combinações é uma característica aceite por todos os estudiosos, divergindo apenas a nomenclatura: Sinclair (1966) diferencia entre *casual* e *significant collocations*, Cowie (1978) *open* e *restricted*, Aisenstadt (1981) *free* e *restricted*, Carter (1987) destaca quatro níveis: *unrestricted*, *semi-restricted*, *familiar* e *restricted* (*apud* Herbst, 1996: 385).

Atenemos agora na classificação de colocações que Cowie (1986) propõe. Este autor distingue, dentro das colocações, as *colocações livres* (ou *abertas*) e as *colocações restritas*. As primeiras são combinações facultativas de palavras cujos elementos constituintes mantêm o seu significado. Estes podem ser substituídos, sem alteração semântica nos outros elementos. Por exemplo, uma pessoa pode comer arroz, pudim, bolo ou chocolate. A possibilidade de complementos diretos do verbo *comer* é praticamente infinita. Semelhantemente, *comer* pode ser substituída por uma série de sinónimos, como *devorar*. As colocações restritas podem ser definidas como combinações de duas ou mais palavras usadas em um dos seus significados regulares e não-idiomáticos, seguindo certos padrões estruturais. São restritas na comutabilidade, não só através da valência gramatical e semântica (como os constituintes das combinatórias livres), mas também através do uso (Aisenstadt 1979, 71 *apud* Pöll, 1996: 19).

Para determinar se uma combinação é uma colocação restrita, Aisenstadt

(1979 *apud* Cowie, 1992) elenca alguns testes que podem ajudar na identificação destas, no caso de colocações **V + N**.

O primeiro teste está relacionado com a especialidade semântica do verbo. O verbo pode ser deslexicalizado, tal como *fazer* em *fazer uma proposta* (Moon, 1988:109 *apud* Cowie, 1992). A unidade da colocação é demonstrada pela existência de um verbo equivalente (*propor*). Em outros casos o verbo pode ter um sentido figurado, mas que, de tanto uso, perdeu a sua força analógica (*tirar uma fotografia*). Noutros casos, o verbo é usado num sentido técnico ou semi-técnico (*instaurar um processo, intentar uma ação ou interpor um recurso*).

O segundo teste é um teste à comutação, feito para demonstrar se o verbo é o único que pode ser usado num determinado contexto ou se, pelo contrário, existem uma série de sinónimos. No caso de *abrir/ iniciar uma sessão* trata-se de um verbo comutável. Já em *tirar uma fotografia* trata-se de um verbo não comutável.

Por último, podemos-nos interrogar se o nome, em relação ao verbo, é a única possibilidade ou se existem outros itens relacionados que o poderiam substituir (*resolver um problema/ uma questão*) (Cowie, 1992: 5-7).

Uma sub-classe de colocações restritas são as chamadas *delexical collocations*. Tipicamente, são constituídas por um verbo gramaticalizado e um complemento direto. O verbo pertence a uma classe de verbos muito frequentes, como *ter, fazer, tomar, dar*, entre outros. São exemplos, *fazer a cama, dar um conselho, fazer uma reclamação, tomar banho*. Seguindo a designação de Gross (1981), muitos linguistas chamam-lhes “construções de verbo suporte”, uma vez que a função do verbo é dar suporte ao nome com o qual co-ocorre, estabelecendo, assim, a ligação entre o nome e o sujeito da frase. As construções de verbo suporte são expressões analíticas complexas que se integram, pela natureza do seu núcleo predicativo, no domínio dos predicados nominais.

Vejamos o processo de formação das construções de verbo suporte: o componente verbal é o portador de tempo, número, pessoa e modo, existe um enfraquecimento lexical (entre verbo auxiliar e verbo pleno), mas não se encontra totalmente esvaziado de significado e deixa de poder ser substituído (**pôr atenção/ dar atenção*). O componente nominal torna-se o núcleo do conteúdo lexical de toda a construção, embora não desempenhe a função proposicional da referência, ou

seja, não é um argumento/complemento convocado pelo constituinte verbal. O seu conteúdo apresenta uma certa afinidade com a categoria semântica do verbo, pois trata-se geralmente de substantivos que expressam um processo, ação, estado. Estas perífrases lexicais têm em comum com os compostos serem combinações regulares e analisáveis, mas diferenciam-se destes por falta de capacidade formal, ou seja, não se podem intercalar elementos e/ou alterar a ordem dos componentes (*dar muita atenção, a atenção que se dá*). Estão a meio caminho entre a formação de lexemas compostos e sintagmáticos. Consideram-se colocações na medida em que conservam regularidade semântico-sintáctica e são consagrados pelo uso social (*pôr a mesa é uma colocação, dar uma mesa não*) (Zuluaga, 2002: 61).

Este tipo de colocações representa, sem dúvida, uma dificuldade para os aprendentes de línguas. Isto porque sobre a variação dos verbos suporte imperam fortes restrições, que se traduzem no facto de construções, aparentemente possíveis face ao sistema, não serem permitidas pela norma (*estar na dúvida, *estar na certeza*). Por outras palavras, há lacunas, no domínio das construções de verbo-suporte, não explicáveis sistematicamente (Athayde, 2001: 17), o que faz com que sejam um desafio para o aluno estrangeiro.

Cowie chama ainda atenção para um outro grupo de colocações que também trazem grandes problemas aos aprendentes de língua estrangeira: as *overlapping collocations*. Por exemplo, o verbo *quench*, em inglês, ocorre com os nomes *fire* e *thirst*. O verbo *extinguish* também pode ser usado com *fire*, mas não com *thirst*. Ao contrário, o verbo *slake* ocorre com *thirst*, mas *slake fire* não é uma combinação aceitável (1986 *apud* Fontenelle, 1994: 43-44). Este autor divide ainda as colocações em dois grandes grupos: as colocações lexicais e as colocações gramaticais. As colocações lexicais são uma classe aberta (não-finita) de, por exemplo, um verbo e um nome, ou um adjetivo e um nome. As *colocações gramaticais*, ao contrário das primeiras, têm um elemento de uma classe aberta e um elemento de uma classe fechada (tipicamente mas não necessariamente, uma preposição). Por exemplo, o verbo *depend* ocorre com a preposição *de* e não *em*³¹.

Em seguida, apresentaremos a perspectiva de Mel'cuk (1992: 182). Este autor

³¹ Alguns linguistas consideram também como colocação gramatical, as várias construções sintáticas (verbos que requerem *-ing form* ou infinitivo) (Fontenelle, 1994: 44).

descreve no *Dictionnaire Explicatif et Combinatoire* que uma colocação, ou semi-frasema, é uma combinação de dois ou mais lexemas **A** e **B**, cujo significante é a soma 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 selecionado livremente (Sanromán, 2001: 170). O significado da colocação inclui intacto o significado de um dos seus dois lexemas constituintes, neste caso **A**. **A** é livremente escolhido pelo falante, estritamente por causa do seu significado. Mas o outro componente do seu significado, isto é **C** é problemático: é expresso por **B**, é escolhido dependendo de **A** (isto faz com que o significado de uma colocação não seja estritamente construído).

A formulação "**B** expressa **C** dependendo de **A**" cobre os quatro tipos mais importantes de colocações (Mel'čuk, (2001 [1998]): 30):

"1) EITHER '**C**' $\neq$ '**B**' i.e. **B** 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 selected by **A** to support it in a particular 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 with **A** by any otherwise possible synonym of **B**;

OR *b*. '**B**' includes (an important part of) the signified '**A**', that is, it is utterly specific, and thus **B** is "bound" by **A**].

Nos casos 1), o significado de **B** não é o mesmo que se encontra no dicionário. No caso de 1) *a*, **B** encontra-se vazio de significado e serve apenas como suporte de **A** que conserva o seu significado. Exemplos deste caso, são as construções com verbo suporte, *dar um conselho*, *tomar em consideração* ou *fazer um favor*. Em 1) *b* **B** não se encontra vazio, mas **B** apenas expressa **C** em combinação com **A**, ou com um conjunto pequeno de outros lexemas. Por exemplo, na expressão

sorriso amarelo, *amarelo* é sinónimo de *forçado*, apenas neste determinado contexto. Outros exemplos deste tipo de colocações são: *ponte levadiça*, *chave mestra* ou *parede mestra*.

Nas colocações número 2), B mantém o mesmo significado que apresenta no dicionário, mas no caso 2) *a*, este significado só é atualizado em combinação com o lexema A. Atentemos na expressão *ódio mortal*. Neste contexto *mortal* significa *intenso*, mas fora deste contexto específico não pode ser utilizado como sinónimo de *intenso*, mas também não pode ser substituído por outro sinónimo, por exemplo **ódio letal*. Outros exemplos de expressões semelhantes são: *café forte*, *profundamente comovido* ou *fumador inveterado*. Em 2) *b*, C (“curvo como bico de águia”) aparece no dicionário como acepção de B, mas só em combinação com o lexema A. Na colocação, *nariz aquilino*, *nariz* está incluído semanticamente na definição de *aquilino*, uma vez que este adjetivo apenas se usa com nariz. Outros exemplos deste tipo de colocação são: *manteiga rançosa*, *cavalo alazão*, *cabelo grisalho*.

O significado de “C”, expresso irregularmente (= *a* B) dependendo de A, corresponde às chamadas **Funções Lexicais** (LF) (cf. 2.4). As colocações de tipo 1*a* e 2*a* incluem-se nas LF *standard* e as de tipo 1*b* e 2*b* seguem LF não-standard. O lexema A, que mantém o seu significado intacto, tem uma posição dominante dentro do significado da colocação e determina a expressão “C” através de B, é o argumento da LF correspondente; é chamada palavra chave da LF ou da colocação (Mel’čuk, 1995).

No entender de Grossman e Tutin (2002:5-6), dado que as colocações não constituem um grupo homogéneo de estruturas, uma vez que as suas realizações variam no *continuum* entre o mais e o menos fixo, estes autores agrupam-nas em subconjuntos de acordo com a forma como o seu significado se constitui:

- (i) o subconjunto das “**colocações opacas**” (*sorriso amarelo* (pt.), *piada picante* (pt.), *colère bleue* (fr.)), em que, não obstante o sentido interpretável da base, o colocativo – pouco produtivo – tem um significado diferente daquele que

apresenta fora da construção³²;

- (ii) as “**colocações transparentes**” (*lavar os dentes* (pt.) ou *heridos leves* (esp.), *grièvementement blessé* (fr.)) que se constituem como o caso prototípico de colocação – apesar de a escolha do colocativo não ser previsível, tornando difícil a codificação para falantes não-nativos, o colocativo apresenta o mesmo significado que carrega fora da colocação, facilitando o processo de decodificação;

- (iii) “**colocações regulares**” (*ano bissexto* (pt.), *nariz aquilino* (pt.), *cabelo louro* (pt.)) são aquelas que estão mais próximas das expressões livremente produzidas. A associação “base + colocativo” é interpretável, transparente e motivada, estando sujeita a restrições de seleção semântica (*nariz aquilino* (pt.), *manteiga rançosa* (pt.), *blondes Haar* (al.)). Usando a terminologia coseriana, poderíamos designá-las por “solidariedades lexicais”³³. Neste grupo incluem-se também combinatórias, cujo colocativo tem um sentido genérico (*grande tristeza* (pt.), *enorme alegria* (pt.), *grande tristesse* (fr.))³⁴.

2.4. Funções lexicais

Partindo do princípio que em inúmeras colocações a arbitrariedade não é total e que, muitas vezes, as associações lexicais favorecem um certo tipo de relações semânticas entre colocativo e base (como as séries de intensificadores de Bally (1909)), Mel’cuk (1995) tenta estabelecer padrões de combinabilidade de uma determinada palavra com outras palavras com que normalmente co-ocorre, para denotar um determinado significado. As funções lexicais foram inicialmente introduzidas por Žolkovskij e Mel’čuk no âmbito da Teoria Sentido-Texto³⁵. Estas funções são usadas para descrever fenómenos lexicais através do conceito de função

³² Trata-se de colocações semanticamente menos transparentes, sendo já possível considerá-las expressões semi-idiomáticas (na terminologia de Burger (2003) são consideradas “Teil-Idiome”).

³³ “La solidaridad es la relación entre dos lexemas (pertenecientes a campos diferentes) de los que el uno está comprendido, en parte o en su totalidad, en el otro, como rasgo distintivo (sema) que limita su combinabilidad” (Coseriu, 1977: 143-161).

³⁴ Este conjunto integra: sequências sintagmáticas como *ano bissexto*, *cabelo louro*, em que o sentido do colocativo só é ativado quando se combina com uma base específica; combinatórias cujo colocativo integra um paradigma que parece poder ser definido com o auxílio de traços semânticos – *grande* ou *enorme*, por exemplo, são considerados pelos autores intensificadores *standard* dos N que designam emoção (*grande tristeza*; *enorme alegria*). Cf., a este propósito, Athayde (2007: 113).

³⁵ Para uma descrição completa desta teoria, ver Mel’čuk (1997).

no sentido matemático. As funções lexicais podem ser definidas como relações de significado entre uma palavra-chave e outras palavras ou combinações fraseológicas de palavras. A forma genérica desta função é $f(X) = Y$, onde X é a palavra-chave e Y é o colocativo selecionado para expressar o significado denotado por $f(X)$. Um exemplo de função que significa “muito, em grande quantidade, grande nível”, é a função lexical *Magn*. Aplicando esta função ao nome *fumador*, chegamos ao adjetivo *inveterado*, porque um fumador inveterado é uma pessoa que fuma muito e *inveterado* tem uma relação privilegiada com o nome *fumador*, excluindo outros adjetivos. Esta relação pode ser exemplificada da seguinte maneira:

Magn (fumador) = inveterado

Semelhantemente, podemos dizer *Magn* (chuva) = torrencial.

Na opinião de Mel’cuk, é possível descrever as co-ocorrência mais típicas em 50 funções lexicais. Por exemplo, **Son** (típico de), **Oper** (verbo transitivo vazio, tornando-se a palavra-chave o objecto), **Liqu** (destruição de), **Mult** (multitude de), **Sing** (entidade única). Os seguintes exemplos ilustram o potencial destas funções:

Son (cão) = ladrar

(cavalo) = relinchar

Oper (suicídio) = cometer

(pressão) = exercer

Liqu (casamento) = anular

(apagar) = ficheiro

Mult (peixe) = cardume

(ovelhas) = rebanho

Sing (areia) = grão

(fatia) = bolo

Apesar de muitas das funções lexicais quererem capturar as propriedades

colocacionais, algumas focam apenas relações semântico-lexicais básicas, como a sinonímia, antonímia, hiponímia. Assim, as funções lexicais abarcam todas as relações sintagmáticas (vertical) e paradigmáticas (horizontal) da palavra entrada no dicionário. No entanto, este tipo de descrição tornaria cada entrada no dicionário imensa (Fontenelle, 1994: 46).

2.5. Relações paradigmáticas e sintagmáticas

Sobre a questão das relações paradigmáticas, Halliday (1996^a *apud* Corpas Pastor, 1996: 77) advertiu para a existência de séries lexicais formadas por palavras que apresentam o mesmo traço colocacional. Estas séries, apresentadas pelo autor (Halliday, 1996b: 151 *apud* Corpas Pastor, 1996: 77), podem ser gramaticalmente heterogêneas: *argued strongly, the strenght of his argument, his argument was strengthened* (ing.), etc. Mitchell (1971: 51 *apud* Corpas Pastor, 1996: 77) dá os seguintes exemplos: *He is a heavy drinker, He is putting in some heavy drinking e He is drinking pretty heavily*. Nestes casos o que se “coloca” é o significado lexical das unidades correspondentes. Considerações deste tipo levam Mitchell (1971 *apud* Corpas Pastor, 1996: 77) a afirmar que não são as unidades lexicais que se colocam, mas os significados lexicais destas, aos quais ele chama *roots* e que representa graficamente com o símbolo matemático:

√HEAV-; √DRINK-.

Baseados no exemplo de Corpas Pastor (1996: 78) para o espanhol, encontramos o seguinte caso em português: √VEEMÊNCIA-; √PROTESTO-:

A acusação **protestou veementemente** contra a decisão.

É este **protesto veemente** que quero aqui deixar em nome da bancada do Partido Social-Democrata.

Recentemente, a APT **protestou com veemência** contra as retenções das devoluções do IVA às empresas.

Exemplos retirados do *Corpus de Referência do Português Contemporâneo*.

Apesar da frequente semelhança semântica entre os elementos integrantes de uma série lexical, existem co-hipónimos relacionados com os mesmo hiperónimo que nem sempre compartilham a mesma extensão colocacional. Por exemplo, *prestar* coloca-se com *atenção*, mas não com o sinónimo *concentração*.

Terminaremos esta seção com a perspetiva sintagmática que, segundo alguns autores, é a que mais estreitamente relacionada está com a noção de colocação (Corpas Pastor, 1996: 78). A extensão colocacional de uma palavra determina significativamente os seus padrões lexicais combinatórios, ou seja, as relações sintagmáticas da dita unidade. Uma das questões essenciais a este respeito é o conceito de distância colocacional (*collocational span*). Segundo Hausmann (1985: 127), a distância entre os colocados indica o grau de proximidade sintagmática entre os componentes de uma colocação.

Na literatura não existe um consenso em relação à distância entre as unidades lexicais para que estas possam ser consideradas uma colocação. No entanto existem várias perspetivas, Greenbaum (1988: 114 *apud* Corpas Pastor, 1996: 78) considera que os elementos da colocação nem sequer necessitam de aparecer na mesma frase, podendo surgir em frases ditas por vários falantes num diálogo. Jones e Sinclair (1974 *apud* Corpas Pastor, 1996: 78) estabelecem quatro posições à direita ou à esquerda do núcleo. Berry-Rogghe (1973 *apud* Corpas Pastor, 1996: 78) também aplica as quatro posições mas usa apenas duas no caso dos adjetivos. Haskell (1972 *apud* Corpas Pastor, 1996: 78) aplica uma distância de três, Al-Madi (1986 *apud* Corpas Pastor, 1996: 78) de um, mas extensível a cinco ou seis, dependendo dos casos. Smadja (1989 *apud* Corpas Pastor, 1996: 78) utiliza cinco posições à esquerda.

Por outro lado, a proximidade sintagmática ou a *mutual expectancy* entre os colocados de que fala Firth (1968^a *apud* Corpas Pastor, 1996: 78) pode também dever-se a causas extralinguísticas. Neste sentido, as restrições podem ser determinadas pela relação de proximidade das entidades denotadas no mundo real. Quanto maior for a associação no mundo físico, maior a frequência de ocorrência apresentarão na língua (Schwalm, 1991 *apud* Corpas Pastor, 1996: 79). No exemplo, em espanhol, de Corpas Pastor (1996: 79) o verbo *glasear* (cobrir com glacé) surge

apenas com nomes, como *biscocho* ou *tarta*. Se esta técnica pudesse ser aplicada a outros produtos, certamente que a capacidade combinatória deste verbo se ampliaria.

As restrições entre os elementos depende também das convenções linguísticas a dois níveis. No primeiro nível, encontram-se as convenções estabelecidas a partir das regras do sistema, do tipo /+hum/ em, por exemplo, *A lâmpada tem sonhos. Estas restrições correspondem *grosso modo* às regras de subcategorização e às regras de restrição seletiva da gramática generativa. Porém estas restrições não conseguem explicar as idiosincrasias de combinações como, *momento crucial* ou *cancelar uma reserva*. Está pois em jogo o segundo nível de restrições: o nível colocacional que é determinado por preferências de uso específicas.

Neste capítulo, começámos por definir colocação. Seguidamente, discutimos algumas das suas principais características, como o seu estatuto de fronteira entre as estruturas fixas e livres, a sua transparência semântica e imprevisibilidade, discutimos a função da recorrência e, por último, dilucidámos o estatuto dos seus componentes e as configurações sintáticas que podem apresentar. Na secção 2.4., apresentámos as tipologias mais conhecidas para classificar o fenómeno colocacional. Nas duas últimas secções, expusemos brevemente as funções lexicais preconizadas por Mel'čuk e o papel das relações sintagmáticas e paradigmáticas na construção das colocações.

Capítulo 3. As colocações em *corpora* de aprendizagem de L2

Neste capítulo, vamos apresentar alguns dos trabalhos mais pertinentes sobre o uso de colocações em *corpora* produzidos por aprendentes de uma língua estrangeira (secção 3.2). Esses estudos serão descritos em termos da definição de colocação que utilizam (secção 3.3), do processo de anotação (secção 3.4) e da taxonomia de erros que usam para classificar os problemas encontrados nas colocações (secção 3.4.1). Abordaremos também, numa subsecção, as dificuldades que se apresentam nesta tarefa (secção 3.4.2).

Antes da descrição desses trabalhos, na secção 3.1, fazemos uma pequena reflexão sobre o papel que as colocações assumem no ensino/aprendizagem de uma L2³⁶, se estas devem ser ensinadas e quais os materiais de ensino mais apropriados a esta tarefa.

³⁶ Genericamente, o termo L2 é usado para classificar “a aprendizagem e uso de uma língua não-nativa dentro de fronteiras territoriais em que ela tem uma função reconhecida” (por exemplo, habitantes de ex-colónias africanas que têm como L1 uma língua africana e como L2 Português) (Leiria, 2004:1); o termo LE é usado para designar a aprendizagem (normalmente formal) e uso de uma língua em espaços onde essa língua não tem qualquer estatuto sociopolítico (por exemplo, o ensino do português como disciplina curricular na Alemanha). No entanto, no casos dos alunos tidos em consideração para esta investigação, tal distinção não é a mais apropriada, uma vez que a origem dos dados é variada. Temos participantes que decidiram passar algum tempo num país estrangeiro, neste caso Portugal, e aprender português. Encontram-se, pois, numa situação de imersão, combinada com o ensino formal da língua, de modo a acelerar o processo de aquisição. Participam também neste estudo alunos que frequentam cursos de português no estrangeiro e alunos que tinham aulas online. Dada a variedade de contextos, considerou-se ser mais adequado adoptar o termo L2, visto que pode ser genericamente usado para referir qualquer língua que se aprenda depois da primeira língua (L1) ou língua materna (LM), independentemente do estatuto desse idioma em relação a quem a aprende ou ao contexto sociolinguístico em que está a ser aprendido (Richards, 1987: 7 *apud*, Leiria, 2004 :7).

3.1. O papel das colocações no ensino de L2

Words shall be known by the company they keep. (Firth, J. R. 1957: 11)

The meaning of a word is its use in the language. (Wittgenstein 1958: 20)

As palavras não co-ocorrem livremente, existindo vários níveis de restrições combinatórias que variam de língua para língua e que são determinados historicamente. Não existe um regulamento no sistema de uma língua que possa ser usado como regra universal, parecendo a arbitrariedade deste fenómeno empírica para os aprendentes não-maternos. Segundo Rudzka (1981), uma das maiores dificuldades com que se deparam os alunos de uma língua estrangeira, em relação ao vocabulário, é conhecer as suas propriedades de colocabilidade (*apud* Bahns, 1993: 56). Assim, tal como sugere Irsula Peña (1994: 48) e Cruse (2000: 76), os alunos estrangeiros terão que aprender estas unidades plurilexémicas como um todo. Também Howarth partilha a mesma opinião, sublinhando que estas estruturas assumem um papel preponderante no domínio de uma língua estrangeira, particularmente em níveis de proficiência mais altos:

Teaching for the most part leads learners to believe that, in producing stretches of English, they combine items of vocabulary according to their knowledge of syntactic structure. As proficiency develops, those learners additionally come into contact with and memorize an increasing number of idioms, which are learned as fully lexicalized fixed expressions to be used when appropriate. (Howarth, 2001 [1998]: 186)

Porém, o fenómeno fraseológico é composto por um variadíssimo número de estruturas; desde as mais idiomáticas e fixas, às menos idiomáticas e mais livres. As expressões mais idiomáticas são, por definição, menos transparentes e, portanto, será mais complexo apreender o seu significado. No entanto, a sua utilização parece ser menos frequente em comparação com estruturas mais transparentes:

*Idioms are difficult on the perception level, but since they do not appear too frequently in everyday situations the student can quite affectively do without them in his performance. (...) Collocations, however, occur practically in every utterance or sentence. (Korosadowicz-Struzsyńska, 1980:114 *apud* Howarth, 1998b: 18)*

Como sublinha, Howarth, dentro do domínio fraseológico, as colocações assumem-se como fenómeno especialmente frequente e, como tal, de crucial importância para o ensino/aprendizagem de uma língua estrangeira:

Moreover, learners need to understand that restricted collocations make up a significant part of a typical native speaker's production in both speech and writing. Conforming to the native stylistic norms for a particular register entails not only making appropriate grammatical and lexical choices but also selecting conventional collocations to an appropriate extent. Such combinations are not optional stylistic adornments on the surface of text; they are essential for effective communication (...). (Howarth, 2001 [1998]: 186)

Uma proficiência linguística semelhante à de um nativo depende crucialmente do conhecimento de um determinado número de construções “pré-fabricadas” (Cowie 1994: 3168 *apud* Herbst, 1996: 389). Infelizmente, para o aluno estrangeiro não existe nenhuma razão que o leve a construir a colocação *café fraco* em vez de *café débil* (Mackin, 1978: 150 *apud* Herbst, 1996: 386). De um número enorme de expressões perfeitamente possíveis do ponto de vista gramatical e que parecem aceitáveis e naturais, o aluno tem de conhecer as restrições para saber quais são as combinações corretas e quais aquelas que apenas o parecem. Este tipo de erro, resultante da falta de competência fraseológica não recebeu ainda muita atenção por parte dos linguistas (Howarth, 2001 [1998]: 161), apesar de, segundo Gabrys-Biskup (1992: 87), a maior parte dos erros cometidos pelos alunos de L2 serem erros em colocações.

Também os professores têm dificuldade em explicar porque se diz *reach a conclusion*, mas *arrive at findings* ou *produce findings* e não **reach findings*. Tentando estabelecer um paralelo semelhante, em português, dizemos *chegar a uma conclusão*, *chegar a um resultado*, mas *tirar ilações* e não **chegar a ilações*. Como foi já dito, não se trata de um tipo de erro causado por uma infracção das regras gramaticais ou semânticas da língua, mas de um desvio das regras fraseológicas (Howarth, 2001 [1998]: 162). Este aspeto da competência linguística é difícil de prever e os problemas que coloca derivam tanto das características internas da língua alvo, como de interferência da língua materna. Por outro lado, a predominância de técnicas comunicativas no ensino da língua levou a que se desse mais ênfase ao uso criativo do vocabulário para a negociação de significados e se

dedicasse pouco interesse às estruturas pré-fabricadas. Cowie (1992:10-11) critica este método de ensino que incentiva a comunicação, mas que não ensina primeiramente as bases para que possa existir comunicação (“communicate without available resources”).

Em outros casos, os professores de línguas centram-se nos dois extremos do espectro: combinatórias livres e expressões idiomáticas, fazendo crer aos alunos que existem apenas dois modos distintos de construções: a aplicação de regras generativas lexicais às combinatórias livres e o completo “congelamento” das expressões fixas. Esquecem-se das colocações restritas. “Os professores encontram-se pouco informados sobre o fenómeno fraseológico”, sublinha Howarth (2001 [1998]: 186) reportando-se ao caso concreto da aprendizagem do Inglês como língua estrangeira, numa reflexão que se poderia estender à aprendizagem de uma qualquer outra língua.

Porém, começou a surgir uma maior consciência da prevalência de combinações memorizadas ou *ready-made* na linguagem escrita e falada, e um reconhecimento do papel crucial que elas apresentam na aquisição das línguas e na produção da linguagem adulta (Pawley e Syder, 1983; Peters, 1983 *apud* Cowie, 2001[1998]: 1). Hill (2002 *apud* Alonso, Ramos, 2003) cunha o termo “competência colocacional” e Philip (2007: 13) chama a atenção para o facto de que, muitas vezes, os alunos conhecem uma palavra, mas o seu conhecimento é imperfeito ao nível colocacional:

Current pedagogical practice is taking collocation far more seriously than it used to, but it is often restricted to the teaching of new vocabulary rather than the teaching of new meanings of already-learned items. Figurative meanings and delexical forms abound in the language, and when they are used wrongly, the flow of meaning is interrupted, resulting in incomprehension, misunderstanding and, at times, embarrassment or unintended humour. (Philip, 2007: 10)

Podemos concluir que apesar do crescente interesse pela aprendizagem de colocações no ensino das línguas estrangeiras, há ainda bastante espaço para mais investigações nesta área, especialmente para línguas como o Português.

3.1.1. Devem as colocações ser ensinadas?

Perante a situação descrita nas secções anteriores, e tendo em conta que a linguagem estereotipada – ainda que, regra geral, usada de forma não-consciente – ocupa uma parte importante do discurso (falado e escrito) de um falante nativo, qual o valor que deve assumir o ensino-aprendizagem das colocações numa L2?

Existem duas visões sobre esta questão: os investigadores que não acreditam ser possível ensinar este tipo de estruturas e os que, pelo contrário, advogam uma metodologia de ensino.

Ronald Mackin (1978 *apud* Bahns, 1993: 59) insere-se na primeira categoria. Este autor é cético em relação à possibilidade de se ensinarem colocações. Um aprendente de Inglês L2 não pode pensar que conseguirá aprender todas as colocações, pois são muitas. “They are in any case so numerous as to rule out any methodical teaching or acquisition of them”. Apenas com anos de estudo, leitura e observação da língua, os alunos podem adquirir algum grau de competência colocacional.

Já Joanna Channel (1981 *apud* Bahns, 1993: 58), investigadora na área do Inglês Língua Segunda, enfatiza a necessidade de se ensinarem colocações. Waldemar Marton (1977: 43 *apud* Bahns, 1993: 58) partilha a mesma opinião e considera que a mera exposição à língua alvo, não é suficiente para um aluno avançado adquirir conhecimento colocacional. “Se os professores querem guiar os alunos de nível avançado, para um conhecimento quase-nativo da língua estrangeira, têm que prestar mais atenção à aprendizagem dos sintagmas convencionalizados”.

Na perspetiva de Sanromán (2001: 167), a aprendizagem ou ensino de uma L2 constitui-se como um momento de tomada de consciência dos constrangimentos, fundamentalmente de natureza lexical, a que a atividade linguística está sujeita. O que se diz de uma certa maneira numa língua, não se diz, forçosamente, da mesma maneira noutra língua. No que respeita, concretamente, ao fenómeno das colocações, só a comparação de diferentes línguas mostra, com clareza, como a escolha do colocativo não é livre, mas imposta (Sanromán, 2001: 167). Assim, talvez não seja necessário aprender todas as colocações, mas aprendê-las de uma maneira

contrastiva. Assim, as colocações a serem ensinadas devem ser escolhidos tendo como base a língua materna dos aprendentes e devem ser ensinadas fazendo referência à LM. Através de uma metodologia contrastiva, os alunos vão estar mais despertos para encontrar as diferenças e semelhanças³⁷. Também para Marton (1977: 40 *apud* Bahns, 1993: 59) a comparação é melhor maneira de colocar o problema das colocações: “A noção de convencionalidade sintagmática ganha a sua total dimensão, só quando é considerada do ponto de vista contrastivo, ou seja, quando a língua materna e a língua alvo são tidas em consideração. Problemas reais de aprendizagem são causados pelos sintagmas, que do ponto de vista da comparação, lexicalmente não-congruentes, ou, por outras palavras, pelos sintagmas que não têm uma tradução direta equivalente entre os elementos correspondentes.” Se pensarmos no caso do Inglês e Alemão, e no sub-tipo de colocação V + N, uma análise contrastiva demonstra que num grande número de casos, existe uma tradução equivalente direta. Como demonstra a comparação de colocações nestas duas línguas, feita por Bahns, nos casos onde não existe tradução direta, existe uma maior probabilidade de erro (1993: 59-60). Se o professor conhecer a LM dos alunos, este pode explicitar as combinações que não demonstram congruência na língua materna do aluno (estas podem ser uma fonte de erro) e chamar especial atenção para as colocações semanticamente possíveis, mas que não são colocações corretas, especialmente se o forem na LM do aluno (Nesselhauf, 2002: 239).

Segundo Howarth, (1998b: 20-21) é errado achar que a maioria das colocações são co-ocorrências arbitrárias que devem ser aprendidas como um todo. Um estudo comparativo entre nativos e não-nativos, demonstrou que os alunos mais avançados já interiorizaram as colocações mais lexicalizadas. Sobram um grande número de combinatórias menos restritas, demasiadas para serem aprendidas como itens lexicais isolados. Na perspectiva deste autor, é mais eficiente ensinar a natureza do fenómeno e desenvolver consciência. É o bloqueio dos paradigmas que é arbitrário, não a co-ocorrência, ao contrário do que diz a abordagem neo-firthiana

³⁷ Devem ser também referidas as combinatórias livres, visto que nem sempre existe um equivalente colocacional na outra língua do aprendente ou vice-versa.

que sugere ser uma questão de probabilidade. A competência colocacional reflete um elevado nível de proficiência e, apesar deste tipo de erro não ser considerado sério por muitos, demonstra uma expressão escrita fraca e à qual falta clareza. Não se pode comparar a utilização das colocações com o uso de recursos estilísticos que embelezam um texto, trata-se de uma questão de precisão

Nesselhauf, tal como Howarth, é da opinião que os professores têm que dar a conhecer aos alunos o fenómeno colocativo. No entanto, segundo este, não basta. Algumas colocações têm que ser ensinadas e aprendidas explicitamente. Mas como estamos a tratar de uma área de uma enorme vastidão, é necessário usar critérios para definir quais devem ser aprendidas. Nesselhauf (2003: 238) propõe o critério da aceitabilidade, da frequência, a escolha de um registo neutral (ou outro que os alunos usem em especial, como um registo académico), a congruência (a não-congruência deve receber uma especial atenção, mas as congruentes não podem ser negligenciadas, pois também há erros mesmo em estruturas congruentes) e, por último, restrição (quanto mais restritas, mais problemas surgem). Para Lewis (2000), para perceber quais as colocações que devem ser ensinados é fundamental identificar os problemas que os alunos têm quando se deparam com as colocações. Sendo as colocações estruturas, relativamente, transparentes, os problemas não residem na compreensão das mesmas, mas na sua produção. Seguindo esta linha de investigação, existem já alguns trabalhos sobre a produção de colocações por parte de não-nativos (Gabrys-Biskup, 1992; Bahns e Eldaw 1993; Howarth 1998; Granger 1998).

Por último, Hausmann (1984: 404 e ss), aqui exemplificado e explicado pelas palavras de Athayde (2009: 19), propõe que o ensino/aprendizagem de colocações em L2 seja levado a cabo por quatro vias, entendidas como complementares: (i) **via onomasiológica**: aprender, partindo da base. O objectivo deste método é tornar os alunos capazes de utilizar em contextos adequados, palavras banais mas frequentes da linguagem quotidiana, cujo significado não traz, portanto, dificuldades de maior ao aprendente de uma L2. Tomando como exemplo a palavra *mesa*, o aluno tomaria conhecimento de um conjunto de expressões em que tal N ocorre em relação privilegiada: {*pôr + a mesa*} ou {*levantar + a mesa*}; (ii) **via semasiológica**: aprender, partindo do colocativo. O conteúdo semântico de um colocativo como “inveterado”

será mais facilmente apreendido por aprendentes de Português como L2 se se fornecer ao aluno as suas possibilidades combinatórias, em colocações como *fumador inveterado*, *alcoólico inveterado* ou *solteirão inveterado*, por exemplo; (iii) **trabalho com colocações em duas línguas**: se, como vimos, as bases têm equivalentes nas diferentes línguas, mas os colocativos não são, em regra, transponíveis de língua para língua, faz sentido um trabalho “bilingue” em sala de aula. Tal é também defendido pelos autores já mencionados Marton (1977 *apud* Bahns, 1993) e Sanromán (2001). De acordo com Hausmann (1984: 405), os alunos terão de perceber que só há equivalência a nível da colocação e não a nível dos colocativos e (iv) **trabalho com colocações em textos e no dicionário**: é importante levar o aluno a procurar colocações em textos e fazê-lo encontrar as colocações equivalentes na sua língua materna; através do trabalho com o dicionário bilingue, o aluno deverá passar a reconhecer que a unidade de tradução é, em muitos casos, a colocação. Levado a constatar as limitações dos dicionários bilingues – que nem sempre fornecem informação colocacional adequada, uma vez que o lema não é geralmente a base, como se esperaria e como seria útil, mas o colocativo e ainda porque traduzem frequentemente palavras e não colocações –, o aluno de L2 passará a ter uma relação mais esclarecida com esse instrumento de trabalho, evitando erros de utilização. Um último parêntesis para referir Irsula Peña (1994: 142) que compara a falta de informação colocacional ou o carácter incompleto dessas indicações em dicionários bilingues a um medicamento que é tomado sem a prescrição de um médico: em vez de curar, pode causar danos ainda maiores.

Na seguinte secção, vamos justamente centrar-nos nos dicionários e materiais didáticos e nas suas abordagens ao tratamento das colocações.

3.1.2. Materiais de ensino de colocações

Na maioria dos manuais escolares e material de apoio ao ensino-aprendizagem, a dimensão paradigmática do Léxico, isto é, questões de sinonímia, antonímia, relações dos derivados com a palavra-base, etc. têm sido sobrevalorizadas relativamente à dimensão sintagmática (que diz respeito às regras combinatórias) (Hausmann, 1984: 406 *apud* Athayde, 2009: 12). Segundo Sinclair

(1991: 103), as gramáticas não providenciam explicações sobre as várias estruturas onde pode surgir a mesma palavra, no entanto alguns dicionários pedagógicos começam a incluir esse tipo de informação essencial para o domínio de uma língua estrangeira.

A natureza imprevisível (idiossincrática) das colocações faz com que sejam adequadas para serem incluídas num tipo especial de dicionário, não tanto de decodificação (compreensão de textos, função dos dicionários tradicionais), mas para codificar textos (perspetiva da produção). Um dicionário colocacional responderia a perguntas, como *O que se pode juntar ao nome X?* ou *Que adjetivos tipicamente modificam o nome X?* Um dicionário colocacional poderia dar-nos a informação de que o nome *fumador* pode ser usado com o adjetivo *inveterado*. É o nome que determina o sentido do adjetivo e não o contrário. Para propósitos de codificação, faz sentido registar a colocação na entrada do nome, uma vez que é natural que um aluno estrangeiro se questione sobre como pode exprimir a ideia de que uma pessoa é um “grande fumador” (Fontenelle, 1994: 43).

A questão da frequência não é central, quando se trata de elaborar um dicionário. Muitas combinações de palavras, como *um vestido bonito*, são muito frequentes, no entanto, não são específicas, uma vez que o adjetivo *bonito* tem um potencial colocativo enorme e pode combinar-se quase com qualquer tipo de substantivo. O mesmo acontece com o substantivo *vestido*, que pode combinar-se com inúmeros adjetivos, como *feio*, *curto*, *comprido*, *azul*, etc. Esta falta de especificidade não justifica a inclusão desta combinatória num dicionário. Já combinações como *fumador inveterado*, apesar de serem menos frequentes que *um vestido bonito*, devem comparecer num dicionário colocacional, uma vez que apresenta um grande grau de especificidade, pois não existem outras possibilidades colocativas para dizer que uma pessoa é um grande fumador (Hausmann, 1985: 121-122).

Com o desenvolvimento de *corpora* em grande escala e o aumento da sofisticação das ferramentas de processamento de dados, alguns dicionários de intuito pedagógico começam já a seguir a perspetiva previamente apresentada. Relativamente ao Português, podemos referir o *Dicionário Contextual Básico da Língua Portuguesa* (2000) de Bernhard Pöll. Para o Inglês, existe o Oxford

Collocations Dictionary For Students Of English (2002). Para o Espanhol, a referência mais conhecida é o *Diccionario de colocaciones del Español*³⁸ que segue a linha de investigação de Benson *et al.* 1986, Hill e Lewis 1997, Crowther *et al.* 2002 e Mel'čuk e Polguère 2007.

No que toca ao material de ensino, Bahns (1993: 59-60) defende que, para o ensino das colocações, é uma mais valia o aluno ter acesso a uma grande variedade de material didático, como livros de exercícios e de textos. No entanto, quem produz este tipo de material didático, deve ter em mente, que as colocações a incluir devem ser diferentes, dependendo da língua materna a que se destinam³⁹. Como foi já mencionado, alguns estudos revelam que, no que diz respeito às colocações, os alunos tendem a usar a “hipótese da transferibilidade”, uma vez que a maioria dos erros colocacionais são causados pela influência da LM. Como refere Granger (2001: 158), os aprendentes de L2 não são “‘phraseologically virgin territory’: they have a whole stock of prefabs in their mother tongue which will inevitably play a role – both positive and negative – in the acquisition of prefabs in the L2”. Por esta razão é necessário distinguir as colocações que os alunos com uma particular LM “já sabem” (uma vez que há equivalência total), das colocações que, através de uma análise contrastiva, demonstram ser específicas da língua alvo e que, nesse caso, o aluno terá de aprender. Nesta perspetiva, a quantidade de material a incluir num livre de aprendizagem, seria consideravelmente menor.

Apesar de não existir uma metodologia didática unânime em relação ao ensino/aprendizagem das colocações, existem já bastantes investigações que se têm focado no estudo da produção de colocações por parte de alunos estrangeiros. Na seguinte secção, apresentamos alguns destes trabalhos que usam *corpora* de aprendizagem, isto é, textos que foram produzidos por aprendentes estrangeiros de uma língua, normalmente em contexto de sala de aula. São vários os propósitos deste tipo de estudo, mas vamos cingir-nos aos mais próximos da investigação que levamos a cabo nesta dissertação.

³⁸ <http://www.dicesp.com/>

³⁹ O caso das expressões idiomáticas é diferente. Todas devem ser explicitamente ensinadas, uma vez que os alunos não sabem quando existe equivalência total.

3.2. Colocações em *corpora* de L2

Vários trabalhos salientam a importância da criação, utilização e partilha de *corpora* de aprendizagem, sendo este género de *corpora* normalmente utilizado para o estudo e anotação de vários tipos de erros, nomeadamente, erros gramaticais, lexicais, entre outros (Dagneaux *et al.* 1998; Tono, 2003). Várias investigações sobre desvios na produção de colocações têm-se apoiado em *corpora* de aprendizagem, nomeadamente Granger (1998), Nesselhauf (2003, 2005) e Gilquin *et al.* (2007). Martelli (2006) é mesmo da opinião que a utilização de material real tem mais vantagens do que o uso de testes de licitação ou exercícios de tradução para melhor conhecermos os problemas dos alunos e para, assim, lhes proporcionarmos uma ajuda mais bem direcionada, para criar material didático mais especializado ou para a própria compilação de dicionários de aprendizagem.

Nesta secção, fazemos um levantamento de alguns trabalhos que têm como objeto a identificação e análise de colocações em *corpora* constituídos por textos escritos por alunos estrangeiros. Mencionamos investigações, como a de Rodríguez-Fernández *et al.* (2015) que usa *corpora* de aprendizagem para a criação de ferramentas automáticas que facilitem a aprendizagem das colocações por computador (secção 3.2.1). Seguidamente, na secção 3.2.2, apresentamos trabalhos que têm como objetivo a comparação da produção de alunos estrangeiros com a produção de nativos dessa mesma língua (Granger, 1998; Howarth, 1998; Prieto *et al.*, 2009). Na secção 3.2.3, descrevemos ainda investigações que se centram unicamente nas produções de alunos estrangeiros: usando exercícios, utilizando os *corpora* na sua totalidade ou focando-se numa construção sintática ou palavra em específico. Por último, na secção 3.2.4, tendo em consideração a língua portuguesa, analisamos o trabalho seminal de Leiria (2006) e o de Antunes e Mendes (2015).

3.2.1. Tratamento automático dos erros

Os principais objetivos da utilização de *corpora* de aprendizagem são compreender o processo de aprendizagem das colocações e estudar os erros que podem afetar a produção das colocações. Porém, existem também outras

finalidades para o uso deste tipo de *corpora*. O estudo de Rodríguez-Fernández *et al.* (2015), com base num *corpus* de alunos americanos aprendentes de Espanhol, utilizou uma tipologia de erros em colocações com o objetivo de realizar a correção automática desses erros. Vincze *et al.* (2011) e Liu (2002) focaram-se na utilização de *corpora* produzidos por aprendentes de modo a criar uma ferramenta automática para a correção de erros de colocações para aprendentes de Espanhol L2 e de Inglês L2, respectivamente. Alonso Ramos (2013) e Wanner *et al.* (2011) usaram *corpora* produzido por alunos de Espanhol (*Corpus Escrito del Español L2*) para a criação de uma ferramenta de aprendizagem online de colocações (*ColocaTe*) que pretende ser mais dinâmica que os recursos existentes. Estes trabalhos sugerem assim que a área da aprendizagem de línguas assistida por computador centrada nas colocações tem muito a ganhar com a utilização de dados provenientes de *corpora* de aprendizagem produzidos por alunos estrangeiros (Shei e Pain, 2000; Chang *et al.*, 2008).

3.2.2. Comparação com as produções de nativos

Existem também alguns estudos que, como metodologia de análise, comparam *corpora* de aprendentes com *corpora* produzidos por nativos⁴⁰. Salientamos o trabalho de Granger (1998), o de Howarth (1998) e o de Prieto *et al.* (2009).

O trabalho de Granger (1998) foi realizado com franceses aprendentes de inglês e centrou-se na extração semiautomática de colocações compostas por advérbios intensificadores terminados em *-ly* que funcionam como modificadores, tais como *perfectly natural*, *closely linked* ou *deeply in love*. A mesma tarefa foi realizada com um *corpus* produzido por nativos para assim se estabelecer uma comparação em relação às produções.

Howarth (1998) levou a cabo um estudo com alunos estrangeiros de vários *backgrounds* linguísticos que aprendiam inglês em universidades britânicas. O autor estabeleceu uma comparação entre as produções deste alunos e textos produzidos

⁴⁰ Existem também vários trabalhos sobre a utilização de colocações por parte de nativos. Podemos mencionar, por exemplo, os estudos em *corpora* falado de Altenberg (1993) e em *corpora* escrito de Cowie (1991, 1992).

por nativos. Os textos dos nativos foram retirados de livros, artigos de ciências sociais e humanas; os textos dos não nativos são composições de alunos de mestrado em linguística e ensino de Inglês como língua estrangeira, sendo alguns deles professores originários de vários países. Este *corpus* é composto por 10 composições com um total de 25.000 palavras. Usou-se a área das ciências sociais e procuraram-se colocações lexicais, uma vez que, segundo o autor, causam mais dificuldades do que as gramaticais, especificamente as colocações de V + N complemento (tais como, *adopt an approach, follow a procedure* ou *reach an agreement*). O objetivo desta investigação foi identificar e analisar fraseologias *não-standard* em textos académicos, de um mesmo registo, produzidos por não nativos e nativos e estabelecer, assim, uma comparação entre o desempenho dos dois grupos. O que diferencia este estudo de outros é a comparação que se estabelece entre alunos com diferentes línguas maternas e o facto de se tratar de um estudo realizado sobre textos completos e não usar, apenas, listas de colocações.

Prieto *et al.* (2009) têm como objetivo a comparação das colocações usadas por falantes nativos da língua espanhola e aprendentes da mesma. O *corpus* usado para esta tarefa tem um número semelhante de redações produzidas por nativos de inglês e não nativos sobre um determinado tópico, permitindo assim fazer uma análise qualitativa dos resultados obtidos e considerar os dados dos nativos como referência. Este estudo insere-se num projeto que, na primeira fase pretende fazer um diagnóstico – observar os problemas no uso das colocações para, num segundo momento, criar uma plataforma multimédia de ensino/aprendizagem. Dados sobre o grau de domínio de L2, língua materna e conhecimento de outras línguas, foram alguns dos dados recolhidos por este estudo. O *corpus* possui 400.000 palavras, 126 textos escritos por nativos e o mesmo número por estrangeiros (nível médio/avançado). Realizou-se uma análise quantitativa dos dados, mas chegaram à conclusão que não eram suficientes para levar a cabo uma comparação fidedigna.

3.2.3. Estudos com produções em L2

Nas seguintes subsecções, apresentamos alguns estudos desenvolvidos somente com produções de estrangeiros: trabalhos que se focam numa palavra em

específico ou numa determinada construção sintática e, por último, estudos que se baseiam em exercícios.

3.2.3.1. Corpora

Focando-se apenas em *corpora* de não-nativos, Thewissen (2008) utilizou um *subcorpus* do *International Corpus of Learner English*⁴¹, composto pelos aprendentes franceses, ingleses e espanhóis. Este *corpus* é constituído por 223 textos (152 363 palavras), do nível B1 ao nível C2. Esta classificação foi levada a cabo por dois avaliadores profissionais, de acordo com o *Quadro Europeu de Referência*⁴². Cada texto foi anotado com vários erros fraseológicos, nomeadamente colocações, por um linguista nativo. Cada erro anotado é acompanhado por uma etiqueta que descreve a natureza do erro e a sua correção. O objetivo deste estudo era descrever os tipos de erros fraseológicos, mas também perceber se o nível de proficiência e a LM têm influência na performance do aluno.

Philip (2007), no seu estudo sobre a produção de colocações, utilizou um *corpus* de 60.000 palavras, produzidas por estudantes de Inglês, na Universidade de Bolonha do nível B2-C1. Nenhum deles tinha conhecimento específico de linguística. Isto é importante, uma vez que alunos com esse tipo de conhecimento podem estar mais bem equipados para reconhecer linguagem figurativa ou fraseológica. Os textos foram produzidos na sala de aula, como trabalho de casa, e as composições foram feitas para uma parte da avaliação. Como os textos foram produzidos para uma tarefa específica, encontramos uma linguagem e ideias similares entre eles. Isto pode colmatar o problema do tamanho do *corpus*, uma vez que assim se garante a repetição e a identificação de padrões que tornam difícil criar generalizações.

3.2.3.2. Utilização de uma palavra ou construção sintática

Um outro tipo de investigação utiliza *corpora* de aprendizagem para se centrar num determinado verbo, como, por exemplo, o trabalho de Altenberg e

⁴¹ Um *corpus* composto por 16 línguas maternas, que se encontra etiquetado com 56 tipos de erros, dividido por línguas maternas e níveis de proficiência.

⁴² http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

Granger (2001) focado no uso fraseológico do verbo *make*. Outros trabalhos focam-se numa determinada construção sintática, como na investigação de Nesselhauf (2005) com alemães aprendentes de inglês avançado sobre o uso de combinações de V + N.

3.2.3.3. Exercícios

Por último, também Gabrys-Biskup (1992) se debruçou sobre questões semelhantes e levou a cabo uma avaliação para estimar a produção de colocações. No entanto, não usou *corpora* de aprendizagem, mas um exercício no qual os alunos tinham que encontrar equivalentes para as colocações da LM. As respostas foram avaliadas por 3 nativos numa escala de 1 a 4, de inaceitável a completamente equivalente. Apesar deste estudo não se centrar *tout court* em *corpora*, decidimos mencioná-lo uma vez que os objetivos dele são muito semelhantes aos estudos que se centram em *corpora*: encontrar a causa dos erros colocacionais e determinar o papel da LM, e estudar, comparativamente, alunos com uma língua materna mais próxima e outra mais distante do inglês.

3.2.4. Estudos em *corpora* de PL2

Considerando os trabalhos realizados com *corpora* de aprendizagem produzido por alunos aprendentes de Português Língua Segunda, destacamos os trabalhos de Leiria (2006) e Antunes e Mendes (2015). O primeiro centra-se na aquisição lexical de alunos de Português L2. A autora analisou um *corpus* de textos produzidos por alunos franceses, alemães, suecos e chineses. Neste material a autora encontrou “co-ocorrências privilegiadas” com um certo grau de fixidez, como *velhos amigos* ou *gastar dinheiro*, tratando-se, no nosso entender, de colocações. Estas estruturas foram avaliadas tendo em consideração se seriam ou não utilizadas por um falante nativo. O trabalho de Antunes e Mendes (2015), centram-se nas expressões multipalavra encontradas numa parte do *corpus* de Português L2⁴³. As autoras identificam diferentes tipos de expressões multipalavra, corretas e

⁴³ 1<http://www.clul.ul.pt/research-teams/547>

incorretas (incluindo colocações), produzidas pelos alunos estrangeiros. Seguidamente, caracterizam os erros encontrados de acordo com uma taxonomia estabelecida pelas mesmas. Antunes e Mendes (2015) concluem que as colocações são o tipo de expressão multpalavra que apresenta maior número de erros, independentemente da língua materna do aprendente. De acordo com as autoras “as colocações são particularmente difíceis para os aprendentes de Português L2, porque eles demonstram níveis de restrição que não são fáceis de apreender”. Considerando que existe pouca informação disponível em dicionários portugueses, pelo menos comparado com os recursos que existem para Inglês (Antunes e Mendes, 2015), o recurso a listas de colocações em língua corrente pode ser uma ferramenta útil para os alunos.

Depois da descrição de investigações sobre as colocações em *corpora* de aprendizagem, iremos, na próxima secção, refletir sobre as definições de colocação que são usadas nesses mesmos estudos. Esta é uma questão essencial, uma vez que, como foi já *supra* mencionado, não existe uma definição totalmente aceite por toda a comunidade.

3.3. Definição de colocação em estudos com *corpora* de L2

Citando Bahns (1993:57), “regrettably, collocation is a term which is used and understood in many diferente ways”. Apesar da veracidade desta afirmação, nesta secção, faremos um breve resumo das noções de colocação que são utilizadas em vários estudos de análise de erros de colocações em *corpora* de aprendizagem. Começamos por distinguir a definição contextual da definição baseada em cálculos de co-ocorrência (secção 3.3.1). Em seguida, explicamos a definição de Alonso Ramos baseada na Lexicologia Explicativa e Combinatória (LEC) preconizada por Mel’cuk (secção 3.3.2). Na secção 3.3.3, comparamos esta definição abrangente com a definição restrita de Martelli (2006). Descrevemos ainda outros aspetos que são tidos em consideração no processo de definição de colocação, tais como o uso (secção 3.3.4), a utilização de padrões sintáticos (secção 3.3.5) e a existência de um *continuum* onde os vários tipos de combinatórias de palavras se posicionam (secção

3.3.5).

3.3.1. Definição contextual vs definição estatística

Vários autores, como Wanner *et al.* (2011), Alonso Ramos *et al.* (2010), Vincze *et al.* (2011) e Prieto *et al.* (2009), seguem a definição de colocação associada à tradição lexicográfica de Hausmann (1989) e Cowie (1992). Segundo eles, esta é a definição mais apropriada para o trabalho com *corpora* de aprendizagem. Assim, assumem que uma colocação é uma co-ocorrência lexical binária, restrita e idiossincrática de duas unidades lexicais entre as quais se estabelece uma relação sintática direta. A co-ocorrência de um dos itens é escolhida livremente pelo falante, de acordo com o seu significado (base), enquanto que a ocorrência do outro é escolhida dependendo da base e do significado total da co-ocorrência (colocativo), e não simplesmente do significado da unidade lexical isolada. No exemplo, em espanhol, de Prieto *et al.* (2009), o substantivo *dolor* seleciona o verbo *vencer* para expressar o significado “fazer com que algo desapareça”, enquanto que o substantivo *aborrimiento* seleciona o verbo *matar* para veicular o mesmo significado. Nestes casos, *dolor* e *aborrimiento* são bases e *vencer* e *matar* são os colocativos. Esta definição distingue-se das noções que se baseiam em frequências, ou seja, itens lexicais que co-ocorrem com frequência formam uma colocação.

Nesselhauf (2005) posiciona-se também próximo da definição de colocação de Hausmann (1989), Cowie (1992), no sentido em que usa uma definição contextual e não baseada em números de ocorrências. Distingue as colocações das combinatórias livres e das expressões idiomáticas. Para avaliar as expressões que são ou não colocações, usa um dicionário como recurso, o *British National Corpus* e, em casos ambíguos, julgamentos de falantes nativos. A comutabilidade é o critério mais importante para distinguir colocações, ou seja, numa colocação composta por V + N, o nome pode ser usado sem restrições arbitrárias, mas o verbo é, num certo sentido, arbitrariamente restrito a certos nomes.

Por oposição a estas abordagens, e seguindo uma perspetiva historicamente associada a Sinclair, estudos como os de Rodríguez-Fernández *et al.* (2015) optam

por usar métricas automáticas de cálculos de ocorrência para definir colocação⁴⁴. São variadíssimas as métricas utilizadas neste tipo de estudos, sendo as mais conhecidas a medida-t, a informação mútua (MI), a MI3, rácio de verosimilhança, sensibilidade mínima, *logDice* e a *Mi.log-f*. Para mais informação sobre medidas de associação, cf. Kilgarrriff *et al.* (2004).

A utilização de métricas facilita bastante o processo de identificação das colocações. Nestes casos, utiliza-se um valor numérico como limiar de frequência e as potenciais colocações extraídas que obtenham um valor superior são consideradas corretas.

3.3.2. Definição baseada na LEC

Alonso Ramos, no seu trabalho de 2013, dá vários exemplos, em espanhol, francês e italiano, das expressões que entende como colocação: a) *dar PASEO/ faire une PROMENADE/ fare una PASSEGIATA*; b) *FUMADOR empedernido/ gros FUMEUR/ FUMATORE incalito*; c) *diente de AJO/ gousse d'ALL/ spiccio d'AGLIO*. Não existindo uma definição genericamente aceite de colocação, a autora opta por um marco teórico que será uma definição operativa baseada fundamentalmente na co-ocorrência lexical restrita, tal como é definida por Mel'cuk e na Teoria Sentido-Texto. Na Lexicologia Explicativa e Combinatória (LEC) (Mel'cuk *et al.*, 1995), a colocação é um tipo de unidade fraseológica constituída por duas unidades lexicais, a base e o colocativo. Nestes casos, existe uma unidade lexical que restringe a seleção da outra unidade lexical. No caso das colocações, viola-se a propriedade de seleção não restrita, característica dos sintagmas livres: a base da colocação, núcleo semântico da colocação, é selecionada livremente tendo por base o seu significado, enquanto que o colocativo é selecionado em função da base. O significado do colocativo pode ser o que assume fora da colocação, mas pode também assumir um significado que apenas é ativado dentro desta combinatória restrita (ex.: *dente de*

⁴⁴ Neste estudo em particular, utilizou-se uma adaptação do Pointwise Mutual Information (PMI): o NPMIc. Para o PMI, dadas duas variáveis independentes, a probabilidade da sua intersecção é o produto da sua probabilidade. O NPMIc, por sua vez, tem em consideração a assimetria dos elementos que constituem a colocação, propondo assim uma medida de associação assimétrica (Carlini *et al.*, 2014).

alho – dente, somente neste contexto, significa “porção de”). Uma vez que a unidade semântica da colocação é a base, as colocações devem ser descritas lexicograficamente na entrada da base (ex.: *dente de alho* deve ser descrito na entrada do dicionário de *alho*). No momento de codificação, o aluno vai partir de *alho* para depois procurar a palavra que indica “porção de qualquer coisa”. Por esta razão, esta informação deve estar presente na entrada da base. No exemplo da combinatória *saia calça*, se entendermos que se trata de uma saia com pernas, esta deve ser descrita como colocação na entrada do dicionário de *saia*. Pelo contrário, se acharmos que é uma peça de roupa que parece uma saia, mas tem pernas, deve ser tratada como uma locução e deve ter uma entrada lexicográfica própria.

Seguindo esta perspectiva, as colocações não são necessariamente idiossincráticas, nem combinatórias frequentes. A decisão sobre se uma unidade fraseológica é uma colocação ou uma locução deve basear-se na análise do seu significado.

3.3.3. Definição abrangente vs definição restrita

Também Vincze *et al.* (2011) seguem a definição usada por Alonso Ramos e entendem que as colocações são uma combinação binária restrita compostas por duas unidades lexicais, na qual um dos elementos, a base, condiciona a escolha do outro, o colocativo. Consideram colocações combinações em que, por exemplo, o adjetivo tem um uso restrito (*comida rica* (esp.)). No entanto, combinatórias menos idiomáticas, com adjetivos menos restritos, segundo estes autores, devem também ser consideradas colocações (*buena comida, comida fantástica* (esp.)). Esta interpretação é um pouco menos restrita do que a definição de Martelli (2006). Este autor define colocação como a repetida co-ocorrência de itens lexicais; combinações que não podem ser explicadas através de regras sintático-semânticas. Combinação de itens que pertencem a uma classe aberta (exclui combinações gramaticais) que se “atraem” uma à outra e estabelecem uma relação de co-ocorrência no eixo sintagmático. Uma combinação de palavras é considerada uma colocação, se os itens que a constituem forem limitados na sua comutação, sujeitos a variação sintática (não necessariamente adjacentes um ao outro) e se a expressão for semanticamente

transparente (o que exclui todas as combinações idiomáticas).

3.3.4. Definição baseada no uso

Granger (1998), na sua definição de colocação, foca-se essencialmente na questão do uso. Para a autora, as colocações são um fenómeno linguístico onde um item do vocabulário prefere a companhia de um outro item, em vez de um sinónimo deste, por causa de constrangimentos que não ocorrem ao nível da sintaxe ou do seu significado conceptual, mas apenas se explicam pelo uso (J. Van Roey, 1990: 46 *apud* Granger, 1998). Tratam-se de combinações determinadas pela utilização dos falantes de uma comunidade, que demonstram uma relação sintagmática preferencial entre dois elementos num padrão sintático específico. Ambos constituintes contribuem isoladamente para a combinação, mas não partilham o mesmo estatuto. A base é semanticamente autónoma e seleccionada, em primeiro lugar, por causa do seu significado independente pelo utilizador da língua. O segundo elemento, o colocativo, é seleccionado pela base e é semanticamente dependente desta. Alguns exemplos, segundo a autora, que partilham estas características são: *rain, closely linked, apologize profusely* (Granger e Paquot, 2008: 22).

3.3.5. Definição baseada em padrões

Apesar das colocações serem estruturas bastante arbitrárias na associação dos seus constituintes, do ponto de vista formal, estas seguem padrões sintáticos relativamente restritos que foram já elencados por vários autores, como Hausmann (1989), Grossmann e Tutin, (2002). Tendo em mente esta regularidade que as colocações apresentam, Gabrys-Biskup (1992) foca-se nos padrões sintáticos que as colocações podem assumir, nomeadamente V + N, V + Adv, N de N, como auxílio à identificação das colocações. Também do ponto de vista da autora, as colocações são definidas como tendo uma fixidez relativa, tratam-se de estruturas não idiomáticas e o seu significado é sempre transparente.

3.3.6. Definição baseada na noção de *continuum*

Por último, a definição de colocação utilizada por Howarth (1998), tal como em Aisenstadt (1981) e Cowie (1981), segue o modelo soviético. Nesta perspetiva existe um *continuum* entre três categorias de combinatórias: combinações livres, colocações restritas, expressões idiomáticas (figurativas e puras). Num dos extremos, encontram-se as expressões idiomáticas, como os provérbios, e, do outro lado do *continuum*, estão as combinatórias livres, como *lua bonita*. Nas primeiras, o significado total não se reflete através do significado dos componentes, ou seja, o seu significado não é adquirido composicionalmente. Do outro lado, as combinatórias livres possuem um sentido analisável e previsível e não apresentam coesão. Entre estes dois extremos, encontram-se o grupo intermédio das colocações. Segundo Howarth (1998), as colocações distinguem-se por oposições aos dois outros extremos do *continuum*.

Podemos concluir dizendo que, apesar de uma grande parte das investigações apresentar bases muito semelhantes, é possível encontrar pontos que as distinguem. Nomeadamente, Martelli (2006) exclui do escopo das colocações qualquer estrutura mais idiomática, já Vincze *et al.* (2011) não parece ser tão rígido nesse aspeto em particular (secção 3.3.3). Granger (1998) opta por usar critérios de utilização de uma expressão para definir o que é uma colocação (secção 3.3.4), Gabrys-Biskup (1992) usa padrões sintáticos para este mesmo propósito (3.3.5) e Howarth (1998) define colocação por oposição a outras estruturas contíguas (3.3.6). Por último, assumindo uma perspetiva com a qual nos revemos, Alonso Ramos (2013) sublinha a falta de uma definição comumente aceite pela comunidade e usa uma definição operativa (secção 3.3.2).

No nosso entender, a definição contextual é que mais se adequa ao estudo das colocações em *corpora* de aprendizagem. Uma vez que este tipo de *corpus* normalmente apresenta limitações em termos de tamanho, os métodos estatísticos podem revelar-se pouco vantajosos. Porém, num segundo momento, podem ser úteis para realizar uma validação das colocações extraídas em *corpora* de referência. Seguir uma perspetiva operativa, como defende Alonso Ramos, traz também

vantagens práticas, porém acreditamos que a ideia de *continuum*, com a qual Alonso Ramos não concorda, reflete a natureza das colocações, tal como nós as entendemos.

Depois das questões de definição, na seguinte secção, vamos debruçar-nos sobre o processo de anotação dos *corpora* de aprendizagem.

3.4. Identificação de colocações em *corpora* de L2

Na seguinte secção, primeiramente descrevemos os processos de anotação, ou seja, de identificação de colocações em *corpora* de aprendizagem que os estudos previamente elencados utilizaram. Em seguida, na secção 3.4.2, apresentamos as taxonomias para a classificação dos erros encontrados nas colocações. Por último, na secção 3.4.3, apresentamos algumas dificuldades identificadas no processo de anotação das colocações.

3.4.1. Processo de anotação das colocações

Em regra geral, a anotação deste tipo de *corpora* é feita manualmente. Porém, existem alguns trabalhos que optam por utilizar abordagens automáticas que podem facilitar o processo de extração de colocações. Em Philip (2007), apesar dos textos que compõem o *corpus* terem sido corrigidos manualmente, foram usadas concordâncias para procurar exemplos, verificar hipóteses sobre os dados recolhidos manualmente e para medir a produção dos alunos. No caso de Nesselhauf (2005), as colocações também foram manualmente extraídas, mas utilizaram-se padrões sintáticos para acelerar este processo. Apesar de não se tratar de um processo automático, Thewissen (2008) utiliza um *corpus* que foi previamente anotado, analisando apenas 2 dos 56 tipos de erros marcados. O processo de anotação das colocações tornou-se bastante mais célere, uma vez que uma pré-seleção tinha sido efectuada. Assim apenas se trataram os erros em sintagmas lexicais e erros em preposições.

Em relação aos trabalhos que optam por uma abordagem inteiramente manual de tratamento do *corpus*, podemos distinguir os que decidem entregar esta

tarefa a linguistas e os que optam pelo juízo de falantes nativos. Em Rodríguez-Fernández *et al.* (2015), o processo de anotação das colocações no *corpus* foi realizado manualmente por vários linguistas. No entanto, não são explicitadas as diretivas seguidas nesse processo. Porém os autores informam que apenas as colocações que demonstraram acordo entre os anotadores foram consideradas no estudo. Também Thewissen (2008) para a anotação das colocações presentes no *subcorpus* do *International Corpus of Learner English*⁴⁵ utiliza um linguista nativo da língua alvo, neste caso, o inglês. Cada erro anotado é acompanhado por uma etiqueta que descreve a natureza do erro e a sua correção.

Por outro lado, para realizar a anotação das colocações corretas e incorretas numa porção do CEDEL⁴⁶, Lozano, 2009 e Vincze *et al.* (2011) optam por confiar no julgamento de nativos. Neste trabalho, o processo de anotação foi levado a cabo por falantes nativos de espanhol, seguindo um procedimento bem definido e usando a tipologia detalhada de classificação de erros em colocações de Alonso Ramos *et al.* (2010). O cálculo da concordância entre anotadores resultou em valores muito baixos, nomeadamente na interpretação pouco clara de colocação. A avaliação das colocações corretas e incorretas foi realizada por dois anotadores principais, cujas avaliações eram revistas por um terceiro anotador. Anotações sem consenso eram validadas através da procura no *Corpus de Referência do Español Actual* (CREA) e se ocorressem cinco vezes, eram consideradas corretas. Casos que não foram resolvidos por este método, foram enviados para um anotador independente e tratados de acordo com o voto da maioria. Eram ainda realizadas reuniões semanais de discussão da anotação. A concordância entre anotadores aumentou de uma média de 30%, na primeira semana, para uma média de 50% nas últimas semanas.

Apesar da utilização de métodos automáticos ou manuais de extração e da anotação ser feita ou não por linguistas serem questões de extrema importância para o processo de anotação, é de igual importância a taxonomia de classificação de erros utilizada. Na seguinte secção, vamos apresentar as classificações de erros usadas nos estudos de produções de colocações em *corpora* de aprendizagem.

⁴⁵ Um *corpus* composto por 16 línguas maternas, que se encontra etiquetado com 56 tipos de erros, dividido por línguas maternas e níveis de proficiência.

⁴⁶ Tratam-se de trabalhos escritos por aprendizes de espanhol com língua materna inglesa (100 textos, 46.420 palavras).

3.4.2. Taxonomia de análise erros

Depois da descrição do processo de anotação de erros de colocações em *corpora* de aprendizagem, nesta secção, vamos descrever as taxonomias de erros utilizadas nos processos de anotação. Três destas taxonomias serão descritas tendo em consideração os níveis de análise. A última, exposta na secção 3.4.2.4, é baseada numa descrição de aceitabilidade.

3.4.2.1. Análise a três níveis

Lüdeling *et al.* (2005), na sua análise de um *corpus* de produções de aprendentes de alemão, advogam o processo de anotação de erros a três níveis: identificação, descrição e explicação. Esta ideia baseia-se nas ideias de Corder (1974) e Ellis (1994), de acordo com os quais a análise de erros tem quatro passos. Primeiro, tem lugar a recolha dos dados de aprendentes. Segundo, a análise dos erros, num sentido mais restrito, começa com a identificação dos erros. Depois disto, os desvios identificados em relação à hipótese alvo são descritos. Finalmente, no quarto passo, é dada uma explicação para o desvio.

Como resposta a trabalhos como o de Granger (2007), cuja taxonomia agrupa as colocações numa única subclasse de erros lexicais, Ramos *et al.* (2010), seguindo a perspetiva de Corder (1974) e Ellis (1994), propõe também três níveis de anotação: localização, descrição e explicação. Em termos de **localização** do erro, este pode encontrar-se em: a) toda a colocação – colocações incorretas cujo significado deveria ser expresso por uma unidade lexical e unidades lexicais inexistentes em vez de colocações; b) num dos elementos, seguindo a terminologia de Hausmann (1998) – erros de seleção lexical paradigmática, ou seja na base, e erros de seleção lexical sintagmática, isto é no colocativo. Em Alonso Ramos (2013, 8), o nível **descritivo** assume três perspectivas: erros lexicais, gramaticais e de registo⁴⁷ presentes em colocações pragmaticamente não apropriadas. Na verdade, estes são os três aspetos que o aluno tem que dominar para produzir as colocações corretamente: seleção da unidade lexical correta (base ou colocativo), correção gramatical e adequação

⁴⁷ Cada um destes erros ainda se desdobra em mais sub-tipos que não vamos neste contexto descrever.

textual. A última camada de anotação diz respeito à **explicação** da causa do erro. Neste nível em particular, mais do que uma causa pode ser atribuída ao mesmo erro. A primeira distinção é feita entre os erros interlíngua, ou seja, transferências que refletem a LM ou L2, e erros de intralíngua que resultam de um conhecimento incompleto da L2, mas que não demonstram interferência da LM. Esta distinção é feita para os três tipos de erros (lexical, gramatical e registo). Para os erros gramaticais e de registo não se fez mais nenhuma distinção explicativa adicional (apenas interlíngua e intralíngua). Os erros lexicais de interlíngua podem dever-se a: a) importação (forma inexistente em L2); b) extensão (a palavra substituída existe na L2 e, em muitos casos, esta é uma tradução da palavra na LM mas que naquele contexto particular não funciona). Dentro da extensão existe ainda a: b.1) semelhança fonética, por exemplo, o aluno usa *lengua maternal*⁴⁸, em vez de *lengua materna*, devendo-se este erro à semelhança entre as duas palavras. b.2) evitação da L2 por ser muito semelhante à LM. O aluno opta por seleccionar *acudir el teléfono* em vez de *atender el teléfono*. Em relação aos erros de intralíngua, distinguiram-se três subclasses: a) derivação errónea (forma inexistente na L2 como processo de derivação errada), b) generalização (palavra mais vaga ou genérica) e c) eleição errónea (não há propriamente uma explicação). Os erros gramaticais mais comuns estão relacionados com regência (ex.: usar como transitivo um verbo que exige uma preposição).

A taxonomia de Alonso Ramos *et al.* (2010) tornou-se bastante popular nos trabalhos com *corpora* de aprendizagem. Wanner *et al.* (2011) usa-a, sem qualquer alteração, para a análise de um *corpus* de falantes de inglês a aprenderem espanhol. Também Vincze *et al.* (2011) usa as três dimensões: localização (captura se o erro diz respeito a toda a colocação ou apenas uma das partes), descrição (permite, inicialmente, etiquetar o erro como lexical, gramatical e registo) e, por último, explicação. Rodríguez-Fernández *et al.* (2015), na sua investigação, também usa a tipologia de Ramos *et al.* (2010), mas opta por trabalhar os erros gramaticais e lexicais separadamente. No seu trabalho preliminar, ao nível da descrição, analisa apenas os erros gramaticais e, num segundo artigo do mesmo ano, centra-se nos

⁴⁸ Todos os exemplos ilustrativos de Alonso Ramos (2013) encontram-se em espanhol.

erros lexicais.

O nosso trabalho inicial de análise de erros, publicado em Costa *et al.* (2016), assumindo também uma abordagem a três níveis, centramo-nos unicamente na análise de erros lexicais. Em relação à classificação de Alonso Ramos, acrescentámos um tipo de erro “opção melhor” que tem como objetivo captar usos menos comuns de colocações, mas que não se encontram totalmente errados. Ao nível da explicação do erro, incluímos o tipo “variante” que capta, no caso do Português Europeu, o uso da variante do Português do Brasil.

Mais detalhes sobre a taxonomia adotada neste trabalho, serão dados na secção 5.4.2.

3.4.2.2. Análise a dois níveis

Philip (2007), demonstrando interesse apenas pelos níveis da descrição e explicação do erro colocacional, desenvolveu a seguinte taxonomia:

- escolha de um forma decalcada, ex.: *absolutely original* (uso de uma expressão que parece menos marcada do que da L2);
- criação de uma nova colocação com base num sinónimo, ex.: *greatly, vastly original* (o aprendente ativou o tipo semântico da palavra no seu sentido completo e portanto sobre-lexicalizou o colocativo);
- a evitação ocorre quando o aprendente não usa colocações e prefere utilizar, por exemplo, numa construção de V + Adv um advérbio *teddy-bear* como *very* ou *really* ou eliminando mesmo o uso de advérbio;
- Sobre-uso – por exemplo, uso de *extremely* ou *entirely* com muita frequência.

Apesar da organização ser diferente, todos estes tipos de erros encontram um equivalente na tipologia de erros criada por Alonso Ramos.

3.4.2.3. Análise a um nível

Os seguintes trabalhos optam por apenas usar uma das três dimensões

preconizadas por Corder (1974) e Ellis (1994). Martelli (2006) utiliza a primeira fase da anotação, ou seja, a identificação da localização do erro. Os erros em colocações são tratados de acordo com a classe gramatical dos elementos da colocação. Podemos encontrar erros no adjetivo (+ verbo), no verbo (+ nome sujeito ou objeto) no advérbio (+ verbo) ou toda a colocação pode estar errada. Também a taxonomia usada por Nesselhauf (2005) se baseia na localização. Os tipos de erros denominados com a classe morfológica da palavra afectada, podemos encontrar erros no verbo, nome, determinante ou preposição. São também contemplados os casos em que toda a colocação não é apropriada. Existem ainda subdivisões tipológicas de cada um dos tipos de erros iniciais. Em relação ao verbo existem mais 23 tipos de erros (ex.: verbo simples em vez verbo simples; verbo composto em vez de verbo simples), 12 tipos de erros no nome e 17 no determinante.

Howarth (1998) apenas descreve o erro, distinguindo dois tipos: modificação gramatical e substituição lexical. Os seguintes autores centram-se na explicação do erro. Shei *et al.* (2000), tal como Futagi (2008), usa uma taxonomia pouco fina, apenas baseado em erros lexicais que resultam de transferência da LM. Também Thewissen (2008), no seu trabalho com produções de alunos estrangeiros com várias línguas maternas, opta por anotar apenas os erros que demonstram uma possível transferência da LM. A par desta informação, as colocações são também marcadas como aceitáveis ou não e gramaticais ou não.

3.4.2.4. Análise de aceitabilidade

À semelhança do trabalho de Thewissen (2008), também em Prieto *et al.* (2009) as colocações foram marcadas como corretas, incorretas e melhoráveis. Apesar do trabalho de Gabrys-Biskup (1992) não utilizar *corpora* de aprendizagem, a autora também optou por usar um critério de aceitabilidade. Na sua avaliação da compreensão e produção de colocações, os alunos tinham que encontrar equivalentes para as colocações da LM. As respostas foram avaliadas por 3 nativos numa escala de 1 a 4, de inaceitável a completamente equivalente. Para James (1998), as unidades multipalavras, nomeadamente as colocações, podem ter dois tipos de erro: formalmente erradas (não existem) (gramaticalidade) e formalmente

corretas mas inapropriadas ao contexto (aceitabilidade). Para estabelecer este tipo de julgamento, o autor usa dicionários.

Como podemos concluir, as taxonomias de análise de erros colocacionais podem resumir-se a três abordagens. As baseadas nos trabalhos seminais de Corder (1974) e Ellis (1994) que defendem uma análise de erros a três níveis: localização, descrição e explicação (secção 3.4.2.1). Alonso Ramos (2010), Vincze *et al.* (2011), Wanner *et al.* (2011) e Rodríguez-Fernández *et al.* (2015) são alguns dos autores que mantêm esta tradição de análise. Porém, a própria natureza dos dados ou um objetivo específico da investigação podem levar a um desmembramento desta abordagem. Assim, surgem trabalhos que optam por seguir apenas um ou dois dos níveis de análise, como em Martelli (2006) ou Thewissen (2008) (secção 3.4.2.2 e 3.4.2.3). A par destas análises mais detalhadas dos erros, surgem trabalhos que se centram num julgamento de aceitabilidade e de gramaticalidade das colocações (secção 3.4.2.4). Apesar de parecer uma abordagem mais simplista, este tipo de avaliação é bastante útil para trabalhos como os de Prieto *et al.* (2009) ou Gabrys-Biskup (1992).

Nesta última secção, descrevemos alguns problemas específicos que podem surgir na anotação dos dados.

3.4.3. Dificuldades na anotação de colocações

Nesta última secção, vamos apontar alguns desafios encontrados nos processos de anotação. Nomeadamente a questão da delimitação do conceito de colocação, os critérios de aceitabilidade utilizados e a interpretação dos próprios erros.

3.4.3.1. Identificação das colocações

Um dos principais problemas referido tanto por Vincze *et al.* (2011) como por Wanner *et al.* (2011), prende-se com o reconhecimento de uma colocação, ou seja, a dificuldade em estabelecer critérios claros e operacionais para delimitar a noção de

colocação. Isto resulta na dificuldade em distinguir colocações de combinatórias livres e, também, de expressões idiomáticas. Wanner *et al.* (2011), no seu trabalho em espanhol, considera colocações as combinatórias restritas (*comida rica*), no entanto, combinações menos idiomáticas com uma escolha do adjetivo menos restrita também foram consideradas colocações (*buena comida*). Um exemplo da dificuldade em distinguir colocação de um idiomatismo é a expressão *darse cuenta*, que deve ser tratada como uma expressão não composicional, dada a sua estrutura sintática cristalizada. Foi confundida como colocação pelos anotadores, uma vez que o verbo *dar* aparece em construções de verbo suporte (*dar consejos, dar un paseo*). Acontece também colocações passarem despercebidas até surgir a sua versão incorreta. Por exemplo, *país de origem* não foi anotado como colocação até surgir a forma *países maternos*. Expressões com erros são mais prováveis de ser interpretadas como colocações, como *lleno *con historia*. O erro na preposição pode ter chamado a atenção para a expressão. *Depender de* não foi considerada uma colocação, apesar de, por vezes, na literatura serem consideradas colocações gramaticais (Benson *et al.*, 1986). Porém preposições governadas por um dos membros da colocação foram consideradas como parte de uma expressão (ex.: *tener miedo de*) e, quando têm algum erro, este é considerado um erro colocacional gramatical.

3.4.3.2. Critérios de correção

Segundo Alonso Ramos (2013), os critérios utilizados para determinar se uma colocação é correta ou não, nem sempre estão bem definidos. Vincze *et al.* (2011) observou que os anotadores apresentam níveis diferentes de permissividade, não apresentando um grande acordo. Neste estudo, observou-se que os anotadores tendem a ser menos permissivos com não nativos do que seriam com nativos, no que toca a usos criativos ou pouco comuns. Os juízos de aceitabilidade são muito subtis; combinações que um nativo pode dizer que se trata de um uso criativo, a um aprendente esse uso não lhe é consentido. Por exemplo, a colocação *admirador empedernido* pode não se encontrar em *corpora* de referência do espanhol, mas isto não quer dizer que não se possa usar. Aliás começam a ser registados alguns usos; a

expressão perde a sua conotação negativa e *empedernido* assume apenas um valor intensificador com o sentido de “muito intenso.” Este comportamento pode ser explicado como uma estratégia de sobre-correção, que pode levar à desmotivação dos alunos. A procura num dicionário pode nem sempre ser a melhor abordagem para o aluno ultrapassar uma dúvida. Uma pesquisa no Google pode mostrar ao aluno ocorrências mais recentes, uma vez que a língua se encontra em constante evolução e nem sempre estas mudanças se encontram documentadas num dicionário. Por último, surgem também problemas em reconhecer variação, nomeadamente usos no espanhol da América Latina que foram inicialmente marcados como incorretos, mas que num segundo momento, através da procura em *corpora*, foram também detetados usos em outros países de língua espanhola. Estes casos foram anotados como corretos, especificando a variante a que pertenciam (*tomar clases – ir a clase*).

3.4.3.3. Interpretação do erro

Em relação ao nível da interpretação do erro, foram identificados por Vincze *et al.* (2011) três problemas na anotação. Primeiro, a etiqueta atribuída a um erro está dependente da sua correção, casos com mais do que uma correção podem ser problemáticos (*hizo gorditas – ponerse gordas* (colocativo errado) ou engordar (forma analítica errada)). Segundo, combinações erradas que parecem colocações são normalmente traduções literais da LM que não tinham equivalente colocacional em espanhol e que apenas podem ser traduzidas através de uma expressão não fraseológica (*humo de segunda mano – secondhand smoke*). Existem ainda os casos contrários, nos quais os alunos não usam uma colocação, mas a forma correta seria uma colocação em espanhol (*estoy curioso – tengo curiosidad*). Terceiro, em casos em que a fonte do erro era ambígua foram admitidas duas etiquetas (*hice citas – made appointments* – tradução direta ou generalização usando um verbo genérico).

Capítulo 4. *Corpora* e sistemas de extração

Neste capítulo, vamos descrever os *corpora* aos quais recorreremos para construir o nosso corpus de análise de produção das colocações. Na secção 4.1., fazemos uma avaliação dos sistemas automáticos disponíveis para a extração de colocações em Português Europeu. Fazemos uma comparação entre estes sistemas, de modo a que possamos avaliar as vantagens e limitações dos sistemas existentes

4.1. Recolha de *corpora*

Para a extração das colocações, utilizámos dois *corpora* recolhidos no seio de projetos de investigações e outros dois recolhidos por nós. Passamos, de seguida, à descrição destes *corpora*.

4.1.1. *Corpus* de Produções Escritas de Aprendentes de Português Língua Segunda

O *Corpus de Produções Escritas de Aprendentes de PL2*⁴⁹ (PEAPLE2) (Pereira, 2014) foi criado pelo Centro de Estudos de Linguística Geral e Aplicada (CELGA) da Faculdade de Letras da Universidade de Coimbra (FLUC). Este *corpus* é o resultado do projeto *Recolha de Corpora de PL2* que teve início em 2008 e que teve duas fases de recolha: a primeira fase entre Maio de 2009 e Maio de 2010 e a segunda fase

⁴⁹ <http://www.uc.pt/fluc/rcpl2/> (Download do corpus feita a 4 de Dezembro de 2013).

entre Janeiro e Maio de 2011. É composto por produções escritas de 458 aprendentes de PL2 que frequentavam os vários níveis (A1 a C1⁵⁰) dos Cursos de Português para Estrangeiros a funcionar na Faculdade de Letras da Universidade de Coimbra. Os dados foram recolhidos com consentimento informado dos sujeitos participantes e são de dois tipos: textos escritos (cada um dos quais correspondendo a um documento no formato txt ou doc) e uma ficha de informações sobre os perfis dos informantes (apresentados num ficheiro excel). Cada produção escrita foi obtida a partir de um estímulo. Foram usados nove estímulos, extraídos da lista dos inicialmente propostos no projeto *Recolha de Dados de Aprendizagem de Português Língua Estrangeira* (coord. Prof. Doutora Isabel Leiria), tendo cada um deles um código identificativo. Os nove estímulos repartem-se em três grandes áreas temáticas: o indivíduo, a sociedade e o meio ambiente. Alguns exemplos de tópicos/estímulos são *Fale um pouco de como ocupava o tempo quando era mais pequeno, dos seus amigos e das vossas brincadeiras* (o indivíduo), *Todos os países são diferentes a nível cultural e geográfico. Descreva o seu país, observando as particularidades das suas regiões, os principais monumentos e saliente alguns dos hábitos mais frequentes da sua cultura* (a sociedade) e *Nos últimos tempos ouvimos falar muito de agricultura biológica. Diga qual é a sua opinião sobre o assunto* (o meio ambiente). Em alguns casos, o mesmo informante produziu mais do que um texto, partindo de dois ou mais estímulos diferentes.

Estes textos foram recolhidos em manuscritos, no âmbito de exercícios de avaliação (contínua e/ou final), foram depois transcritos de acordo com convenções descritas em detalhe em Leiria, I. 2006 – *Léxico, aquisição e ensino do Português Europeu língua não materna*. Lisboa: FCG/FCT (p. 201). Estes textos foram ainda anonimizados, ocultando os nomes próprios e outros elementos que permitissem a reconstituição da identidade do informante. A cada texto foi atribuído um código de identificação, do qual constam os seguintes elementos: língua materna do aprendente, tipo de curso de PL2 frequentado pelo informante, nível da turma, número de identificação do aprendente e código do estímulo usado para a produção do texto. Ex.: Um texto sobre atividades dos tempos livres (33.1J) produzido por um

⁵⁰ Níveis definidos pelo Quadro Europeu Comum de Referência para as Línguas (QECR).

informante (01) de LM alemã (ALEMÃO) que frequentava, aquando da recolha de dados, uma turma do nível A2 destinada a alunos do programa Erasmus (ER), é identificável através do código: ALEMÃO.A2.01.33.1J.

Os dados foram recolhidos junto de aprendentes que frequentavam três tipos de curso, apresentando perfis diferentes o público de cada um deles. As *unidades curriculares de Língua Portuguesa Erasmus* são exclusivamente frequentada por alunos universitários dos três ciclos, que apenas permanecem em Portugal durante um ou dois semestres letivos. O *Curso Anual de Língua e Cultura Portuguesa para Estrangeiros* atrai um público mais diversificado. Muitos estão em Portugal por razões profissionais ou pessoais, encontram-se em situação de imersão, e não têm, necessariamente, formação universitária (a idade mínima de matrícula é 16 anos). A maioria dos alunos do *Curso de Férias de Língua e Cultura Portuguesa* não reside em Portugal. Tende a ser um grupo heterogéneo do ponto de vista das idades e da formação académica, apresentando, igualmente, motivações muito diversas para a aprendizagem, em regime intensivo, do Português.

Aos sujeitos que participaram na recolha foi pedido o preenchimento de uma ficha individual. Foram assim recolhidas as seguintes informações:

- Dados pessoais: idade, sexo, naturalidade, países em que tenha vivido/ quanto tempo;
- Indicações sobre a história linguística individual: língua materna, língua de escolarização, outras línguas não maternas (LNM) conhecidas, hábitos e domínios de uso do português, ano de início do estudo do português;
- Autoavaliação do nível já atingido em português e na outra LNM que melhor domine no âmbito da: compreensão do escrito, compreensão do oral, interação oral, produção oral, produção escrita.

Para a nossa análise da produção de colocações não usámos a totalidade deste *corpus*, mas apenas as produções dos alunos que tinham como língua materna o Espanhol, o Italiano, o Francês, o Inglês e o Alemão. Excluímos também todos os

alunos que disseram ter mais do que uma LM, mesmo que uma delas fosse uma das *supra* citadas. Dado que queríamos analisar os dados usando apenas três grupos de proficiência, agrupámos os alunos do nível A1 e A2 no nível 1, os alunos do B1 e B2 no nível 2 e os alunos do nível C1 no nível 3. A seguinte tabela mostra os números totais dos alunos que preenchiam estes requisitos e que, como tal, foram usados no nosso estudo bem como o número de textos que eles produziram.

LM	Nível	Nº alunos	Nº textos
Alemão	1	17	21
	2	47	67
	3	0	0
Inglês	1	3	8
	2	27	45
	3	5	6
Francês	1	4	6
	2	22	29
	3	2	3
Italiano	1	27	40
	2	28	42
	3	0	0
Espanhol	1	19	20
	2	33	41
	3	5	7
TOTAL		239	335

Tabela 2 Número de alunos e número de textos por língua materna e nível usados no estudo.

Na tabela 3, podemos ver os tópicos das composições que usámos no nosso estudo e número de textos que foi escrito para cada um deles. Relembro que estes tópicos pertencem as três grandes temas: o individuo, a sociedade e o meio ambiente.

Estímulo	Nº textos
Escreva um texto em que se apresente.	39
Fale dos seus tempos livres.	63
Descreva o seu país: descreve os monumentos e os hábitos.	30
Descreva um episódio sobre contacto de culturas.	13
Fale dos seus gostos alimentares.	7
Escreva uma carta a um amigo que não vê há muito tempo.	81
Quais as vantagens e desvantagens de viver no campo ou na cidade?	48
Fale sobre os meios de transporte: quais usa, quais prefere.	17
Fale do bairro onde mora.	37

Tabela 3 Número de textos escritos por estímulo.

4.1.2. Recolha de Dados de Aprendizagem de Português Língua Estrangeira

A coordenação da recolha dos dados destes *corpus* esteve igualmente a cargo da Professora Doutora Isabel Leiria e resultam do projeto de *Recolha de dados de aprendizagem de português língua estrangeira*⁵¹, realizado ao abrigo do protocolo entre o Instituto Camões e o Centro de Linguística da Universidade de Lisboa. Para questões de abreviatura, chamaremos a este *corpus* CLUL. Este projeto teve como objetivo recolher produções de aprendentes de português língua estrangeira com vista à criação de uma base de dados que pudesse apoiar a investigação na área da língua portuguesa, ajudar na formação de professores e servir de base para a criação de material didático de português língua estrangeira. Tratando-se de material real, este é um instrumento fundamental para a investigação na área da aprendizagem de PL2 e daí o termos utilizado na nossa investigação.

Este projeto teve início em Outubro de 2008 e terminou em Outubro de 2010. O *corpus* é constituído por 470 produções escritas, sendo que cada um dos documentos tem em média 150 palavras, o *corpus* tem cerca de 70.500 palavras transcritas. Estes documentos foram escritos por 397 aprendentes de PL2, falantes de 28 diferentes línguas maternas. Os alunos frequentavam cursos com diferentes

⁵¹ <http://www.clul.ul.pt/en/resources/314-corpora-of-ple> (Download do *corpus* feita a 4 de Dezembro de 2013).

níveis de aprendizagem (A1 – C1), em dezoito universidades dos seguintes países: Alemanha, Áustria, Bulgária, Coreia do Sul, Espanha, Estados Unidos da América, França, Índia, Itália, Polónia, Reino Unido, República Checa e Roménia. A recolha dos dados esteve a cargo dos docentes de português que se encontravam ao serviço do Instituto Camões.

A metodologia definida para a recolha dos dados junto dos participantes incluía: o preenchimento de uma ficha de perfil linguístico à qual foi atribuída um número e a tarefa de redação de um texto que seria identificado com o respetivo número do informante. Este sistema de códigos permitia identificar as produções efectuadas pelo mesmo aluno. Estes materiais foram transcritos e identificados e organizados, estando este trabalho a cargo da Prof. Rita Gonçalves.

Os dados sociolinguísticos dos informantes estão organizados num ficheiro excel. A informação está organizada pela universidade onde foi efectuada a recolha e contém os seguintes itens:

- Dados pessoais: idade, sexo, nacionalidade, curso e respetivo ano de curso e universidade em que o frequenta;
- Percorso linguístico: língua materna, língua de escolarização, outras línguas que estuda/conhece para além do português, avaliação do domínio da língua na qual tem maior proficiência linguística além da LM, segundo os níveis do *Quadro Europeu Comum de Referência para as Línguas* (QECR) nas competências de compreensão oral e escrita, produção oral e escrita e interação oral;
- Percorso linguístico da língua portuguesa: ano de início do estudo do português, outras disciplinas de língua/cultura portuguesa, além do português, que frequenta ou frequentou, contacto com outros falantes do português, avaliação do domínio do português, segundo os níveis do QECR nas competências de compreensão oral e escrita, produção oral e escrita e interação oral;
- Estímulos utilizados.

Tal como no *corpus* descrito anteriormente, as produções foram obtidas a

partir de um estímulo. Foi disponibilizado aos professores uma lista de 83 propostas de tópicos de redação (concebida através da tese de doutoramento da Professora Doutora Isabel Leiria). Tal como no *corpus* anterior, estas propostas de tópico de escrita foram organizadas em três grandes áreas temáticas: o indivíduo, a sociedade e o meio ambiente. Quem decidiu que estímulo apresentar aos participantes foram os professores, de acordo com o nível dos alunos. No entanto, com o intuito de ir ao encontro das preferências dos informante, aos próprios alunos foi dada a escolha entre um/dois estímulos de cada um dos grandes temas.

Os textos obtidos foram transcritos de acordo com as convenções descritas em Leiria, I. 2006 – *Léxico, aquisição e ensino do Português Europeu língua não materna*. Lisboa: FCG/FCT (p. 201), tal como no *corpus* anterior. Também em conformidade com a *Recolha de Corpora de PL2* da Universidade de Coimbra, os nomes próprios e outros elementos passíveis de permitir a reconstituição da identidade do informante foram ocultados com o código XXXXX.

Cada documento está identificado com as iniciais relativas à universidade onde foi efectuada a recolha, nível de proficiência em português do aluno no momento da recolha (os códigos 1, 2 e 3 atribuídos referem-se aos níveis A1-A2, B1-B2 e C1-C2 do QECR), número de informante (atribuído na ficha do perfil linguístico) e código de estímulo. Assim, um texto redigido na Universidade de Rutgers (RU), produzido por um aluno de nível A1-A2 (1), com o número de identificação 07, sob o estímulo 45.2L, tem a seguinte identificação: RU_1_07_45.2L.

Tal como fizemos com o *corpus* anterior, seleccionámos como foco do nosso estudo somente os alunos que tinham língua materna alemã, inglesa, francesa, italiana e espanhola. Também como anteriormente, agrupámos os alunos em 3 níveis de proficiência. Assim, a seguinte tabela mostra o número total de alunos e textos que foram considerados no nosso estudo.

LM	Nível	Nº alunos	Nº textos
Alemão	1	40	41
	2	0	0
	3	0	0

Inglês	1	21	21
	2	8	12
	3	2	4
Francês	1	8	8
	2	0	0
	3	0	0
Italiano	1	47	63
	2	32	35
	3	5	11
Espanhol	1	35	36
	2	21	43
	3	0	0
TOTAL		219	274

Tabela 4 Número de alunos e número de textos por língua materna e nível usados no estudo.

Na tabela 5, podemos ver o número de textos que foi escrito para cada um dos estímulos. Tal como no *corpus* anterior, estes tópicos pertencem a três grandes temas: o indivíduo, a sociedade e o meio ambiente.

Estímulo	Nº textos
Escreva um texto em que se apresenta.	81
Descreva alguém que admira ou detesta.	5
Se ganhasse uma bolsa de estudo, onde gostaria de ir.	3
Quando fora grande, quero ser...	1
O dinheiro não dá felicidade, mas ajuda...	6
Descreva as melhores férias da sua vida.	4
Imagine que ganhou muito dinheiro e tem mês de férias.	16
Convide um amigo para ir de férias.	
Descreva a capital que mais gostou ou gostaria de visitar para uma revista de viagens.	6
Fale de uma festa tradicional à qual costuma ir.	4
Conte um filme ou um espetáculo que gostou.	9
Responda a um anúncio para participar num filme, contando o filme da sua vida.	2
O que pensa sobre os livros electrónicos.	7
Fale sobre os seus hábitos de leitura.	5
Fale sobre as vantagens e desvantagens do telemóvel.	6

Basta um computador ligado à internet para viajar?	7
Como ocupava o tempo quando era pequeno?	7
Descreva o seu país: descreve os monumentos e os hábitos.	10
Descreva um episódio sobre contacto de culturas.	1
Fale sobre a preocupação crescente com o aspeto físico.	2
Frequenta o ginásio? O que pratica e porquê?	1
Fale dos seus gostos alimentares.	17
Fale sobre a questão da saúde e da alimentação.	1
Gostava de se inscrever num curso de cozinha?	4
Escreva uma carta a um amigo que não vê há muito tempo.	7
Quais as suas preferências em termos de restaurantes.	5
Quais os maiores problemas da emigração.	3
Aborde a questão da discriminação dos emigrantes.	2
Imagine que mudou de casa, cidade ou país. Escreva a um amigo a contar.	4
Trocaria a cidade pelo campo?	5
Qual a sua opinião sobre a agricultura biológica.	3
Escreva um texto para uma campanha publicitária para convencer as pessoas a reciclar.	3
Fale sobre os meios de transporte: quais usa, quais prefere.	5
Fale da sua cidade e do seu bairro.	4
O sonho comanda a vida?	9
Centros comerciais ou comércio tradicional?	10
Qual o papel da publicidade no mundo de hoje?	9

Tabela 5 Número de textos escritos por estímulo.

4.1.3. Produções dos Alunos do Curso Semestral de Língua Portuguesa para Estrangeiros

O terceiro conjunto de dados utilizado no nosso estudo foi recolhido por nós na Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa (FCSH – UNL). A FCSH organiza todos os anos o curso semestral de Língua Portuguesa⁵² que tem como principal alvo os alunos Erasmus da Universidade Nova de Lisboa. Os alunos interessados no curso realizam um teste para avaliar os seus conhecimentos de português que servirá para os colocar no nível apropriado. Tal como nas outras duas recolhas de dados, os níveis utilizados são os do *Quadro Europeu Comum de Referência para as Línguas*, ou seja, A1, A2, B1, B2, C1. As aulas destes alunos decorrem com uma frequência bissemanal, tendo cada aula a duração de 2h.

Os dados que utilizámos na nossa investigação foram recolhidos ao longo do

⁵² <http://www.fcsh.unl.pt/clcp/>

1º Semestre de 2011 junto de uma turma do nível A2. Esta turma, constituída por 32 alunos, era essencialmente composta por estudantes que se encontravam em Lisboa a fazer um semestre ou um ano de Erasmus. A maioria estudava na Faculdade de Ciências e Tecnologia no *Campus* do Monte da Caparica. As nacionalidades dos alunos eram alemã, francesa, espanhola, italiana, sueca e holandesa. No entanto, usámos o mesmo critério de seleção dos dados, definido anteriormente, e excluámos o aluno sueco e o aluno holandês. Para questões de abreviatura, chamaremos a este *corpus* FCSH. A tabela 6 mostra-nos o número de alunos por LM e o número de textos que produziram⁵³.

LM	Nº alunos	Nº textos
Alemão	2	2
Francês	1	10
Espanhol	14	91
Italiano	13	65
TOTAL	30	168

Tabela 6 Número de alunos por língua materna.

Os textos *supra* mencionados foram recolhidos ao longo do período lectivo. Foi pedido aos alunos que fizessem composições cuja nota faria parte da avaliação final de cada um. Estas composições foram realizadas na sala de aula como qualquer outro exercício. Assim foram recolhidos 168 textos. Os tópicos das composições faziam parte do currículo do nível A2. Não sendo exatamente os mesmos usados nos *corpora* anteriores, são tópicos bastante próximos. A seguinte tabela resume estes tópicos e mostra-nos o número de composições de cada tópico.

Estímulo	Nº textos
Descreve um caro amigo.	18
Conta como foi o teu último aniversário.	18
Escreve uma carta a um amigo.	21
Descreve a tua casa.	27

⁵³ Relembro que todos estes alunos frequentam o nível A2, o que equivale ao nível 1, seguindo a divisão previamente apresentada.

Conta como é a tua cidade.	19
Descreve um evento cultural no qual tenhas participado recentemente.	2
Faz a descrição física de um amigo.	4
Conta a tua experiência Erasmus.	11
Como foram as tuas últimas férias?	2
Descreve uma festa típica da tua terra.	1
Quais são os teus hobbies?	16
Como é normalmente o teu Natal?	12
Explica como se faz uma receita culinária ao teu gosto.	17

Tabela 7 Tópicos das composições e número de composições.

Apesar dos alunos não terem respondido a um inquérito formal para a realização do nosso estudo, possuímos algumas informações que constavam na folha de identificação que preencheram no início do ano lectivo. Assim sabemos a idade, a língua materna e outras línguas que também sabiam para além do português e o curso que frequentavam. Sobre estes alunos possuímos também a nota, de 0 a 5, que foi atribuída a cada composição e avaliação final de cada aluno (0 – 20).

4.1.4. Textos dos alunos do Curso Online da Ciberescola da Língua Portuguesa

O último conjunto de dados foi recolhido através da Ciberescola da Língua Portuguesa⁵⁴. A Ciberescola é uma plataforma de recursos interativos e cursos online de ensino do português. Os alunos interessados em frequentar as aulas online fazem um teste inicial que avalia os seus conhecimentos escritos e orais e que servirá para os colocar no nível adequado, seguindo os níveis do QECR. Em seguida, as aulas decorrem em sistema de vídeo conferência e os exercícios são realizados numa página da *internet*. As aulas podem ser individuais ou em pequenos grupos.

Os alunos seleccionados para participar neste estudo frequentavam o curso do nível B1 e o curso do nível C1. Ambos os cursos tinham uma duração total de 36 horas e decorreram de 21 de Maio a 17 de Julho de 2013. As nacionalidades dos alunos eram alemã, espanhola, grega e russa. No entanto, usando o critério previamente estabelecido, excluímos os dados dos alunos gregos e russos. A tabela 8

⁵⁴ <http://www.ciberescola.com>

mostra-nos o número de alunos por língua materna e nível e o número de textos produzidos pelos alunos. Tal como nos *corpora* anteriormente descritos, o nível A1 e A2 correspondem ao nível 1, o B1 e B2 correspondem ao nível 2 e o nível C1 e C2 ao nível 3.

LM	Nível	Nº alunos	Nº textos
Alemão	2	1	6
	3	1	1
Espanhol	2	1	6
TOTAL		3	13

Tabela 8 Número de alunos por língua materna.

Os textos recolhidos foram produzidos como trabalho de casa dos alunos, sendo as temáticas dos mesmos relacionadas com a matéria lecionada. A tabela 9 mostra-nos os tópicos e o número de composições por tópico.

Estímulo	Nº textos
A importância da agricultura.	1
Escreve uma carta a um amigo.	2
Divertir-se sem gastar uma fortuna.	1
As férias perfeitas.	3
Conta o último filme que viste.	2
Escreve a história do desenho animado “A linha”.	2
Explica como se faz uma receita culinária ao teu gosto.	2

Tabela 9 Tópicos das composições e número de composições.

Como os cursos não requeriam presença, a maioria dos alunos não se encontrava em Portugal. Apesar dos aprendentes não terem preenchido um inquérito formal, possuímos várias informações sobre os alunos, como Idade, número de anos de estudo da língua portuguesa, formação académica, outras LNM's que também sabiam e se usavam o português fora do contextual lectivo.

Resumindo, a totalidade do *corpus* que constituímos é composta pelas

produções dos alunos de língua materna alemã, inglesa, francesa, italiana e espanhola e que frequentavam cursos do nível A1, A2, B1, B2, C1 e C2, sendo agrupados, tal como anteriormente, em três níveis: 1 (nível A1 e A2), 2 (B1 e B2) e 3 (C1 e C2). A tabela, em baixo, mostra-nos o número total de alunos que participaram nestes estudo, o número de textos e o número de palavras que estes continham.

LM	Nível	Alunos	Textos	Palavras
Alemão	1	59	64	8229
	2	48	73	20737
	3	1	1	523
Espanhol	1	68	147	18057
	2	55	90	19673
	3	5	7	1351
Francês	1	13	24	3002
	2	22	29	8101
	3	2	3	867
Inglês	1	24	29	4424
	2	35	57	14760
	3	7	10	2077
Italiano	1	87	168	24225
	2	60	77	18595
	3	5	11	2318
TOTAL		491	789	146822

Tabela 10 Número total de alunos, de textos e número de palavras

4.1.5. Grupo de controlo

Por último, recolhemos também os dados produzidos pelo grupo de controlo do *corpus* PESTRA criado por Leiria (2006). Este grupo é composto por 50 falantes nativos de português LM. Trata-se de alunos dos primeiro e do quarto ano do curso de Línguas e Literaturas da Faculdade de Letras da Universidade de Lisboa. Foi pedido a este grupo de controlo que produzisse composições com tópicos muito

semelhantes aos que os alunos estrangeiros produziram nos *corpora* descritos em 4.1.1. e 4.1.2. Estes tópicos dividem-se em três temáticas: o indivíduo, a sociedade e o meio ambiente. Os subtópicos e o número de textos produzidos para cada um dele estão representados na tabela 11. Este grupo de nativos produziu 51 textos, com um total de 11.667 palavras.

Estímulo	Número textos
O indivíduo	
Vida pessoal (do autor ou de outrem)	
Características e comportamentos	4
Experiências e recordações	7
Relações com o próximo	2
Vida profissional	
Estudos	1
Procura de emprego	3
Trabalho (profissões e ofícios)	1
Problemas financeiros	2
Vida cultural e lazer	
Projetos de férias, férias e viagens	5
Passatempos	1
A sociedade	
Características de grupos e povos	
Hábitos sociais e comportamentos	10
Hábitos alimentares e saúde	4
Problemas económicos e sociais	
Alcoolismo	2
O meio ambiente	
O campos / A cidade	1
A cidade: o trânsito	3
O bairro	3
A sociedade de consumo	2

Tabela 11 Estímulos das composições produzidas pelos nativos.

4.2. Sistemas de extração para Português

No contexto desta investigação, pareceu-nos pertinente fazer um pequeno estudo sobre as ferramentas de extração mais utilizadas para o português. Assim,

nesta secção, estabelecemos os parâmetros de avaliação seguidos e analisamos quatro ferramentas disponíveis online para a extração automática de candidatos a colocação em Português Europeu: Wortschatz⁵⁵, DeepDict⁵⁶, *Corpus* de Referência do Português Contemporâneo⁵⁷ e Sketch Engine⁵⁸. Os resultados deste trabalho estão publicados com maior detalhe em Costa (2016).

No entanto, como viemos a concluir, a abordagem automática não se verifica a mais adequada para o nosso trabalho com *corpora* de PL2 e optámos por realizar a identificação das colocações manualmente.

4.2.1. Parâmetros de avaliação

Uma vez que a pesquisa de colocações se realiza, em todos estes programas, através da base da colocação, seleccionámos de um *corpus* de frequências de Português, disponível na Linguateca⁵⁹, dez substantivos: cinco com grande frequência na língua e outros cinco com baixa frequência. Estes nomes foram, de seguida, submetidos aos já mencionados motores de pesquisa. Para a avaliação do desempenho das ferramentas usámos alguns parâmetros que nos parecem importantes para a construção de um *corpus* de colocações. Assim, no nosso entender, o **tempo de processamento** deve ser breve, uma vez que normalmente se realizam múltiplas pesquisas. Existir a possibilidade de definir um **limiar de frequência** é também muito importante porque permite fazer um rastreio inicial entre as colocações mais salientes e simples combinatórias de palavras. A pesquisa deve ser realizada pelo **lema** porque utilizar apenas uma forma de uma palavra pode produzir resultados muito limitados. Determinar a **janela de procura de palavras** é bastante relevante porque, em alguns casos, os dois constituintes da colocação não ocorrem um a seguir ao outro. Nesse caso, é necessária uma procura mais alargada e não apenas num arco de duas palavras. Por exemplo, na frase “O chá é extremamente forte”, apesar da colocação ser descontinua, um sistema de extração

⁵⁵ <http://corpora.informatik.uni-leipzig.de>

⁵⁶ <https://gramtrans.com/deepdict/>

⁵⁷ <http://www.clul.ul.pt/pt/recursos/183-reference-corpus-of-contemporary-portuguese-crpc>

⁵⁸ <https://www.sketchengine.co.uk>

⁵⁹ http://www.linguateca.pt/COMPARA/listas_freq.php

deve ser capaz de a encontrar. Dado que a utilização de **padrões sintáticos** tem demonstrado bons resultados na tarefa de extração de colocações, seria também importante um sistema ter esta valência. Gostaríamos também que a ferramenta distinguisse entre **variedades do Português** e que não utilizasse **textos repetidos nos seus corpora**, de modo a garantir a representatividade dos candidatos extraídos e as suas estatísticas. Seria também interessante ter acesso às colocações em **contexto** e não apenas a uma lista de co-ocorrências e o valor estatístico da sua associação. Gostaríamos também de poder fazer **download** das frases, onde se encontraram os candidatos a colocação, mas também poder fazer **upload** de outros *corpora* e fazer pesquisas sobre eles. Em relação às **métricas**, gostaríamos de ter acesso, pelo menos, às mais utilizadas para a extração de colocações, como a informação mútua⁶⁰, o rácio de verosimilhança⁶¹ e o coeficiente de Dice⁶². Finalmente, seria também uma mais valia a utilização de uma **ferramenta gratuita**.

4.2.2. Ferramentas

Nesta secção, apresentamos as ferramentas analisadas: Wortschatz, DeepDict, *Corpus* de Referência do Português Contemporâneo e Sketch Engine. A palavra *momento* foi utilizada com um propósito ilustrativo em todos os exemplos.

4.2.2.1. Wortschatz

O Wortschatz é um sistema desenvolvido pela Universidade de Leipzig que utiliza o Leipzig Corpora Collection⁶³. Este *corpus* é composto por mais de 250 línguas, todos os seus dados são pesquisáveis e, até um milhão de frases, podem ser descarregadas. Os dados do português consistem num *corpus* constituído por artigos

⁶⁰ Esta medida relaciona o número de co-ocorrências entre duas palavras com o número de ocorrências de cada palavra individualmente, medindo assim a sobreposição destas.

⁶¹ Partindo de duas hipóteses, previamente definidas, esta medida diz qual das duas hipóteses é a mais provável. Esta medida tem em consideração a frequência de cada um dos elementos da colocação, a frequência das duas palavras em conjunto e o número de palavra do *corpus* onde ocorrem.

⁶² Esta medida estima o grau de coesão que existe entre duas palavras. Para mais informação sobre medidas de associação, cf. Kilgariff *et al.* (2004).

⁶³ <http://corpora.uni-leipzig.de/>

de jornais retirados da internet em 2011 e pela Wikipedia⁶⁴, também retirada da internet em 2011. Existem 2.540.587 frases e 53.879.750 tokens para a variedade do Português Europeu. A variedade do Português do Brasil tem 25.008.883 frases e 486.724.987 tokens. A variedade de Macau possui menos frases (391.371) e menos tokens (8.672.381). A pesquisa pode ser feita em uma das variedades ou em todas.

Várias ferramentas são utilizadas no pré-processamento do *corpus*, nomeadamente tokenização, cálculos de frequências de palavras e cálculos de co-ocorrências de palavras. São usadas também técnicas de pós-processamento, como etiquetação morfológica e lematização. Estas ferramentas estão disponíveis gratuitamente e podem ser descarregadas do website do Wortschatz, tal como a maioria dos *corpora*. A pesquisa é feita através da introdução de uma palavra-chave. Esta é a única ferramenta avaliada que não permite a definição de um limiar de frequência, tal como não permite pesquisas pelo lema de uma palavra. Na figura 5, podemos observar o resultado para a pesquisa da palavra *momento*. Em cima, temos informação sobre o número de co-ocorrências, posição e a classe da frequência. Em baixo, podemos ver as frases onde ocorre a palavra pesquisada. Seguidamente, são apresentadas as co-ocorrências, as co-ocorrências à esquerda e à direita. Se clicarmos numa das palavras co-ocorrentes, podemos ver frases onde esta aparece, mas não temos acesso à colocação propriamente dita em contexto frásico. Finalmente, não é possível descarregar os resultados.

⁶⁴ <https://pt.wikipedia.org/>

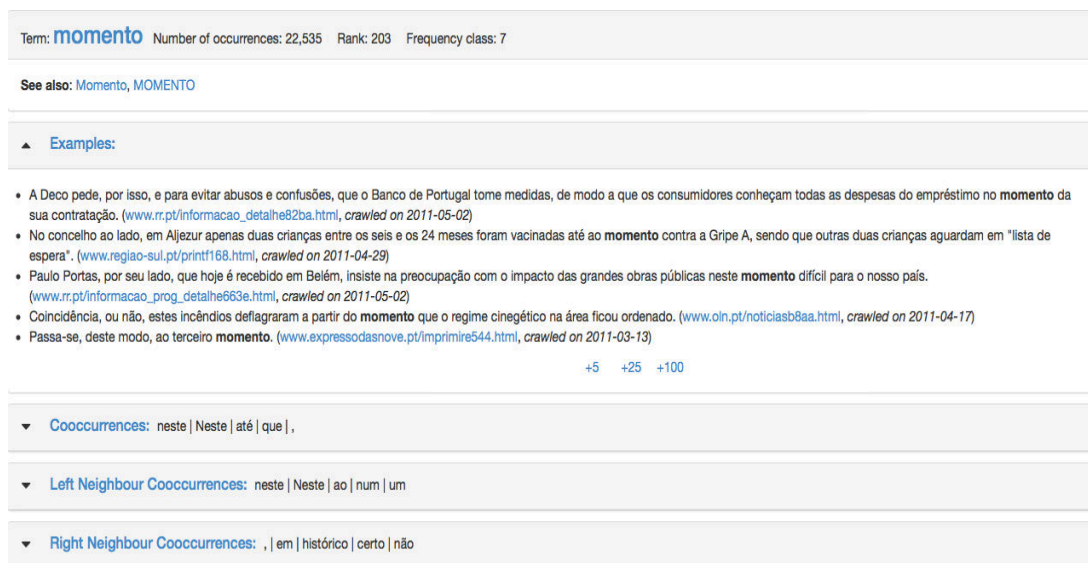


Figura 5 Resultados do Wortschatz para a palavra *momento*.

4.2.2.2. DeepDict

DeepDict é uma ferramenta gratuita que, através da informação que possui, permite construir complexas entradas de dicionário e visualizar contextos de uma determinada palavra em poucos segundos. As relações entre as palavras são baseadas numa Gramática de Restrições⁶⁵, análise de dependências e funções gramaticais, e não apenas em co-ocorrências, como acontece com o Wortschatz. No caso do Português, o analisador sintático utilizado é o PALAVRAS⁶⁶ e o *corpus* é a Floresta Sintá(c)tica⁶⁷. Este *corpus* é constituído por uma coleção de frases que foram morfossintaticamente analisadas, resultando num *treebank*⁶⁸ de 1.000.000 palavras recolhidas do CETEMPúblico (Santos, 2001).

Para procurar candidatos a colocação, escrevemos a palavra sobre a qual queremos fazer a pesquisa, discriminamos a sua classe (nome, verbo, advérbio e adjetivo) e a língua em que queremos trabalhar. Existem também outras opções avançadas disponíveis, como estabelecer um limiar de frequência lexical (usado para filtrar palavras mais raras), um mínimo de ocorrências (usado para incluir ou excluir

⁶⁵ http://beta.visl.sdu.dk/constraint_grammar.html

⁶⁶ <http://linguateca.dei.uc.pt/Floresta/InicialFloresta.html>

⁶⁷ <http://linguateca.dei.uc.pt/Floresta/InicialFloresta><http://www.linguateca.pt/floresta/princip.html>

⁶⁸ Chama-se *treebank* a um *corpus*, cujas frases foram analisadas e etiquetadas sintaticamente ou semanticamente.

relações raras) e um mínimo de frequência relativa (estabelece um limiar de co-ocorrência). Em relação à procura por janela de palavras, nem o DeepDict nem o Wortschatz apresentam essa possibilidade. A figura 6 mostra como os resultados são apresentados, considerando a palavra *momento*.

momento (noun)		
countable, total of 261299 relations		
Hide Frequencies		
Premodifiers: 4.03:9 mau · 3.93:9 bom · 2.76:9 primeiro · 2.48:9 último · 3.33:8 preciso · 3.07:8 determinar · 1.82:9 grande · 2.71:7 breve · 1.58:8 actual · 1.07:8 dar · 1.52:7 excelente · 2.16:6 exacto · 0.7:7 único · 1.59:6 raro · 0.62:5 belo · 0.01:4 derradeiro	PP postmodifiers: 3.71:6 rel-ADV 4.01:7 de glória 3.2:7 de tensão 1.65:8 de forma 1.65:8 de vida 2.04:7 de votação 1.89:7 de carreira 2.83:6 de viragem 2.69:6 de euforia 1.67:7 de crise	Adjectival postmodifiers: 5.38:8 oportuno · 3.72:9 difícil · 4.58:8 crucial · 3.52:9 alto · 4.4:8 decisivo · 3.41:8 histórico · 3.15:8 certo · 3.33:7 exacto · 2.23:8 actual · 5.13:5 culminante · 1.92:8 importante · 2.99:6 emocionante · 1.78:7 crítico · 2.72:6 ideal · 1.5:7 presente · 1.85:6 delicado · 0.82:7 final · 0.71:7 político · 2.64:5 inesquecível · 2.37:5 hilariante · 0.17:7 único · 1.06:6 adequado · 1.97:5 propício · 1.91:5 marcante · 0.79:6 inicial

Figura 6 Resultados do DeepDict para a palavra *momento*.

4.2.2.3. CRPC

O *Corpus* de Referência do Português Contemporâneo (CRPC) é um *corpus* electrónico do Português Europeu e variedades (Brasil, Angola, Cabo Verde, Guiné-Bissau, Moçambique, São Tomé e Príncipe, Goa, Macau e Timor-Leste). Contém 311,4 milhões de palavras e abrange vários tipos de textos escritos (literários, jornalísticos, técnicos, etc.) e material oral (formal e informal). As pesquisas podem ser feitas online na subparte escrita do *corpus* (309 M). Os textos foram tokenizados usando o *LX tokenizer* (Branco, 2004). A etiquetagem morfosintática foi treinada com um etiquetador baseado em memórias (Daelemans, 1996). Finalmente, a lematização foi feita com a versão portuguesa do lematizador MBLEM (van den Bosch, 1999).

Para começar a pesquisa, temos que seleccionar o *corpus*, Português Europeu ou o *corpus* inteiro, incluindo as variedades. A procura é feita através da introdução de uma palavra na caixa de pesquisa (apesar da pesquisa também poder ser feita através de um padrão sintático). Podemos escolher o número de resultados por página e também podemos restringir a nossa pesquisa a um *corpus* específico

(direito, jornais, livros escolares, etc,). Esta é a única ferramenta, por nós analisada, que permite este tipo de especificação. Seguidamente, a ferramenta apresenta as ocorrências da palavra pesquisada em contexto. Podemos criar uma base de dados com as possíveis colocações, utilizando as listas de palavras co-ocorrentes. A figura 7, mostra os resultados para a palavra *momento*.

Collocation controls						
Collocation based on:	Word form	Statistic:	Log-likelihood			
Collocation window from:	3 to the Left	Collocation window to:	3 to the Right			
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5			
Filter results by:	specific collocate:	and/or tag:	(none)	Submit changed parameters	Go!	
There are 44,078 different words in your collocation database for "[lemma="momento"%c]". (Your query "{momento}" returned 131,381 matches in 80,302 different texts, ordered randomly) (30.967 seconds - retrieved from cache)						
No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood value
1	neste	174,686	444.47	47,567	34252	367279.952
2	em	3,031,659	7713.733	26,751	20915	29048.331
3	s	20,881,197	53129.979	86,393	44204	19061.758
4	no	1,893,953	4818.962	15,976	13094	16207.397
5	num	182,247	463.708	5,033	4423	15006.405
6	que	7,874,287	20035.284	38,014	28649	13201.751
7	partir	68,422	174.093	3,271	2912	13150.521
8	até	311,154	791.698	5,665	5103	12656.72
9	oportuno	6,560	16.691	1,289	1218	8928.535
10	Nesse	52,127	132.632	1,803	1515	6127.394

Figura 7 Resultados do CRPC para o substantivo *momento*

Podemos decidir o tamanho da janela de procura (por exemplo, três palavras para o lado esquerdo e duas palavras para o lado direito) que pretendemos usar. Os controlos permitem-nos também, ao contrário dos outros sistemas descritos, mudar as medidas estatísticas (informação mútua, rácio de verosimilhança, etc) e o limiar de frequência. Se clicarmos na frequência colocacional observada, podemos ver as frases onde essa combinação de duas palavras, em particular, ocorre. Contrariamente a outras ferramentas, podemos descarregar estas frases e todas as estatísticas. Em relação ao tempo de processamento da pesquisa da palavra *momento*, o Wortschatz e o DeepDict demoraram menos de seis segundos, enquanto que o CRPC demorou aproximadamente 46 segundos.

4.2.2.4. Sketch Engine

O Sketch Engine é um *software* pago que funciona como interface para fazer pesquisas online num *corpus*. Permite pesquisar concordâncias para qualquer

palavra, sintagma ou construção gramatical, e mostrar o comportamento gramatical e colocacional de qualquer palavra. Pesquisas mais específicas podem ser feitas com parâmetros mais complexos.

Possui 200 *corpora* em 82 línguas, mas é possível fazer *upload* de qualquer texto. Aliás, esta é a única ferramenta que possui esta mais valia. Estão disponíveis *corpora* para duas variedades do Português: Europeu e do Brasil. O analisador sintático utilizado, tal como no DeepDict, é o PALAVRAS.

Existem várias possibilidades para a utilização deste *software*, como criar um *thesaurus*, listas de palavras, comparações de ocorrências de duas palavras, mas, no nosso caso, vamos centrar-nos nas opções que permitem extrair colocações: “word sketch” e “concordance”. Elas são explicadas em mais detalhe, nos próximos parágrafos.

Word sketch

Selecionar a opção “word sketch”, permite-nos visualizar o comportamento gramatical e colocacional de uma determinada palavra. Escrevemos a palavra, na caixa de pesquisa, e especificamos a sua classe morfológica. Nas opções avançadas, escolhemos o *corpus* de Português Europeu (ptTenTen11 com 3.245.834.337 de tokens provenientes de páginas web). Podemos escolher uma frequência ou pontuação mínima (pontuação é entendida como LogDice⁶⁹) e o valor mínimo de co-ocorrência. Podemos também escolher agrupar⁷⁰, ou não, os colocativos e decidir o número máximo de itens nessas relações gramaticais. As colocações podem ser ordenadas de acordo com a sua saliência⁷¹ ou, simplesmente, através da frequência absoluta. O “word sketch”, para além de usar saliência estatística e lematização, utiliza também padrões gramaticais. Em vez de se olhar arbitrariamente para as

⁶⁹ Esta medida baseia-se só na frequência das palavras w1 e w2 e no bigrama w1 w2; não é afectada pelo tamanho do *corpus*.

⁷⁰ Se se escolher a opção de agrupar, os colocativos de um “word sketch” são apresentados em grupos de acordo com a sua posição no *thesaurus* distribucional. As palavras do *thesaurus* são aglomeradas de acordo com a sua pontuação de semelhança distribucional.

⁷¹ A saliência é uma medida estatística que quantifica quão saliente uma palavra ou um lema são, num determinado contexto, dada a frequência da palavra e o contexto. É medido através do LogDice.

palavras que ocorrem ao lado da palavra pesquisada, cada relação gramatical em que esta palavra participa é tida em consideração. Para Português, existem 11 tipos de relações gramaticais. O “word sketch” disponibiliza uma lista dos colocativos em cada relação gramatical na qual a palavra participa; os resultados são apresentados em grupos de relações sintáticas (*object_of* (objeto_de), *subject_of* (sujeito_de) ou *n_modifier* (nome_modificador). O CRPC também permite pesquisas através de padrões sintáticos (por exemplo, *N + Adj*) ou uma palavra (por exemplo *momento + Adj*, porém não usamos esta especificidade na nossa avaliação. O DeepDict não permite pesquisas através de padrões, mas os resultados são apresentados em grupos de relações (por exemplo pré-modificados, sintagmas preposicionais pós-modificadores e pós-modificadores adjetivais).

A figura 8 mostra os resultados para a palavra *momento*.

momento

(noun)

ptTenTen [2011, Freeling v3] freq = 1,915,768 (414.07 per million)

object_of	313,641	2.10	subject_of	105,975	1.60	n_modifier	445,627	2.10	modifies	11,932	0.10	y_o	26,131	5.00
viver	25,312	10.00	exigir	1,016	5.94	difícil	16,577	9.82	desfrute	102	7.83	importante	40	5.30
determinar	16,965	9.15	pensar	719	5.89	certo	12,353	9.34	lembrança	205	6.67	termômetros	35	5.26
aproveitar	10,311	8.97	parecer	876	5.62	histórico	16,751	9.31	recordação	78	6.47	acusado	38	4.32
proporcionar	7,476	8.50	poder	5,365	5.60	oportuno	6,823	8.92	diagrama	36	5.59	narrador	23	4.11
chegar	12,455	8.26	ir	13,297	5.33	importante	15,247	8.83	sucessão	88	5.56	senhor	190	3.84
atravessar	2,604	7.59	sentir	619	5.07	marcante	6,178	8.64	conveniência	46	5.56	melhor	87	3.78
relembra	1,991	7.49	chegar	1,099	5.01	decisivo	5,790	8.52	alternância	27	5.37	prioridade	67	3.65
registrar	2,917	7.29	corretar	224	4.96	atual	7,425	8.48	magnitude	33	5.36	suspeito	25	3.62
esperar	3,005	7.00	manifestar	207	4.95	lazer	5,911	8.46	retrato	65	5.04	4o	14	3.62
curtir	1,524	6.98	falar	705	4.82	crise	9,593	8.41	palco	117	4.93	único	48	3.56
aguardar	2,004	6.92	acontecer	689	4.73	especial	12,006	8.35	saudade	78	4.90	assaltante	15	3.48
reviver	1,067	6.71	requerer	241	4.71	inesquecível	4,480	8.27	protagonista	47	4.76	autor	169	3.45
recordar	1,142	6.62	surgir	373	4.62	reflexão	5,427	8.22	capricho	15	4.72	condutor	27	3.43
compartilhar	1,084	6.49	conseguir	632	4.58	crucial	3,956	8.09	circunstância	63	4.70	participante	64	3.26
haver	13,920	6.46	pronunciar	98	4.58	feliz	4,067	8.01	partilha	46	4.65	entrevistador	9	3.24
passar	4,702	6.35	servir	275	4.50	único	6,145	8.00	recriação	14	4.65	utilizador	32	3.23
retratar	862	6.22	ocorrer	490	4.48	descontração	3,347	7.92	vivência	44	4.63	5o	10	3.23
atar	1,051	6.19	cogitar	79	4.46	mágico	3,673	7.89	gerador	44	4.57	aluno	209	3.21
marcar	1,537	6.13	aparecer	298	4.45	alegria	4,171	7.88	espera	64	4.39	bandido	19	3.20
partilhar	744	6.05	aproximar	126	4.44	delicado	3,389	7.85	cristalização	9	4.31	motorista	38	3.14
representar	1,944	5.92	equipar	343	4.42	presente	4,154	7.64	evocação	10	4.29	espectador	19	3.13
dar	8,239	5.91	perguntar	138	4.41	crítico	3,115	7.43	fotografia	86	4.25	vereador	78	3.12
captar	671	5.86	coincidir	78	4.36	alto	3,784	7.41	eternidade	13	4.25	foco	55	3.10
propiciar	693	5.84	achar	329	4.36	emocionante	2,321	7.36	captura	23	4.18	2o	15	3.10
alternar	544	5.76	ser	26,805	4.31	bom	3,632	7.29	foto	225	4.14	jogador	125	3.10

Figura 8 Resultados do “word sketch” para a palavra *momento*.

Como já foi mencionado, as potenciais colocações são apresentadas em grupos, de acordo com a relação gramatical em que ocorrem. A primeira pontuação, ao lado de cada candidato a colocativo representa a frequência da palavra. Se clicarmos aí, podemos ver frases do *corpus* onde o potencial colocativo e a palavra procurada co-ocorrem. O segundo número representa a frequência dentro do grupo

de uma determinada relação. Tanto a lista de candidatos, como os exemplos do *corpus* podem ser descarregados em xml ou txt.

Concordance

A outra hipótese disponível no Sketch Engine para a extração de candidatos a colocação é a opção “concordance”. Escrevemos uma palavra na caixa de pesquisa, selecionamos o *corpus* e obtemos todas as ocorrências dessa mesma palavra no *corpus*. O passo seguinte será construir a lista de possíveis colocações. Isto é feito através da criação de uma lista de todas as palavras estatisticamente associadas com a palavra pesquisada (nó). Neste menu de pesquisa, podemos selecionar o atributo (palavra, classe morfológica, lema, etc.). Podemos especificar quantas palavras para os dois lados do nó devemos considerar, quando se fizer a procura dos possíveis candidatos. Por norma, costuma-se considerar as três palavras à direita e as três palavras à esquerda do nó. Podem-se estabelecer limiares de frequência dos candidatos no *corpus* inteiro ou num subtipo específico. Podemos escolher as estatísticas que queremos visualizar e a estatística com a qual queremos ordenar os nossos resultados. Mais detalhes sobre as várias métricas disponibilizadas podem ser encontrados em Kilgariff (2004). Esta é a ferramenta de extração que disponibiliza mais informação estatística, apesar do CRPC também utilizar muitas métricas.

A figura 9 mostra os resultados para a palavra *momento*. Para vermos os resultados ordenados segundo uma métrica específica, basta clicar sobre a métrica pretendida. Podemos visualizar as co-ocorrências no *corpus*, carregando em “P”. Tanto os resultados estatísticos, como o *corpus* resultante da pesquisa podem ser descarregados.

Collocation candidates									
Page 1		Go	Next >						
		Frequency	T-score	MI	MI3	log likelihood	min. sensitivity	logDice	MI.log_f
P N	em+este	73,828	270.593	7.922	40.266	680,817.958	0.020	9.186	88.805
P N	partir	10,944	103.031	6.047	32.883	70,557.757	0.005	7.206	56.241
P N	em+um	15,784	123.562	5.922	33.814	99,265.330	0.005	7.164	57.247
P N	proporcionar	3,488	58.473	6.654	30.190	25,327.746	0.008	7.154	54.279
P N	em+aquele	3,545	58.890	6.517	30.100	25,074.195	0.007	7.089	53.269
P N	qualquer	11,239	104.171	5.846	32.758	69,377.319	0.005	7.039	54.526
P N	actual	2,366	48.276	7.058	29.474	18,492.902	0.006	7.031	54.835
P N	convívio	1,711	41.211	8.074	29.556	15,787.846	0.005	6.957	60.118
P N	atravessar	1,852	42.757	7.275	28.985	15,030.761	0.005	6.864	54.744
P N	marcante	1,587	39.652	7.751	29.015	13,927.506	0.004	6.809	57.127
P N	difícil	3,423	57.578	5.978	29.460	21,678.899	0.005	6.743	48.655
P N	inesquecível	1,310	36.073	8.225	28.936	12,362.293	0.004	6.653	59.045

Figura 9 Resultados do “concordance” para a palavra *momento*.

4.2.3. Comparação entre os sistemas

Nas próximas secções, avaliaremos cada uma das ferramentas em relação às colocações que conseguiram extrair, os *corpora* criados através da extração e as estatísticas resultantes desta tarefa.

4.2.3.1. Extração de colocações

Como foi previamente dito, as pesquisadas em cada um dos sistemas foram feitas usando dez nomes como base das colocações. Foram analisados apenas os primeiros dez resultados da extração de cada uma das ferramentas. A configuração utilizada foi, quando o sistema assim o permitiu, uma frequência mínima de 5 e foram consideradas as 3 palavras à direita e à esquerda do nó. A tabela 10 mostra as colocações extraídas e validadas por um linguista. A última coluna mostra exemplos de colocações que foram extraídas por mais do que um sistema. Como podemos ver, o sistema que conseguiu extrair mais colocações foi o Word Sketch, demonstrando que combinar estatísticas com padrões gramaticais pode ser uma estratégia melhor para extrair colocações, do que, simplesmente, fazer contagens de ocorrências e frequências, como acontece com o Wortschatz. Em relação às palavras que foram selecionadas como possíveis colocativos, mas que na realidade não o eram, encontramos entre elas artigos (*o momento*), preposições (*na verdade*) e palavras que pertencem ao mesmo campo semântico (*autógrafo* e *fotografia*). Observámos

também que as palavras com baixa frequência são as que demonstram mais restrições na escolha dos colocativos, selecionando apenas um (*diagnóstico precoce* e *escolaridade obrigatória*). Como podemos ver na tabela 12, os cinco sistemas concordam na escolha desses colocativos (*precoce* e *obrigatório*).

Palavra	WoC	DD	CRPC	SE-1	SE-2	Exemplos
momento	0	5	1	2	1	oportuno (2); decisivo (2)
fim	0	1	1	4	3	lucrativo (4); pôr (2)
verdade	1	1	0	5	2	absoluto (3)
certeza	1	2	3	4	2	absoluto (4); ter (4)
força	1	0	0	3	3	de vontade (3)
adversidade	1	2	3	6	5	superar (3); climatérico (3)
autógrafo	0	0	2	6	2	pedir (3); dar (2)
fumador	2	4	2	3	2	inveterado (5); passivo (5)
diagnóstico	1	1	1	2	1	precoce (5)
escolaridade	1	1	2	1	2	obrigatório (5)
TOTAL	8	17	14	36	23	

Tabela 12 Comparação entre as extrações do Wortschatz (WoC), DeepDict (DD), CRPC, Word Sketch (SE-1) e Concordance (SE-2).

4.2.3.2. Recolha de *corpora* e estatísticas

Em primeiro lugar, o propósito da nossa investigação era apenas trabalhar com dados de Português Europeu. Esta necessidade foi colmatada, uma vez que todas as ferramentas distinguem entre variedades. O Deepdict, em particular, apenas possui dados de Português Europeu. Considerámos também importante que a natureza dos *corpora* usados não fosse muito repetitiva porque isso pode enviesar os resultados estatísticos, no entanto só encontrámos frases repetidas no CRPC. Queríamos também poder visualizar as colocações no contexto do *corpus*, mas este critério só não foi cumprido pelo DeepDict, que apenas mostra o colocativo em contexto e não a inteira colocação. Além de poder ter acesso às colocações em contexto, também gostávamos de poder descarregar as frases onde elas ocorrem, mas apenas o Sketch Engine e o CRPC permitem o *download* dos dados. O Sketch

Engine, aliás, é a única ferramenta que permite fazer *upload* de um qualquer *corpus* e usar as suas ferramentas para fazer pesquisas sobre esse mesmo *corpus*.

Em relação à informação estatística disponível, DeepDict calcula a frequência relativa de uma determinada relação e a frequência absoluta numa escala de 0 a 9. Utilizando a opção “word sketch” do Sketch Engine, podemos obter a frequência da combinatória e a frequência dentro de um determinado tipo (cluster). Na opção “concordance”, os candidatos a colocações podem ser ordenados de acordo com diferentes métricas. O CRPC dá informações sobre o número total de ocorrências em relação ao *corpus*, a frequência expectável do colocativo, a frequência observada do colocativo e o número de textos onde existiram ocorrências. Em seguida, podemos mudar a métrica e visualizar a ordenação dos resultados através de outras métricas. O Wortschatz só dispõe do número de ocorrências. Finalmente, também tivemos em consideração se a ferramenta era gratuita; das quatro ferramentas avaliadas apenas o Sketch Engine é a pagamento. O DeepDict é pago para algumas línguas, mas não para Português.

A tabela 13 resume a avaliação de todos os sistemas. Apesar do Sketch Engine dispor de duas modalidades de pesquisa, aqui serão apresentadas em conjunto, uma vez que são bastante semelhantes.

	WoC	DD	CRPC	SE
Tempo de processamento para a palavra momento	≈ 1s	≈ 2s	≈ 46s	≈ 6s
Limiar de frequência	X	✓	✓	✓
Procura por lema	X	✓	✓	✓
Window span	X	X	✓	✓
Procurar por padrão	X	X	✓	✓
Distinção de variedades	✓	✓	✓	✓
Não haver duplicações no <i>corpus</i>	✓	✓	X	✓

Mostrar as co-ocorrências no <i>corpus</i>	X	✓	✓	✓
Download dos resultados	X	X	✓	✓
Upload de <i>corpora</i>	X	X	X	✓
Uso de várias métricas	X	X	✓	✓
Gratuito	✓	✓	✓	X

Tabela 13 Comparação entre as características dos quatro sistemas.

Neste capítulo, realizámos uma breve avaliação dos quatro sistemas de extração de colocações mais utilizados para o Português Europeu: DeepDict, Sketch Engine, CRPC e Wortschatz.

Começámos por descrever cada uma das funcionalidades das ferramentas, seguidamente, utilizámos os quatro sistemas para extrair colocações, usando como base das potenciais colocações dez substantivos selecionados de um *corpus* de frequências. Recolhemos os *corpora* que foram criados através destas pesquisas e, finalmente, guardámos todas as informações estatísticas disponibilizadas pelos sistemas.

Tendo por base os passos anteriormente descritos, podemos concluir que o Sketch Engine, apesar de não ser gratuito, demonstra grande potencial como extrator de colocações, sendo capaz de encontrar mais colocações do que as outras ferramentas. Este extrator consegue melhores resultados do que os outros sistemas em vários aspetos da nossa avaliação, nomeadamente permitindo um ajuste bastante fino dos parâmetros de pesquisa (limiar de frequência, procura por janela de palavras, procura por lema ou padrão). O *corpus* do Português está dividido entre a variedade Europeia e do Brasil e, pelo que nos foi permitido observar até agora, não tem frases repetidas. Os resultados podem ser observados de acordo com diferentes métricas, permitindo perceber qual delas melhor serve os propósitos da investigação em questão⁷². Finalmente, as co-ocorrências são visualizadas no seu contexto, os resultados podem ser descarregados e é possível fazer *upload* de diferentes *corpora* para a aplicação. Devemos sublinhar que na impossibilidade de

⁷² McEnery (2006) sugere que a MI é geralmente usada para propósitos lexicográficos, enquanto que a MI3 é provavelmente mais útil num contexto de aprendizagem de segunda língua.

utilizar uma ferramenta a pagamento, o CRPC também demonstra bons resultados, tendo apenas a desvantagem de existirem repetições no *corpus*.

Apesar de estas ferramentas demonstrarem resultados promissores para a extração de colocações de *corpora*, elas não servem o propósito da nossa investigação. Em primeiro lugar, a maioria destas ferramentas permite apenas fazer pesquisas partindo da base da colocação e não encontrar toda a colocação, como é do nosso interesse. Em segundo lugar, nesta dissertação trabalhamos com textos que apresentam vários erros gramaticais e ortográficos que dificultam o seu tratamento automático e consequentemente a extração automática de colocações. No entanto, esperamos que esta avaliação de ferramentas de extração seja elucidativa para quem pensa usá-las e que sirva também como contributo para um possível melhoramento das mesmas.

Capítulo 5. Processo de anotação das colocações

Como foi previamente dito, apesar das ferramentas automáticas de extração de colocações demonstrarem resultados promissores, para a nossa investigação estas não se configuraram como a melhor opção. Assim, o trabalho de identificação das colocações no *corpus* de aprendentes foi realizado manualmente. Apesar da identificação das colocações ser fulcral para o nosso trabalho, foram necessárias outras etapas no processo de anotação que descreveremos neste capítulo.

Nas primeiras secções, começamos por apresentar alguns aspetos da definição de colocação que usámos no processo de anotação das colocações. Na secção 5.1, estabelecemos as orientações gerais seguidas, em 5.2 dilucidamos aspetos importantes para a definição de colocação e por último, em 5.3, elencamos os critérios definitórios usados.

Na secção 5.4, descrevemos o processo de anotação manual das colocações no *corpus*, apresentamos a taxonomia de classificação de erros que definimos e utilizámos. Por último, apontamos algumas das dificuldades com as quais nos deparámos na identificação das colocações, na classificação das colocações corretas e incorretas e na categorização dos erros presentes nestas últimas.

5.1. Orientações gerais

O interesse sobre o estudo das colocações é grande e existem cada vez mais trabalhos sobre elas, porém continua a não existir uma definição unanimemente

aceite. Apesar do debate sobre as fronteiras entre os vários tipos de estruturas fraseológicas ser de grande interesse teórico, optámos por não lhe atribuir um papel central nesta dissertação. Alonso Ramos tem uma posição semelhante, defendendo mesmo que o debate teórico sobre se uma construção é ou não colocação é pouco útil (2006: 29).

Neste trabalho, como orientação geral, para a identificação e descrição das colocações seguiremos o ponto de vista da produção do aprendente estrangeiro. Sendo as colocações muitas vezes transparentes, as dificuldades dos aprendentes de uma L2 encontrar-se-ão na codificação e não na decodificação. Assumimos a visão essencialmente prática de Alonso Ramos, e adotamos o ponto de vista do lexicógrafo que desenvolve um dicionário de colocações e que tem como referência um aprendente de uma língua segunda, ou seja, o papel mais importante é atribuído ao aluno e às suas necessidades. Assim uma colocação corresponde a uma co-ocorrência de palavras que os alunos não esperam encontrar juntas e que, portanto, deverá ser registada num dicionário.

Para além da visão do aprendente, consideramos que uma perspetiva contrastiva entre línguas pode também ser uma boa ferramenta para encontrar as colocações. Mesmo um falante nativo nem sempre tem consciência do carácter de uma determinada expressão, uma vez que domina a sua língua com todos os seus automatismos, muitas vezes é a perspetiva contrastiva que leva à tomada de consciência. Assim, a comparação entre línguas pode tornar visível a falta de liberdade na combinatória das unidades lexicais, podendo esta passar despercebida quando se tem em conta uma perspetiva apenas monolíngue.

A comparação entre línguas, mesmo sendo elas próximas (ex.: Francês, Espanhol e Português), pode demonstrar vários tipos de diferenças na construção das colocações. Nomeadamente, como podemos ver nos exemplos de Alonso Ramos (2006:35) em espanhol, colocações muito semelhantes podem ter significados diferentes (**vontade espantosa* vs *ganas espantosas*), uma colocação pode, em outra língua, ser equivalente a uma unidade lexical e vice-versa (*tomber amoureux* (fr.) – *namorarse* (esp.)) ou, como neste exemplo em francês, ter uma estrutura actancial diferente (*avoir (quelqu'un) en haine* (literalmente, *ter alguém em ódio*)). Finalmente, esta abordagem contrastiva tem como objetivo não só chamar a

atenção dos alunos para as diferenças entre as colocações, mas também encontrar semelhanças entre as línguas.

Depois de identificadas as duas linhas norteadoras, na seguinte secção vamos apresentar os aspetos definitórios mais importantes para a identificação das colocações.

5.2. Noções necessárias para a definição de colocação

Nesta secção, vamos dilucidar alguns aspetos das colocações que serão necessários para o processo de anotação das colocações nos *corpora* de aprendizagem.

Começamos por abordar o carácter mais ou menos transparente e arbitrário das colocações. Seguidamente, centramo-nos no papel que a co-ocorrência restrita e a frequência assumem na definição de colocação. Distinguimos as colocações das combinatórias com as quais fazem fronteira: as combinatórias livres e as locuções. Por último, descrevemos os constituintes da colocação e as configurações sintáticas que estas podem assumir.

5.2.1. Transparência (não fixidez semântica)

No nosso entender, a transparência das colocações deve ser interpretada à luz da aprendizagem de uma L2 e tendo em consideração a distância entre línguas. Assumindo o ponto de vista da compreensão, uma colocação pode ser mais ou menos transparente, dependendo se o sujeito conhece o colocativo em co-ocorrência com uma determinada base. *Sorriso amarelo* é pouco transparente para um aprendente que não conhece o sentido do adjetivo *amarelo* nessa colocação; não obstante, *vinho branco* ou o seu equivalente em francês *vin blanc* são completamente transparentes. Em contrapartida, do ponto de vista da produção, o falante não nativo não pode prever quando é que um colocativo coincide com o seu equivalente na língua materna e quando não: *vin rouge*, em francês, traduz-se para *vinho tinto* e não para *vinho vermelho*.

Seguimos a perspectiva de Alonso Ramos (2006), para quem a definição de

colocação não se centra na sua opacidade ou transparência, mas na seleção restrita de certas unidades lexicais para expressar um dado sentido em combinação com outras unidades lexicais. Assumindo esta perspectiva mais heterogénea, incluímos na categoria de colocação sintagmas completamente transparentes (*ter fome*) e mais opacos (*sorriso amarelo*). Em ambos existe alguma restrição, apesar de *ter* poder ter uma combinatória muito mais ampla. Também Mel'cuk (2003) inclui na categoria de colocação expressões como *peur bleue* ou *colère noire*, onde o adjetivo não é imediatamente interpretável.

Podemos então dizer que nem todas as colocações têm o mesmo grau de transparência semântica. A transparência ou opacidade semântica não caracterizam pois toda a classe.

5.2.2. Arbitrariedade (não previsibilidade)

Para Hausmann uma combinatória é previsível se é regida pelas regras do sistema, em contrapartida, é imprevisível se as suas restrições de combinatória não se explicam através de leis gramaticais e semânticas. Na perspetiva da produção/codificação, neste exemplo em espanhol, a seleção de *intachable* não é livre, uma vez que não podemos escolher livremente o adjetivo que se combina com *comportamiento* para expressar o sentido de “tão bom que não pode ser criticado”. Trata-se então de uma combinatória restrita uma vez que este adjetivo é apenas selecionado para comportamentos mas não para *gusto* (*impecable*) ou *honor* (*immaculado*) (Alonso Ramos, 1993: 151).

Diferentes autores acreditam que a arbitrariedade na construção de colocações nem sempre é total e tentam encontrar regras semânticas para justificar determinadas combinações em colocações. A associação lexical por vezes é favorecida por certas relações semânticas entre base e colocativo, como as séries de intensidade de Bally (1909). Porém as regras semânticas que se podem encontrar são tão complexas e tão hiperespecíficas que não seriam úteis nem para os aprendentes de uma língua nem para um sistema automático.

Na perspetiva de Alonso Ramos, o critério da imprevisibilidade resulta pouco eficaz. Atentemos no exemplo *grande distância*. Trata-se de uma combinatória

bastante previsível do ponto de vista da produção, no entanto, um não-nativo pode não saber como intensificar “distância” e usar por exemplo o adjetivo *largo* (**larga distancia*). Novamente, a autora reitera que a característica mais importante para a definição de colocação é a restrição na seleção lexical, ou seja, a base determina a seleção do elemento que se combina com ela, o colocativo, para assim veicular um determinado significado. Em *quebrar uma promessa*, *quebrar* é o nome selecionado por *promessa* para veicular o significado “deixar de ter”. Claro que existe a possibilidade de haver uma comutação e o colocativo ser substituído por outros itens lexicais. No nosso entender, a ênfase deve recair na imprevisibilidade ou arbitrariedade que caracteriza a escolha do colocativo e esta arbitrariedade deve ser vista de uma perspectiva contrastiva. Uma combinação aparentemente trivial ou transparente pode ser considerada uma colocação, se a sua tradução literal não é correta noutra língua. O significado veiculado por uma colocação pode não ser constituído pelo mesmo colocativo noutra língua. Isto pode acontecer mesmo em línguas próximas, por exemplo em espanhol diz-se *jugar un papel*, em português uma tradução literal (*jogar um papel*) não é correta, mas sim *desempenhar/fazer um papel*.

5.2.3. Co-ocorrência restrita

Para Mel’cuk, trata-se de uma colocação sempre que produzindo uma combinatória de duas palavras, a escolha de uma delas não for livre, mas restrita e imposta pela outra. Podemos dizer que são usadas pelos falantes de uma língua como blocos pré-fabricados. É o uso que as torna combinações estáveis, o que justifica a restrição combinatória, embora um dos membros da colocação seja mais restrito que o outro. Esta restrição pode desencadear uma especialização semântica, neutralizando-o semanticamente ou atribuindo-lhe um significado metafórico.

Efetivamente, como exemplifica Corpas Pastor (2001: 93 *apud* Cardoso, 2012: 33), quando o verbo espanhol *prestar* se coloca com os substantivos *ayuda*, *apoyo*, *atención*, *asistencia*, *auxilio*, *colaboración*, *servicio*, *declaración*, *juramento*, *atención* ou *interés*, entre outros, perde a sua identidade semântica básica – ceder algo a outra pessoa, com o sentido de que nos seja devolvido por ela – para adquirir um

significado geral e gramaticalizado, funcional e auxiliar. Já o verbo *levantar* adquire um significado metafórico quando se combina com substantivos como *castigo, sanción, prohibición, veda, interdicción*. Para Mendívil (1991: 717 *apud* Cardoso, 2012: 34), trata-se de preferências lexicais, já que em várias colocações podem-se utilizar outros verbos de conteúdo sinonímico (*superar/vencer uma dificuldade*).

Porém nem todas as colocações admitem sinónimos. Atentemos nos exemplos de Muñiz (2004) e Blanco (2006), aqui traduzidos para português, uma combinação como *a ovelha bale* ou *o barco atraca* são restritas semanticamente. Esta combinação é bastante previsível, pois o verbo *balir* está semanticamente restrito a *ovelha* e o mesmo acontece com *barco* e *atracar*. São combinatórias composicionais porque os seus significados são adquiridos através do significado das partes. No caso de *the problem lies* (o problema reside) a restrição é lexical, mas, em ambos, trata-se de colocações.

É certo que as colocações constituem, prototipicamente, um fenómeno de natureza essencialmente lexical, não semântica: o colocativo é selecionado lexicalmente pela base, no sentido em que se exige um unidade lexical em concreto. *Coger* e *tomar* (em espanhol) são sinónimos em combinatórias livres, mas nas colocações não são substituíveis. Em português *levar, transportar* ou *acompanhar* são sinónimos. No entanto, dizemos *levar o país à ruína* e não **transportar/*acompanhar o país à ruína*. Não se trata pois de uma decisão do falante: é a língua que exige ao falantes que quer expressar “som típico produzido pelas ovelhas” que diga *balir*. No nosso entender, o ponto em comum entre estas estruturas é que ambas podem causar problemas a um aprendente estrangeiro da língua e ambas devem ser descritas na entrada lexicográfica do nome, tendo em mente a produção.

5.2.4. Frequência

Para Sinclair a frequência é de primordial importância para a definição de colocação: "collocation" is the co-occurrence of two items in a text within a specified environment. "Significant collocation" is a regular collocation between items, such as that they co-occur more often than their respective frequencies and the length of

text in which they appear would predict” (2004 [1970]: 10).

Apesar de, anos mais tarde, semantistas e lexicógrafos explorarem amplamente a posição de fronteira das colocações entre as expressão idiomática e as combinação livres, outros, porém, em vez de se focarem nesta posição intermédia da colocação, continuam a centrar-se no seu carácter habitual, usual e frequente. Cruse (1986: 40), por exemplo, define colocação como sequências de unidades léxicas que co-ocorrem habitualmente, mas que, não obstante, são completamente transparentes no sentido em que cada constituinte lexical é também um constituinte semântico (ex.: *heavy drinker, torrential rain, fine weather*)

Para alguns autores a co-ocorrência frequente não é um traço exclusivo das colocações, ou seja, existem combinações com alta frequência que não são colocações.

El hecho de que dos lexemas coocurran frecuentemente no es prueba de que exista una colocación. La coocurrencia de los lexemas puede estar determinada por su significado y esto es independiente de que ambos lexemas aparezcan frecuentemente en los textos. (Alonso Ramos, 1994-95: 14-15)

Assim a frequência não é um critério determinante para decidir se uma estrutura é ou não uma colocação.

Una combinación como mirar un árbol será probable e incluso frecuente pero no por ello esta combinación es una colocación. Está formada según las reglas de la sintaxis, es completamente libre: el verbo mirar se puede combinar con todo lo que puede ser mirado. Y esta libertad abarca a todas las lenguas en donde exista un verbo equivalente a mirar. (Alonso Ramos, 1994-95: 15)

Também nós partilhamos a opinião da autora *supra* citada que, apoiando-se em Hausmann, defende que o facto de uma combinação ser frequente e possível não determina que seja uma colocação, na medida em que as palavras que compõem a unidade lexical podem ser substituídas por palavras semanticamente equivalentes.

5.2.5. Construções de fronteira

Na fronteira entre a pré-construção e a combinatória livre, as colocações constituem uma problemática real para a linguística e para a linguística aplicada, uma vez que este fenómeno continua mal definido. Nestas subsecções, apresentamos algumas das diferenças que distinguem as colocações das combinatórias livres e das construções completamente fixas.

5.2.5.1. Diferença entre combinatória livre e colocação

Hausmann estabelece a diferença entre colocações e combinações livres através da visão saussuriana de *langue* e *parole*. Segundo o autor, como as colocações apresentam uma certa fixidez e algum grau de especialidade semântica representam “une unité, non de la parole, mais de la langue”. Já uma estrutura, como *comprar um livro*, trata-se de uma construção não específica e, como tal, pertence à *parole* (Hausmann, 1985: 118).

Mel’cuk (2008 *apud* Alonso Ramos, 2006) também distingue as colocações das combinatórias livres, fazendo uma análise do comportamento das colocações no eixo sintagmático e paradigmático. Segundo o autor, as colocações distinguem-se dos sintagmas livres porque violam a propriedade paradigmática que diz respeito à seleção⁷³, ou seja, um dos constituintes da colocação é selecionado em função do seu sentido, enquanto que o outro é selecionado em função do primeiro. *Atracar um barco* viola somente a propriedade paradigmática: o sentido de “aproximar-se da costa” expressa-se com *atracar* porque se predica de *barco*. No entanto, a seleção da unidade lexical *atracar* não é independente de *barco*. No entanto, a propriedade sintagmática é perfeitamente respeitada: os significados dos dois elementos do sintagma unem-se respeitando as regras do espanhol.

Segundo Grossmann e Tutin (2002: 7), nas associações completamente livres do tipo de *argument intéressant, envie de chocolat*, o sentido não é somente

⁷³ O sintagma AB é livre no eixo paradigmático se a seleção dos constituintes A e B pelos falantes for não restrita, ou seja, A é selecionado independentemente de B e vice versa. No lugar de A e B, o falante pode utilizar outros signos A' e B' que significam aproximadamente o mesmo.

analisável, mas é também previsível. O mesmo não acontece com as colocações, uma vez que a escolha do colocativo é muitas vezes imprevisível.

As combinatórias livres apresentam uma maior flexibilidade combinatória. Por oposição às combinações livres, nas colocações o colocativo não pode ser substituído por qualquer outro para expressar um determinado sentido. Em *chuva torrencial*, *fumador inveterado*, embora os colocativos tenham um valor de intensidade, a eleição do adjetivo é determinada pelo uso. Podem ser aceites sinónimos, mas mesmo assim a sua seleção é restrita (*cometer um crime*, *perpetrar um crime*/**executar um crime*). Já as combinatórias livres apresentam um menor grau de restrição combinatória determinada pelo uso (Cardoso, 2012: 38).

Também do ponto de vista sintático e morfológico, as colocações apresentam mais restrições. Por vezes constroem-se de maneira atípica (muitas vezes arcaica), em comparação com as regras da língua, como atesta, por exemplo, a ausência frequente do determinante nas construções com verbo suporte: em português *ter fome* ou *ter sono* e nos exemplos em francês dos autores Grossmann e Tutin (2002: 16) *avoir faim*, *avoir soif*, *rendre visite*. Apesar deste comportamento não ser sistemático (*perpétrer un délit*).

As colocações apresentam também menos flexibilidade ao nível da distribuição sintática do que as combinatórias livres. Por exemplo, os colocativos adjetivos (epítetos) que têm um funcionamento predicativo não podem aparecer como atributos. Numa combinação livre, esse contexto é praticamente sempre possível com os adjetivos que têm essa função semântica (Gross, 1988). Atentemos no comportamento das colocações *peur bleue* e da combinação livre *peur immense* em francês e em português a colocação *ódio mortal* e a combinação livre *grande ódio* em contexto.

Jean a eu une peur bleue.

O João sentiu um ódio mortal.

Jean a ressenti une peur immense.

O João sentiu um grande ódio.

**La peur de Jean a été bleue.*

**O ódio que o João sentiu era mortal.*

La peur de Jean a été immense.

O ódio que o João sentia era grande.

Como podemos ver no terceiro conjunto de frases, as colocativos em causa não funcionam como atributos, não acontecendo o mesmo com os adjetivos das combinatórias livres (quarto grupo de frases). No entanto, este comportamento nem sempre é a regra. A colocação *steak bleu* apresenta menor fixidez sintática que *peur bleue*. *Léo aime le steak quand il est bleu*. Também o mesmo acontece com *argumento infundado*: *O argumento que o Miguel utilizou era infundado*.

Comparemos agora o comportamento das combinatórias livres com as colocações verbais, nomeadamente as colocações com verbo suporte, face a alterações sintáticas operadas nas frases onde estas se encontram. Para esta observação vamos utilizar as expressões *ficar em pânico* (construção de verbo suporte) e *ficar em casa* (*ficar* é um verbo pleno):

- Restrições na anaforização:

Ele ficou em casa. Ele ficou lá.

Ele ficou em pânico. *Ele ficou lá/aí.

- Restrições nos determinantes

Ele ficou na tua casa.

*Ele ficou no teu pânico.

- Impossibilidade de interrogação

Onde é que ele ficou? Em casa.

*Onde é que ele ficou? Em pânico.

Segundo Cardoso (2012: 36), nas colocações de V + N existe uma certa tipicidade

estabelecida entre o nome e o verbo. Entre *tocar* ou *afinar* e *guitarra* existe essa relação, enquanto que entre *limpar* ou *guardar* e *guitarra* não existe uma relação típica. Ao contrário das colocações, que expressam uma relação típica entre os seus componentes, as combinações livres carecem de tal relação.

Por último, as colocações são memorizadas pelos falantes de uma língua e são psicologicamente salientes (Bahns, 1993: 57), portanto são reproduzidas e não produzidas como as combinatórias livres.

5.2.5.2. Diferença entre locução e colocação

Segundo Grosman e Tutin (2012: 7), as colocações opõem-se intuitivamente às expressões completamente fixas, como *pomme de terre* ou *nid d'ange*, nas quais o significado do todo é dificilmente previsível partindo do significado das partes e a interpretação semântica não se faz de maneira composicional. Estas estruturas beneficiam de múltiplos nomes: “*expression figées*”, “*locutions*”, “*unités phraséologiques*” (Bally, 1909), “*phrasèmes complets*” (Mel'cuk et al., 1995), “*unités polylexématiques*” (Corbin, 1997).

No plano semântico distinguem-se claramente das colocações⁷⁴. A principal característica semântica das locuções é a não-composicionalidade do sentido da expressão fixa, isto é, o significado do todo não pode ser calculado a partir do significado das partes, perdendo cada elemento a sua função referencial. Na expressão *bater as botas*, o significado da expressão não pode ser inferido através do significado de *bater* e de *botas*. O referente da expressão não alude nem a *bater* nem a *botas*. Os processos linguísticos, como a metáfora ou a metonímia podem muitas vezes explicar a génese destas expressões. Podemos concluir que a colocação normalmente tem um significado, enquanto que a locução tem dois: o literal (*deitar a toalha ao chão*) e o idiomático (*desistir*). No entanto, existem algumas colocações menos transparentes, que apresentam mais dificuldades na descodificação, não se caracterizando pela composicionalidade semântica, apesar da composicionalidade formal lhes permitir alguma flexibilidade morfológica e sintática (Cardoso, 2012: 42),

⁷⁴ Embora algumas colocações possam ser mais opacas, como *sorriso amarelo*, ainda assim o significado de um dos constituintes é ainda perceptível (sorriso).

nomeadamente as colocações opacas que mencionam Grossman e Tutin (2002:5-6),

Existem colocações que também são locuções, podendo admitir interpretações metafóricas (neste caso, o seu significado não é obtido pela soma das suas partes). Uma vez usadas em sentido figurado, as colocações transformam-se em locuções, de acordo com uma perspectiva semântica. *Abrir uma fresta* tem um sentido literal, mas noutro contexto pode ser interpretado como “conseguir uma oportunidade”. Neste caso, a colocação ganha uma interpretação idiomática. Na teoria de Wotjak (1998: 258 *apud* Cardoso: 37), é natural que a colocação se converta numa expressão idiomática, uma vez que se encontra em vias de lexicalização.

Apesar de existirem estas exceções, as colocações, por norma, são composicionais e podem ser variáveis, do ponto de vista combinatório e morfológico. Em relação às transformações sintáticas, as colocações apresentam uma maior flexibilidade formal. As locuções não admitem transformações sintáticas, como a modificação adjetival, a nominalização ou a transformação para a passiva. Já as colocações admitem a substituição de um componente, nominalização, a pronominalização do substantivo em função de complemento direto⁷⁵ (*desempenhar um cargo: desempenhá-lo*), tal como a modificação por um adjetivo (*fazer uma aterragem forçada*) ou a transformação para a passiva (*o órgão foi transplantado*); dada a sua composicionalidade sintática, um dos seus componentes pode ser extraído e funcionar na frase autonomamente (Koike, 2001: 34 *apud* Cardoso, 2012: 41). As locuções perdem a sua idiomaticidade na presença de outras combinações, não admitindo a sinonímia. O mesmo não acontece com as colocações (*matar/saciar a sede*), no entanto existem algumas exceções. O sentido idiomático das locuções é obtido dentro de uma determinada fixidez estrutural, algo que as distingue das colocações, sendo estas parcialmente fixas e as locuções quase totalmente fixas. Na visão de Corpas Pastor (2001: 94), isto implica que “uma determinada colocação possa apresentar variantes diatópicas, diafásicas e diastráticas, bem como corresponder a modificações particulares do falante para obter determinado efeito estilístico”.

⁷⁵ Porém algumas locuções também podem.

No que toca ao número de constituintes, as colocações são compostas, normalmente, por dois lexemas, as locuções admitem construções com um maior número de constituintes.

Resumindo, as colocações podem ser distinguidas das locuções através de critérios de composicionalidade formal, da admissão de transformações sintáticas e de variações e, por último, através do número de lexemas.

5.2.6. Carácter binário e assimétrico

Para Hausmann (1985), tal como para Mel'cuk (2003) e Grossmann e Tutin (2002) a colocação é essencialmente composta por duas palavras ou dois itens lexicais⁷⁶: um item lexical (palavra simples ou frasema⁷⁷) L e de um constituinte C. Nesta relação, como já foi dito:

- **C** (o colocativo) é selecionado na produção para exprimir um sentido ganho em co-ocorrência com **L** (a base);
- O significado de **L** é o habitual.

Como vimos, as colocações são determinadas pelo uso e representam relações sintagmáticas preferenciais entre dois lexemas num padrão sintático específico (cf. 5.2.7.). Os dois contribuem para a combinação mas não têm o mesmo estatuto: “Dans la collocation, le statut des deux partenaires combines n’est pas égal” (Hausmann, 1989: 1010).

Codificar uma colocação parte necessariamente da base, ou seja, a realização do sentido do colocativo depende funcionalmente da base. A base é semanticamente autónoma e livremente selecionada de acordo com o seu significado habitual pelo falante. O segundo elemento, o colocativo, é semanticamente dependente da base. Seguimos então a perspetiva de Mel'cuk (2003), para quem a relação entre os constituintes da colocação é claramente assimétrica, uma vez que a base conserva o seu sentido habitual. Nos exemplos

⁷⁶ Hausmann (2004; 2007) refere as “collocational chains” (pay close attention) que têm mais do que dois constituintes.

⁷⁷ Ou seja, uma expressão completamente fixa na terminologia mel'cukiana.

chuva torrencial, *torrencial* é o colocativo selecionado pela base *chuva* para caracterizar uma chuva particularmente forte.

Para além destes dois elementos, tal como em Tutin (2008:9), consideramos que as colocações podem ser compostas por duas palavras lexicais e uma gramatical. No nosso entender, as preposições devem então fazer parte da descrição das colocações, enquanto que palavras puramente funcionais, apesar de deverem ser descritas, não são centrais para o nosso trabalho de análise de erros.

Em relação à distância entre os constituintes da colocação, estes não têm necessariamente de se encontrar juntos, podendo surgir uma ou mais palavras entre eles. Este espaço chama-se “distância colocacional” (*collocational span*). Muitos autores defendem a distância de quatro palavras antes e depois da palavra núcleo, mas também há quem defenda cinco. Nós seguimos a posição de (Koike, 2001: 146 *apud* Cardoso, 2012:32), segundo o qual o critério da distância colocacional carece de fundamento linguístico pois a colocação não demonstra um vínculo puramente formal medível através da distância, mas sim um vínculo baseado no significado lexical.

5.2.7. Classificação sintática das colocações

Apesar da arbitrariedade muitas vezes caracterizar as colocações, é de notar que as elas seguem padrões sintáticos precisos. Por exemplo, as colocações construídas em torno de um nome incluem adjetivos ou verbos, mas provavelmente não surgirão advérbios ou conjunções. Hausmann, notando nesta regularidade apresentada pelas colocações, apresenta também uma primeira classificação dos tipos de colocação. Para tal considera as categorias gramaticais dos colocados e as relações sintáticas que se estabelecem entre si.

Estando a base e o colocativo sintaticamente relacionados, é possível encontrar construções prototípicas em várias línguas. Hausmann (1989: 1010) descreve os seguintes seis tipos, que podem servir para várias línguas:

- a) Nome + Adjetivo: *célibataire endurci*, *fumador inveterado*, *sono profundo*

- b) Nome (sujeito) + Verbo: *la colère s'apaise, vento/tempestade amaina, problema reside*
- c) Verbo + Nome (objeto): *retirer de l'argent, assumir responsabilidades, dar um abraço, desempenhar um cargo*
- d) Verbo + Advérbio: *il pleut à verse opor-se categoricamente, amar perdidamente*
- e) Adjetivo + Advérbio: *grièvement blessé, profundamente agradecido, fermement convencido, estreitamente associado*
- f) Nome + Preposição + Nome: *une bouffée de colère,*
- g) Adjetivo + Preposição + Nome – *louco de ciúmes, verde de raiva*
- h) Verbo + Preposição + Nome – *jogar às cartas, levar ao forno, morrer de tédio.*

No caso de a), a ordem não é relevante, podendo esta ser Nome + Adjetivo (*solteiro inveterado, pele clara*), como Adjetivo + Nome (*grande quantidade, melhor amigo*). A diferença entre b) e c) está relacionada com a função sintática do substantivo, sendo num caso sujeito (*sol nasce, tempo voa*) e noutro objeto (*levantar dinheiro, fazer dieta*). Outros exemplos de Verbo + Advérbio e Adjetivo + advérbio são *chover torrencialmente* e *gravemente ferido*, respetivamente. Outro exemplo de Nome + Preposição + Nome em português é *cabelo aos caracóis, mudar de vida* ou *procurar na internet*.

Para além deste elenco de formas sintáticas, incluímos também: Adjetivo + Preposição + Nome (*louco de ciúmes*), seguindo a perspetiva de Athayde (2009) e Verbo + Preposição + Nome (*pôr em causa, dar um passeio*). Este último subtipo é também mencionado por Corpas Pastor (1996: 66-76).

Para além das colocações formadas por duas unidades lexicais, individuamos também as colocações complexas, compostas por uma unidade lexical e uma locução, por exemplo *transportes públicos frequentes* ou *participar no programa Erasmus*.

Em relação à contiguidade dos dois elementos da colocação, esta nem sempre ocorre (Heid, 1994). Certas associações de Nome + Adjetivo (epíteto), como *cabelo grisalho* podem aparecer separadamente (o cabelo dele é grisalho): os

elementos da colocação podem formar um sintagma, mas isso pode também não acontecer. Também nas colocações verbais, como já foi visto, certas alternâncias sintáticas são possíveis (passivas, orações relativas), se bem que este caso seja mais comum nas combinatórias.

Apesar das associações de partes do discurso não serem arbitrárias, por exemplo um advérbio e um nome, permite formar um sintagma nominal (Adjetivo + Nome) ou um sintagma verbal (Verbo + Nome), pode também acontecer que a combinação do colocativo com a base, apesar de contíguos, não formem um constituinte no sentido tradicional. É, por exemplo, o caso da associação Nome (Sujeito) + Verbo em *le problème reside (dans cet aspect)*, na qual o sujeito e o verbo não formam uma unidade. Como já várias vezes foi sublinhado (Heid, 1994; Iordansjake, Kim & Polguère, 1996), a base da colocação não é sistematicamente a cabeça do sintagma, como podemos constatar nos seguintes exemplos: *rendre visite*, base – *visite*, cabeça do sintagma – *rendre*; *vert de peur*, base – *peur*, cabeça do sintagma – *vert* (Grossman e Tutin, 2012).

5.3. Critérios usados para a definição de colocação

Depois de termos apresentado os aspetos mais importantes a ter em consideração quando delimitamos o conceito de colocação, definimos critérios para nos ajudarem na complexa tarefa de identificação das colocações. São esses critérios que sistematizamos nesta secção.

1. A definição usada neste trabalho será, como em em Athayde (2009), de acordo com Grossmann e Tutin (2002: 3), como “a co-ocorrência lexical privilegiada de dois (ou mais) elementos linguísticos que entre si estabelecem uma relação sintáctica”.
2. Os dois lexemas que constituem a colocação estabelecem entre si uma relação hipotáctica (ou de dependência). A base é a palavra que determina a escolha do outro elemento co-ocorrente – o colocativo. E o colocativo é o constituinte determinado. No

exemplo, *chuva torrencial*, *torrencial* é o colocativo selecionado pela base que acrescenta um novo traço semântico à base *chuva*.

3. Normalmente são semanticamente transparentes e, como tal é facilmente apreendido o seu significado, a escolha dos seus constituintes é normalmente arbitrária e dificilmente previsível e, por essa razão, a sua produção pode criar dúvidas aos aprendentes de uma língua estrangeira.
4. Imaginando um *continuum*, que de um lado engloba estruturas livres e, do outro, expressões completamente fixas, entenderemos as colocações como estruturas que se encontram entre estes dois extremos.
5. Estão guardadas na mente dos falantes de uma comunidade tal com as unidades lexicais e são reproduzidas como tal.
6. As colocações seguem estruturas sintáticas regulares, assim estudaremos as colocações que assumem os seguintes padrões sintáticos: V + N (sujeito), V + N (objecto), N + A, Adv + Adj, V + Adv, V + Prep + N, A + Prep + N e N + Prep + N.
7. Do ponto de vista sintático, são bastante flexíveis, admitindo várias alterações à sua estrutura: podem apresentar variantes, nas quais se altere o valor categorial dos elementos constituintes, mas não o seu valor lexical (*ajuste de contas /ajustar as contas*).
8. Podem apresentar algum grau de variação (*conseguir/ter uma oportunidade*).

Resumindo, trata-se de estruturas de fronteira, que se encontram entre a combinatória livre e a combinatória fixa, o seu significado é normalmente transparente, existe uma dissimetria entre os seus dois constituintes, a escolha do colocativo é muitas vezes imprevisível, geralmente seguem uma estrutura sintática regular e, por último, têm um uso recorrente nas línguas.

5.4. Anotação manual de colocações

Depois de termos definido e delimitado o conceito de colocação, neste subcapítulo, vamos descrever todo o processo de anotação das colocações. Começamos, na secção 5.4.1., por explicar o processo de identificação manual das colocações presentes no *corpus* descrito na secção 4.1. Na secção 5.4.2., apresentamos a taxonomia utilizada para classificar os erros das colocações. Por último, na secção 5.4.3. apresentamos alguns dos problemas que individuámos no desenrolar do processo de identificação das colocações e classificação dos seus erros.

5.4.1. Identificação de colocações

Cada colocação presente no *corpus* foi identificada e assinalada como correta ou incorreta. Às colocações incorretas foi associada a sua versão correta, a sua forma sintática, bem como informações sobre a língua materna do estudante e outras línguas estrangeiras que o aluno também soubesse. Em seguida, anotámos as colocações incorretas considerando as três dimensões da taxonomia que iremos descrever na secção 5.5. (localização, descrição e explicação).

Para validar o processo de anotação, um segundo anotador também reviu as colocações corretas e incorretas selecionadas pelo anotador principal. Este processo de validação consistiu em várias reuniões em que o anotador principal e o segundo analisaram as colocações selecionadas pelo primeiro. Após uma discussão aprofundada, foi alcançado um acordo para cada uma das colocações. Após essa revisão, o tamanho do *corpus* diminuiu em 7%.

5.4.2. Taxonomia de erros de colocações

Depois de identificadas as colocações presentes no *corpus* e de cada uma delas ter sido classificada como correta ou incorreta, o passo seguinte que realizámos foi a categorização dos tipos de erros que afectavam as colocações incorretas.

A taxonomia de erros que usámos no processo de anotação e que aqui apresentamos foi inspirada na classificação de erros em colocações desenvolvida por Alonso Ramos. Esta escolha justifica-se uma vez que, de acordo com o nosso conhecimento, esta é a única taxonomia de erros desenvolvida exclusivamente para colocações. Muitas das taxonomias existentes tratam os erros em colocações como um mero erro lexical (Nesselhauf, 2005; Martelli, 2006; Granger, 2007). Por outro lado, outras optam por tratar os erros em unidades multipalavra, não fazendo uma diferenciação entre os vários tipos de estruturas que esse termo abarca (James, 1998).

Apesar desta taxonomia ser a que mais detalha os erros em colocações, optámos por realizar algumas alterações que nos pareceram mais apropriadas tendo em conta os objetivos do nosso estudo. Apesar das alterações, mantivemos os três níveis de análise que Corder (1974) e Ellis (1994) preconizaram e que Alonso Ramos também utiliza.

O **primeiro nível captura a localização do erro**, podendo este encontrar-se numa das partes da colocação ou afetá-la por inteiro. O **nível descritivo** tem como objetivo **detalhar com precisão o erro**. E, por último, o **nível da explicação** pretende encontrar uma possível **causa para o erro**.

Seguidamente, vamos descrever com mais pormenores cada um dos níveis de análise.

5.4.2.1. Localização

O nível da localização, tal como foi dito, pretende identificar a localização do erro dentro da estrutura da colocação. O erro pode encontrar-se em cada um dos

membros que compõem a colocação, ou seja, usando a terminologia de Hausmann, na **base** ou no **colocativo**. Esta é uma perspectiva que contrasta com estudos anteriores. Segundo Wanner *et al.* (2011: 2), alguns trabalhos que têm como objetivo a detecção automática de erros em colocações e apresentação de possíveis correções parecem implicitamente assumir que (a) a base se encontra sempre correta e que (b) a colocação com o significado que se pretende veicular existe sempre na língua alvo, portanto o aluno só pode fazer erros no uso do colocativo. Assumimos a posição crítica de Leo Wanner e, tal como ele, distinguimos tanto erros na base como no colocativo. Seguem exemplos, retirados do nosso *corpus*, destes dois tipos de erros:

Exemplo de erro: base

Erro: mandar recordas

Correção: mandar lembranças

Exemplo de erro: colocativo

Erro: tomar o sol

Correção: apanhar sol

Existem também alguns casos nos quais o colocativo e a base foram erradamente selecionados. Ao nível da localização, estes problemas foram anotados como contendo um erro na base e outro no colocativo.

Exemplo de erro: base e colocativo

Erro: fazem-nos regalos

Correção: dar uma prenda

Finalmente, existem ainda dois tipos de erro que afetam a colocação como um todo, e não a base ou o colocativo separadamente. Esta categoria inclui os casos em que se usa uma combinação que se assemelha estruturalmente a uma colocação, mas a forma correta para veicular um determinado significado seria uma unidade lexical.

Observámos também o fenómeno inverso, uma palavra é usada, mas, na verdade, o uso de uma colocação seria mais correto. Neste caso e considerando o

primeiro exemplo, a forma correta não resulta, pois, numa colocação (*lanchar* é uma unidade lexical, não uma colocação). No segundo exemplo, o mesmo não acontece, uma vez que a forma correta corresponde a uma colocação.

Exemplo de erro: colocação

Erro: fazer a merenda

Correção: lanchar

Exemplo de erro: colocação

Erro: o avião leva-se

Correção: o avião levanta voo

A seguinte figura, em baixo, resume os tipos de localização de erros que acabámos de descrever.

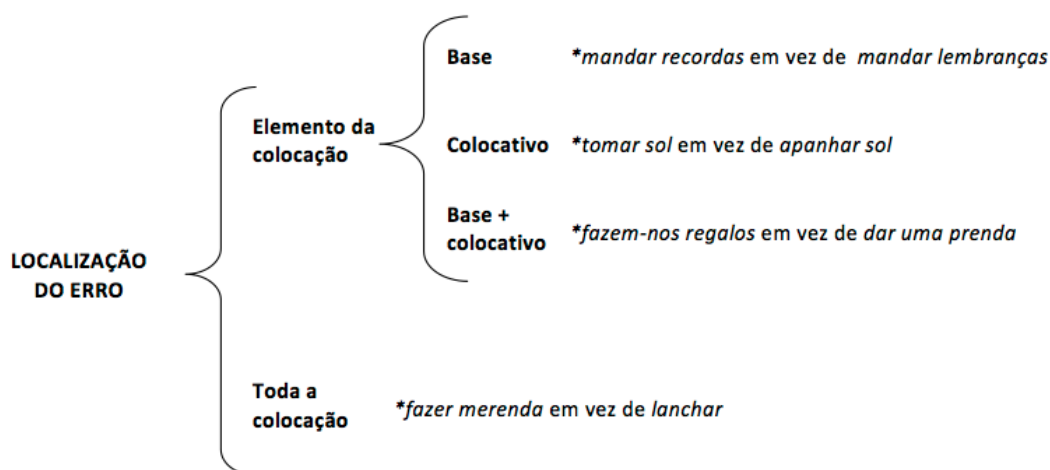


Figura 10 Resumo da categoria localização do erro.

5.4.2.2. Descrição

Ao nível da descrição, Alonso Ramos distingue três tipos de erros: os de registo, os gramaticais e os lexicais. Os erros de registo são usados para classificar usos incorretos de colocações corretas. Por exemplo, em espanhol, na frase *Yo tengo el deseo personal de ser bilingue*, a colocação *tener un deseo* está correta do ponto de vista lexical e gramatical. No entanto, observa-se um problema de registo, uma

vez que esta colocação deve ser usada num contexto formal ou mais emotivo, o que não é o caso na frase produzida pelo aluno. Os erros gramaticais são usados para classificar o uso incorreto das regras gramaticais de uma língua, no caso da autora, o Espanhol. Alguns exemplos são o uso incorreto do determinante (*tienen el derecho de* em vez de *tienen derecho de*) ou falta de concordância de género (*días festivas* em vez de *días festivos*) ou número (*tienen prejuicio* em vez de *tienen prejuicios*). Os erros lexicais dizem respeito a usos incorretos tanto da base como do colocativo, podendo estes erros ser, por exemplo, de criação de uma palavra (*derechos mujeriles* em vez de *derechos de las mujeres*) ou de substituição de uma palavra por outra que também existe na língua alvo (*gastar todo el año* em vez de *passar todo el año*).

O estudo de Alonso Ramos tem como objetivo a criação de uma plataforma *online* de ensino/aprendizagem de colocações que prevê a geração de uma ferramenta automática de correção de erros em colocações. Tendo em mente estas tarefas, será necessário corrigir todo o tipo de erros que possam afetar as colocações, podendo eles ser de registo, gramaticais ou lexicais. Ora, o nosso estudo tem como objetivo avaliar a competência colocacional dos alunos estrangeiros e, portanto, decidimos centrarmo-nos apenas nos erros lexicais. Erros gramaticais, como *días festivas*, em espanhol, podem ser solucionados através de um dicionário que contenha o género da palavra *día*. Ou seja, na nossa opinião, os erros gramaticais *per se* não afetam gravemente a produção colocacional e são também mais fáceis de solucionar, através da consulta de uma gramática ou de um dicionário, do que erros que envolvam a própria seleção lexical. Em relação aos erros de registo, no nosso *corpus*, não nos foi possível observar casos que recaíssem nesse nível de análise.

No nível lexical, tal como em Alonso, distinguimos entre erros de **Substituição, Criação, Síntese, Análise e Sentido diferente**.

A **substituição** refere-se à troca de uma palavra por outra que também existe em português. Esta troca pode referir-se à base:

Exemplo de erro: substituição (base)

Erro: alimentacion muito delicioso

Correção: comida deliciosa

Pode também ser uma troca do colocativo, o que, segundo alguns estudos (Vincze et al. 2011, Wanner *et al.* 2011, Martelli, 2006), é o erro mais frequente.

Exemplo de erro: substituição (colocativo)

Erro: jogar música

Correção: tocar música

Por último, os dois membros da colocações (base e colocativo) podem também sofrer uma substituição. Estes casos são dificilmente identificados e o contexto é muito importante para percebermos o que o aluno queria realmente dizer, mas é natural que alguns destes casos passem despercebidos.

Exemplo de erro: substituição (base e colocativo)

Erro: os teatros organizan numerosas actos

Correção: encenar uma peça

A **criação** refere-se ao uso de uma palavra como base ou colocativo, ou ambos, que não existe em português. A seleção errada da base, através de uma substituição ou criação, poderia simplesmente ser considerada um problema lexical e não um problema de seleção colocacional, uma vez que a base é seleccionada livremente e o colocativo é determinado pela base. No entanto, como a base faz parte da colocação, optámos por tratar os erros na base também como erros de colocação.

Exemplo de erro: criação (base)

Erro: carnação clara

Correção: tez clara

A criação pode também acontecer ao nível do colocativo:

Exemplo de erro: criação (colocativo)

Erro: comida à vontade

Correção: comida à descrição

Por último, a criação pode referir-se a ambos os constituintes da colocação:

Exemplo de erro: criação (base e colocativo)

Erro: ojos oscuros

Correção: olhos escuros

A **síntese** acontece quando o aluno utiliza uma palavra (unidade lexical) em vez de usar uma colocação.

Exemplo de erro: síntese

Erro: probar

Correção: pôr à prova

A **análise** trata o processo inverso, ou seja, o aluno cria uma expressão com uma estrutura semelhante à de uma colocação em vez de usar uma palavra. A correção destes casos não resulta numa colocação. Como foi já mencionado, na secção anterior, consideram-se que os erros de análise e síntese afetam a colocação como um todo.

Exemplo de erro: análise

Erro: tomar o almoço

Correção: almoçar

O **sentido diferente** classifica os casos em que o aprendente usa uma colocação correta em português, mas esta não é apropriada para aquele determinado contexto. Como podemos ver no exemplo em abaixo, o aluno escreveu *fazer uma fotografia* (fisicamente produzir uma fotografia) em vez de *tirar uma fotografia* (com uma máquina fotográfica). Ambos os sentidos existem, mas não são sinónimos. Na maioria dos casos, tanto a colocação correta como a errada partilham

a mesma base. No entanto, como se trata de duas colocações diferentes, o erro não é marcado apenas no colocativo, mas em toda a colocação.

Exemplo de erro: sentido diferente

Erro: fazer uma foto

Correção: tirar uma foto

Incluímos também na nossa taxonomia os erros de regência tratados no nível de análise gramatical por Alonso Ramos. No entanto, optámos por incluí-los no nível lexical. Consideramos que estes fazem parte da estrutura da regência lexical ou verbal da base ou do colocativo e, como tal, deve ser marcados num destes dois elementos. Ao nível dos problemas de regência, individuamos apenas o uso incorreto da preposição. Assim optámos por denominar este erro de **preposição**. O uso incorreto de uma preposição pode dever-se a uma substituição de uma preposição por outra, a uma inserção de uma preposição quando esta não faz parte da estrutura da colocação e, por último, à elisão de uma preposição que é obrigatória na estrutura da colocação.

Exemplo de erro: preposição (substituição)

Erro: voltar em realidade

Correção: voltar à realidade

Exemplo de erro: preposição (inserção)

Erro: jogar de futebol

Correção: jogar futebol

Exemplo de erro: preposição (elisão)

Erro: centro cidade

Correção: centro da cidade

Acrescentámos à nossa taxonomia um último tipo de erro que abarca estruturas que não estão realmente erradas, mas que poderiam assumir uma formulação mais precisa e mais utilizada pelos falantes do português – **opção melhor**. Alonso Ramos distingue os erros de registo que também podem ser tratados

nesta categoria. Estes casos podem surgir tanto ao nível da base, do colocativo ou afetar a colocação na sua totalidade.

Exemplo de erro: opção melhor (base)

Erro: tipos de música

Correção: géneros de música

Exemplo de erro: opção melhor (colocativo)

Erro: cozinhar o jantar

Correção: fazer o jantar

Exemplo de erro: opção melhor (colocação)

Erro: companheiro de apartamento

Correção: colega de casa

Finalmente, optámos por admitir a possibilidade de atribuir mais do que uma etiqueta a uma colocação. Por exemplo, no seguinte caso *ver* é substituído por *controlar* e no caso de *casela de posta electronica* houve uma criação, uma vez que esse nome composto não existe em português.

Exemplo de erro: substituição e criação

Erro: *controlar a casela de posta electronica*

Correção: ver o email

A figura, em baixo, resume os tipos de erros que acabamos de apresentar.

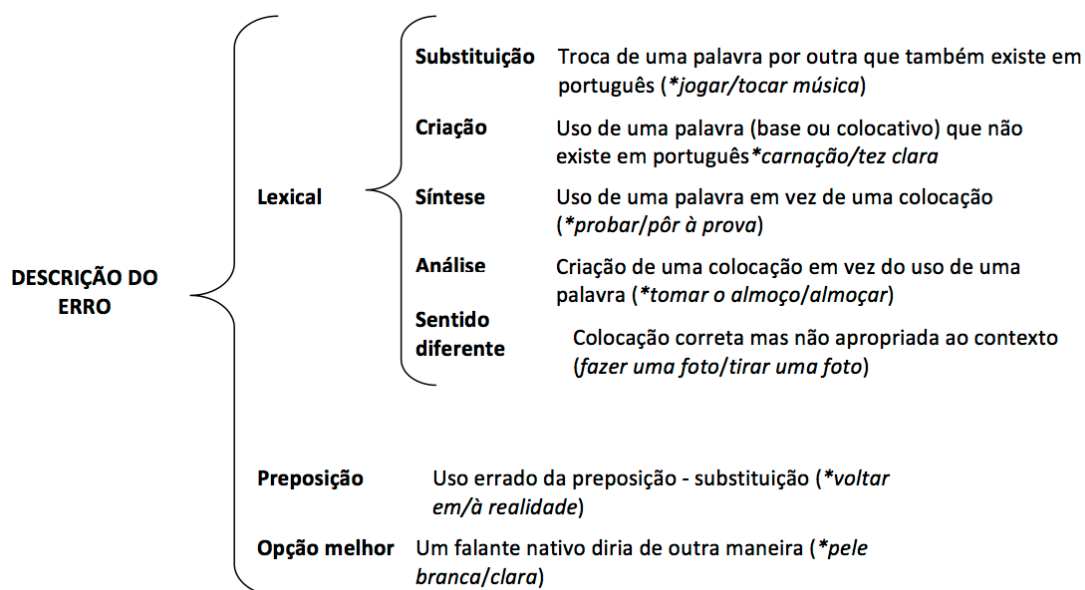


Figura 11 Resumo da categoria descrição do erro.

A seguinte tabela sintetiza a localização que cada um dos tipos de erros que descrevemos, nesta secção, pode assumir.

Descrição Localiz	Base	Colocativo	Colocação
Substituição	✓	✓	X
Criação	✓	✓	X
Síntese	X	X	✓
Análise	X	X	✓
Sentido diferente	✓	✓	✓
Preposição	✓	✓	X
Opção melhor	✓	✓	✓

Tabela 14 Localização dos tipos de erros: substituição, criação, síntese, análise, sentido diferente, preposição e opção melhor.

5.4.2.3. Explicação

Contrariamente a autores como Granger (2007: 467), que limitam a análise dos erros em colocações ao nível descritivo, optámos por seguir a perspectiva de Alonso Ramos que decide explicitar a fonte do erro, particularmente no nível lexical. Segundo a autora, interpretar os erros tem grande utilidade para a criação de material didático que tenha como alvo as dificuldades individuais de cada aluno.

Do ponto de vista da explicação dos erros, estes distinguem-se entre dois tipos: erros de “interlíngua” (transferência da LM do aprendente ou outra L2) e erros de “intralingua” (transferência da língua alvo da aprendizagem). Do ponto de vista lexical, os erros de interlíngua dividem-se em duas subclasses: **importação** e **extensão**.

Os erros de **importação** acontecem quando o aluno cria, na L2, uma unidade lexical, partindo de uma unidade lexical da LM ou de outra língua que o aluno possa saber. Esta importação pode ser direta (cópia direta da L2) ou não, fazendo-se alguma adaptação à língua que se está a aprender, neste caso o português.

Apesar de aqui explicitarmos estes dois tipos de importação, ambos os casos foram anotados no *corpus* apenas como importação. Neste tipo de erro, optámos por explicitar qual a língua da qual o aluno fez a importação.

Exemplo de erro: importação com adaptação (de italiano)

Erro: restos romanos (resti romani)

Correção: vestígios romanos

Exemplo de erro: importação direta (de italiano)

Erro: nolegjar um carro (noleggiare un auto)

Correção: alugar um carro

No caso de **extensão**, o aluno estende o significado de uma unidade lexical que realmente existe na L2. Em alguns casos de extensão, a unidade lexical que o aluno seleciona na L2 é uma tradução válida dessa mesma palavra na LM. No entanto, naquele contexto específico, não é a opção correta. No exemplo que se segue, **praticar teatro*, o aluno escolheu o colocativo *praticar* em vez de *fazer* (*fazer*

teatro). O verbo *praticar* normalmente usa-se com desportos com um sentido muito semelhante ao que o aluno pretendia exprimir, no entanto, neste contexto específico, não é o correto. Foi um aluno de língua materna alemã que produziu esta colocação e, uma vez que em alemão também se diz *fazer teatro* (Theater machen), neste caso não houve qualquer influência da língua materna.

Exemplo de erro: extensão

Erro: praticar teatro

Correção: fazer teatro

Para Alonso Ramos existem ainda mais dois tipos de extensão: **semelhança fonética** e **evitação da LM**. A semelhança fonética acontece quando uma palavra na L2 é usada porque é muito semelhante ao seu equivalente na LM. Em espanhol, neste exemplo, o aluno usa o colocativo *maternal*, na colocação **lengua maternal*, em vez do colocativo *materna* (*lengua materna*). A evitação acontece quando uma forma na L2 é evitada porque é muito semelhante ao equivalente na LM. Trata-se de um caso de hipercorreção. Neste exemplo, em espanhol, aluno descarta o uso do verbo *convertirse* na colocação *convertirse al cristianismo* e opta por usar o verbo *cambiar*, resultando na estrutura errada **cambiar al cristianismo*. Na nossa taxonomia, optámos por não distinguir estes dois subtipos, no entanto casos como os que foram descritos são considerados extensões.

Os erros de intralingua podem ser divididos em quatro subclasses: **derivação errada**, **sobregeneralização**, **escolha errada** e **variante**.

Acontece uma **derivação errada** quando o aluno produz uma forma inexistente na L2. Alguns destes casos resultam de uma analogia com uma forma na L2, noutros não existe uma lógica aparente.

Exemplo de erro: derivação errada

Erro: engravar a música

Correção: gravar uma música

No caso de **sobregeneralização**, o aluno usa uma unidade mais vaga ou genérica na L2 do que seria necessário.

Exemplo de erro: sobregeneralização

Erro: fazer arte

Correção: criar arte

Chamamos **escolha errada**, tal como em Alonso Ramos, aos casos em que o aluno seleciona uma palavra errada, mas onde não existe uma razão aparente para essa escolha, nem uma visível intervenção da LM ou de qualquer outra L2.

Exemplo de erro: escolha errada

Erro: olhar música

Correção: ouvir música

Por último, consideramos também na nossa taxonomia a categoria **variante**, sendo que este tipo de erro não é considerado por nenhuma das taxonomias previamente descritas (3.4.). Este erro acontece quando o aluno opta por utilizar uma colocação ou unidade lexical que é usada na variante da língua alvo. No nosso estudo, detetámos alguns usos da variante do Português do Brasil em vez da variante europeia. De certa maneira, este tipo de erro pode ser considerado um erro de interlíngua, uma vez que espelha uma transferência de uma variante da LM.

Exemplo de erro: variante do brasil

Erro: fazer regime

Correção: fazer dieta

A seguinte figura resume as explicações para os erros que incluímos na nossa taxonomia.

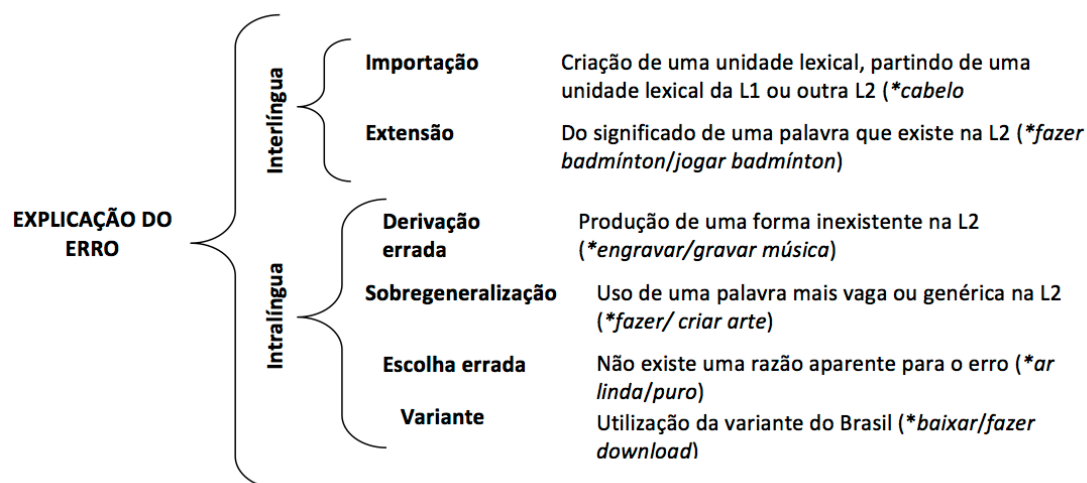


Figura 12 Resumo da categoria explicação do erro.

A seguinte tabela resume a relação entre os tipos de erros de acordo com a sua descrição e a sua explicação.

Descri Explic	Imp	Exten	Deriv	Sobreg	Escolha errada	Vari	Inter/Intra
Substituição	✓	✓	✓	X	✓	✓	X
Criação	✓	X	✓	X	✓	X	X
Síntese	✓	X	X	X	✓	X	X
Análise	✓	✓	X	X	X	✓	X
Sentido diferente	✓	✓	X	X	X	X	X
Preposição	X	X	X	X	X	X	✓
Opção melhor	✓	✓	X	✓	X	✓	X

Tabela 15 Relação entre a descrição dos erros e a sua explicação: importação, extensão, derivação, sobregeneralização, escolha errada, variedade e interlíngua e intralingua.

Ao nível da explicação, os erros que encontramos ao nível do uso da preposição, como podemos ver na tabela 15, foram apenas descritos como sendo de interlíngua, ou seja, erros que demonstram influência da língua materna ou de outra

L2, ou de intralíngua, isto é, erros que resultam da própria língua alvo, neste caso o português.

Exemplo de erro: interlíngua (influência do espanhol)

Erro: sair pela noite

Correção: sair à noite

Exemplo de erro: intralíngua

Erro: apanhar do avião

Correção: apanhar o avião

5.4.3. Desafios na anotação de colocações

Nesta última secção, descrevemos os principais desafios no processo de anotação das colocações. Primeiramente, estes problemas estão relacionados com a identificação da colocação como tal e com o julgamento de correção e incorreção que foi necessário fazer para cada uma das estruturas. Por último, para além destas duas questões, individuámos também problemas na interpretação e classificação do tipo de erros presente nas colocações.

5.4.3.1. Problemas na identificação de colocações

Como já foi discutido, as colocações são um fenómeno da língua que cria várias dificuldades, começando pela sua identificação. Distinguir e delimitar uma colocação de uma co-ocorrência livre e de uma combinatória completamente fixa não é uma tarefa fácil. Seguimos os critérios que tínhamos previamente estabelecido na secção anterior, mas, no entanto surgiram casos que nos levantaram dúvidas.

Existem exemplos nos quais é bastante claro identificar uma colocação, como no exemplo dado por Wanner *et al.* (2011:5) *boa nota*, uma vez que as características semânticas do nome *nota* pedem uma qualificação adjetival. Outros casos, como o nome *comida* podem criar mais problemas, uma vez que não requerem necessariamente um adjetivo. Na perspectiva do *continuum* entre combinatórias livres e completamente fixas, consideramos *comida processada* uma combinatória fixa, *comida quente* uma combinatória livre e *comida deliciosa* uma

colocação.

Apontando alguns casos mais concretos de dificuldades, tal como Wanner *et al.* (2011: 5), também nós, numa primeira análise, considerámos uma colocação a expressão *dar conta* (aperceber-se). Esta expressão não deveria ter sido considerada uma colocação, dada a sua natureza não composicional e a sua estrutura sintática completamente fixa. Esta interpretação errada deve-se ao facto do verbo *dar* ser muitas vezes utilizado em construções de verbo suporte, como *dar um passeio*, *dar um conselho* que são colocações.

Também é comum que colocações corretas passem despercebidas no processo de anotação, até a sua versão incorreta surgir. Por exemplo, *procurar casa* não tinha sido marcado como colocação até surgir a versão *buscar casa*. Estes casos confirmam a teoria de que a comparação entre línguas tem um papel fulcral para a identificação de colocações.

Ao mesmo tempo, pode também acontecer que outro tipo de erro, especialmente quando demonstra uma clara influência de outra língua, seja considerado como erro colocacional. A frase *Ele aprendeu-me a ler*, demonstra uma clara influência do Francês (*Il m'a appris à lire.*). No entanto não deve ser confundida como sendo uma colocação, apesar do erro.

A grande frequência pode também levar a julgamento errado na identificação das colocações, ou seja, expressões recorrentemente utilizadas pelos alunos, como *ir à praia*, podem também passar por colocações, apesar de não o serem.

5.4.3.2. Problemas no julgamento de correção

Como já foi reportado, nomeadamente em Vincze *et al.* (2011: 282), as maiores dificuldades em avaliar se uma colocação está correta ou incorreta estão relacionadas com a permissividade do anotador e com questões de variabilidade na língua. O facto de se tratar de textos escritos colocam também alguns limites às interpretações da intenção do aluno que podem também influenciar alguns critérios de correção.

Está também documentado que os anotadores tendem a ser menos permissivos com produções de não nativos do que de nativos e são mais rígidos em

relação a usos da linguagem mais criativos ou menos comuns. Para superar este problema e validar algumas utilizações menos frequentes usámos o motor de pesquisa *Google* e o site *linguee.pt* para assim confirmarmos algumas produções que não encontraríamos num *corpus* de referência de português.

A permissividade parece-nos também estar relacionada com a distância entre as línguas. Um aluno alemão que escreva *cabelo corto* em vez de *cabelo curto*, pode levar-nos a interpretar este erro como um simples erro ortográfico que não está relacionado com o conhecimento colocacional do aluno. No entanto, se um aluno italiano ou espanhol fizer semelhante erro, será interpretado como uma importação da língua materna, uma vez que *curto* se diz *corto* em italiano e em espanhol. Para tentarmos minimizar a questão da língua materna, as colocações corretas e incorretas foram revistas mais do que uma vez e questões duvidosas foram validadas por um falante nativo da língua em questão.

Erros de número e género não foram considerados erros. Por exemplo, a expressão **navegar no net*, demonstra apenas um problema no conhecimento no género da palavra *net*, problema este que não tem impacto no conhecimento colocacional do aluno e que pode ser resolvido através da consulta de um dicionário.

Chamamos também atenção para a questão das variedades do Português Europeu. Estes casos podem não ser considerados erros, dependendo do propósito da investigação. No nosso caso, optámos por considerá-los usos incorretos e atribuir-lhes a etiqueta “variação” ao nível da explicação do erro. Alguns exemplos são *comida gostosa* (comida saborosa), *fazer regime* (fazer dieta) ou *baixar* (fazer download). O primeiro mostra um erro no colocativo, o segundo na base e o último caso corresponde ao uso de uma unidade lexical que em Português Europeu é substituída por uma colocação. Surgiram também alguns casos de colocações que inicialmente pensámos que não ocorriam em Português Europeu, mas que na verdade foram atestadas pelos *corpora* de referência (*dar uma risada*).

5.4.3.3. Problemas na interpretação dos erros

Para além dos dois problemas previamente enumerados, existem também dificuldades na própria interpretação dos tipos de erros presentes nas colocações.

Neste trabalho, individuámos três principais tipos de dificuldade.

O primeiro problema está relacionado com a correção que selecionamos para cada erro. O tipo de erro que identificámos está intrinsecamente relacionado com a sua correção. Expressões que admitam mais do que uma correção, podem resultar problemáticas. Esta é uma questão que surge sobretudo nas construções de verbo suporte. A expressão *fazer uma fiesta* pode ser corrigida para a colocação *fazer uma festa* e, neste caso, descrevemos o erro como uso incorreto da base (*fiesta*), mas pode também ser corrigida pelo verbo *festejar* e, nesse caso, trata-se de um erro de análise. Por oposição às expressões que podem assumir mais do que uma correção, surgem também algumas combinatórias, como *fazer vistas* ou *cabelo nívio*, para as quais não conseguimos encontrar uma correção mesmo tendo em conta o contexto, a língua materna do aluno e outras línguas que eles também soubessem.

Em segundo lugar, algumas combinações incorretas produzidas pelos alunos que se assemelham a uma colocação do ponto de vista estrutural são normalmente traduções literais da língua materna que não têm um equivalente colocacional em português. Por exemplo a expressão *como muito legéro* (em italiano *mangiare leggero*) corresponde a uma colocação em italiano, mas só pode ser traduzida em português por uma frase complexa, uma vez que não existe uma colocação (lit. *comer leve*) para o exprimir o sentido de “comer alimentos de digestão leve”. Estes casos foram marcados como sendo importações e para os descrever usámos a etiqueta de análise, uma vez que se criou uma colocação na língua alvo, quando deveria ter sido usada uma combinação livre de palavras. Ocorre também o contrário; algumas expressões usadas pelos alunos não se constituem como colocações, e existe uma colocação em português mais apropriada para expressar o mesmo sentido. Por exemplo, a expressão *fritar um ovo* está correta, porém, em português, podemos também dizer *estrelar um ovo*. Como estes casos não são propriamente erros, optámos por utilizar a etiqueta que acrescentámos à taxonomia “opção melhor”.

Por último, como foi já mencionado, existem casos onde a fonte do erro não é completamente clara e podem existir interpretações ambíguas. Por exemplo, a colocação incorreta *tomar o autocarro* pode ser interpretada como uma tradução direta do inglês (*take a bus*). Porém poderia também ser interpretado como um erro

de extensão de significado, uma vez que tanto *tomar* como *apanhar* são ambas traduções possíveis do verbo *take* em inglês. No entanto, neste contexto em particular, apenas *apanhar* é usado com meios de transportes. Wanner *et al.* (2011: 5-6) também documenta esta questão e admite a atribuição de duas etiquetas ao mesmo erro. Nós, no entanto, seguimos a solução de Vincze (2015) e, nestes casos, damos preferência aos erros de interferência da língua materna.

Capítulo 6. Análise das colocações em PL2

Nesta secção, iremos analisar a produção das colocações por parte dos alunos de Português L2. Começamos, na secção 6.1, por analisar os valores gerais de ocorrência das colocações, a sua estrutura sintática, os nomes e os substantivos que as compõem e, por último, as colocações mais frequentes. Comparamos também estes parâmetros com a produção colocacional nativa. Seguidamente, na secção 6.2, centramo-nos nas colocações que revelaram problemas de correção. Assim, vamos olhar com detalhe para os tipos de problemas que apresentam. Na secção 6.3, propomos duas métricas de avaliação da competência colocacional: a precisão e a proliferação. Em 6.4, comparamos a produção de colocações nos aprendentes e nos não-nativos, usando as medidas previamente estabelecidas. Utilizando também estas métricas calculadas para cada um dos alunos, analisamos como estas se comportam em relação a variáveis individuais, tais como o número de anos de aprendizagem ou a formação académica de cada um dos aprendentes (secção 6.5). Por último, em 6.6, estabelecemos uma comparação entre os erros dos alunos e os erros colocacionais produzidos no contexto da tradução automática

6.1. Colocações no *corpus* de PL2

Nesta secção, vamos começar por analisar a totalidade das colocações presentes nos textos produzidos pelos alunos de PL2, observando os seus valores de ocorrências corretas e não corretas (secção 6.1.1). Nas secções 6.1.2, 6.1.3 e 6.1.4

apresentamos, respetivamente, as estruturas sintáticas mais usadas na construção das colocações, os nomes e substantivos mais comuns que as compõem e, por último, as colocações mais frequentes. Estes parâmetros serão também comparados com a produção nativa em cada um deles.

A análise apresentada é feita em traços gerais, tendo em consideração a totalidade dos dados. Nas secções 6.4 e 6.5 faremos, depois, uma análise mais detalhada, tendo em consideração a produção colocacional individual de cada aluno.

6.1.1. Descrição geral da produção de colocações

Começamos a análise das colocações encontradas nos textos dos alunos através de uma apreciação global das colocações produzidas. Atentemos na tabela geral que mostra o número de colocações (corretas e incorretas) encontradas por nível de PL2 e por língua materna, o número de alunos de cada nível, e o número de palavras e textos escritos.

LM	Nível	Alunos	Textos	Palavras ⁷⁸	Coloc. corretas	Coloc. incorretas	Total colocações
Alemão	1	59	64	8229	142	42	184
	2	48	73	20737	316	60	376
	3	1	1	523	9	0	9
Espanhol	1	68	147	18057	403	118	521
	2	55	90	19673	298	74	372
	3	5	7	1351	26	4	30
Francês	1	13	24	3002	53	18	71
	2	22	29	8101	119	12	131
	3	2	3	867	7	2	9
Inglês	1	24	29	4424	32	28	60
	2	35	57	14760	184	56	230
	3	7	10	2077	24	4	28
Italiano	1	87	168	24225	511	153	666
	2	60	77	18595	379	83	462
	3	5	11	2318	39	3	42
TOTAL		491	794	146455	2542	657	3189

Tabela 16 Número de alunos, palavras, textos, colocações corretas, incorretas e total de colocações por nível e LM.

⁷⁸ Como palavra entende-se unidade linguística dotada de sentido que pertence a uma (ou mais) categoria(s) sintática(s) e que é delimitada por espaços em branco.

Como podemos concluir através da análise da tabela 16, o número de colocações (em relação ao número de palavras) decresce nos dois últimos níveis em relação ao primeiro. Esta tendência contraria as conclusões de estudos, como o de Nesselhauf (2005) que conclui que alunos com um nível de proficiência mais alta numa língua estrangeira, normalmente, demonstram uma utilização de colocações mais elevada em comparação com alunos de um nível inicial. Nos nossos dados, porém observamos uma tendência contrária, com a exceção dos alunos de língua materna inglesa. Estes resultados podem ser explicados através do uso consciente de uma estratégia de evitação das colocações. Quando confrontados com uma falta de conhecimento, os alunos do nível superior, supostamente, têm mais conhecimento da língua para poderem substituir uma colocação por uma palavra ou expressão diferente.

Num estudo inicial de Howarth (2001 [1998], 180) é sugerido que quanto mais proficientes os alunos são, mais erros colocacionais dão, em comparação com alunos menos proficientes, pois têm mais confiança no uso da língua e correm mais riscos. Tal observação não é corroborada pelos nossos dados. Como podemos concluir pela observação do gráfico 1, ao longo dos três níveis, a proporção de colocações corretas aumenta em relação ao número de colocações erradas que diminui.

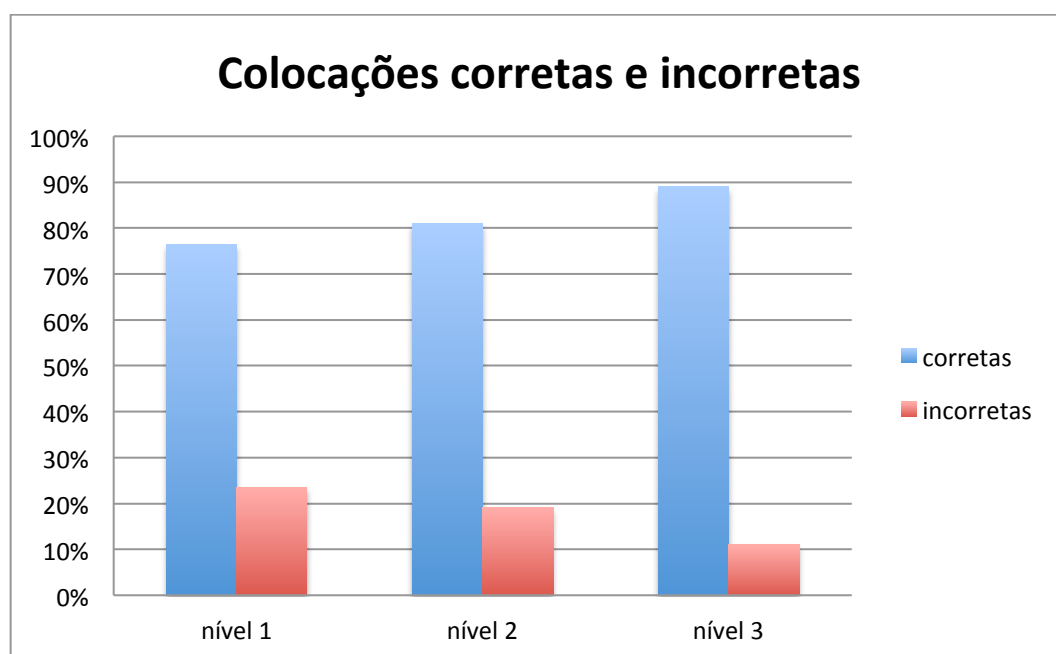


Gráfico 1 Percentagem de colocações corretas e incorretas por nível.

Com já foi mencionado, Nesselhauf (2005) advoga que existe um aumento quantitativo na produção de colocações nos níveis mais proficientes, no entanto, ainda segundo o autor, este aumento não implica um aumento qualitativo da produção de colocações. Também Howarth, num pequeno estudo inicial de 1993, conclui que os alunos mais proficientes dão mais erros em colocações do que os alunos menos proficientes, uma vez que teriam mais confiança para experimentar e arriscar nas combinações. No entanto, os nossos dados parecem não corroborar a hipótese destes autores. Considerando o número total de colocações, a proporção de colocações incorretas parece diminuir nos níveis mais altos. Como podemos observar no gráfico 2, apenas nos alunos de língua materna francesa, o número de colocações incorretas não diminui no terceiro nível e o número de colocações corretas não aumenta.

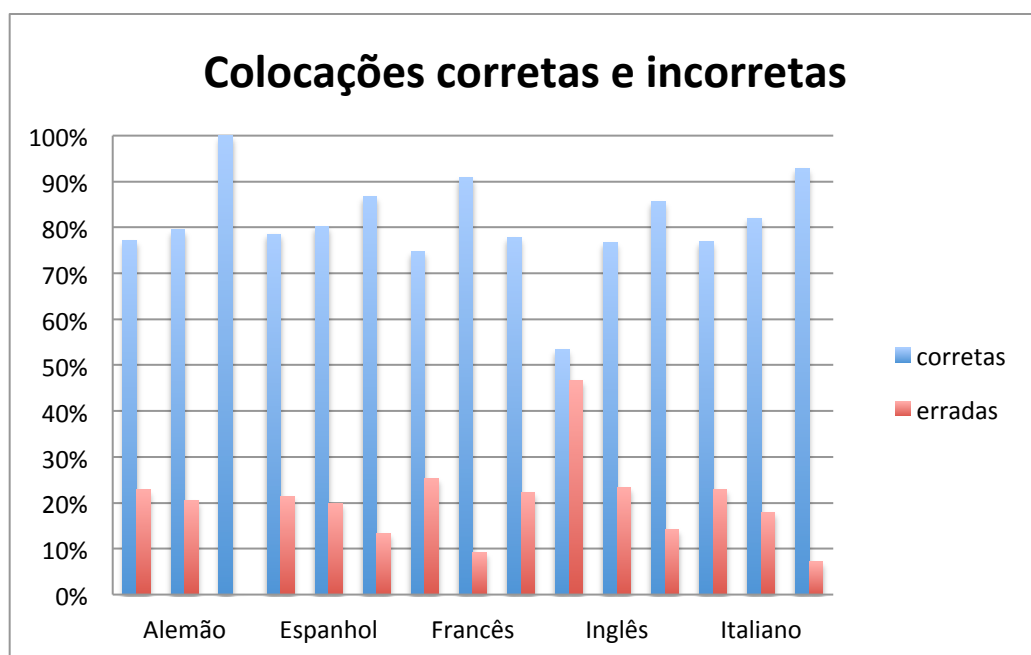


Gráfico 2 Percentagem de colocações corretas e incorretas por nível e LM dos aprendentes.

6.1.2. Padrões sintáticos

Prestemos agora atenção às estruturas sintáticas utilizadas pelos alunos na produção de colocações. Como podemos observar através do gráfico 3, a construção sintática V + N foi a mais utilizada pelos alunos na produção total de colocações, independentemente da sua língua materna ou do seu nível de proficiência. A

segunda estrutura mais utilizada é N + A ou a sua variante A + N. No entanto, entre os alunos italianos do nível 3 e os alunos do nível 1 e 2, surgem também algumas utilizações frequentes de V + Prep + N. Como podemos comprovar através da observação do gráfico 3, esta é a terceira combinatória mais utilizada na totalidade das colocações.



Gráfico 3 Padrões sintáticos das colocações utilizadas pelos alunos.

Também Martelli (2006), nos seus dados, encontra sobretudo utilizações destas duas estruturas sintáticas (V + N e N + A). Philip G. (2007) advoga que combinações que refletem a observação do mundo real, como V + N ou N + A, assumem uma ocorrência mais natural e portanto mais frequente no discurso dos alunos.

No entanto, como foi já dito por Nesselhauf (2005), a frequência em produções escritas não se reflete necessariamente numa utilização correta. Atentemos no gráfico 4, que representa a proporção das construções sintáticas das colocações incorretas.

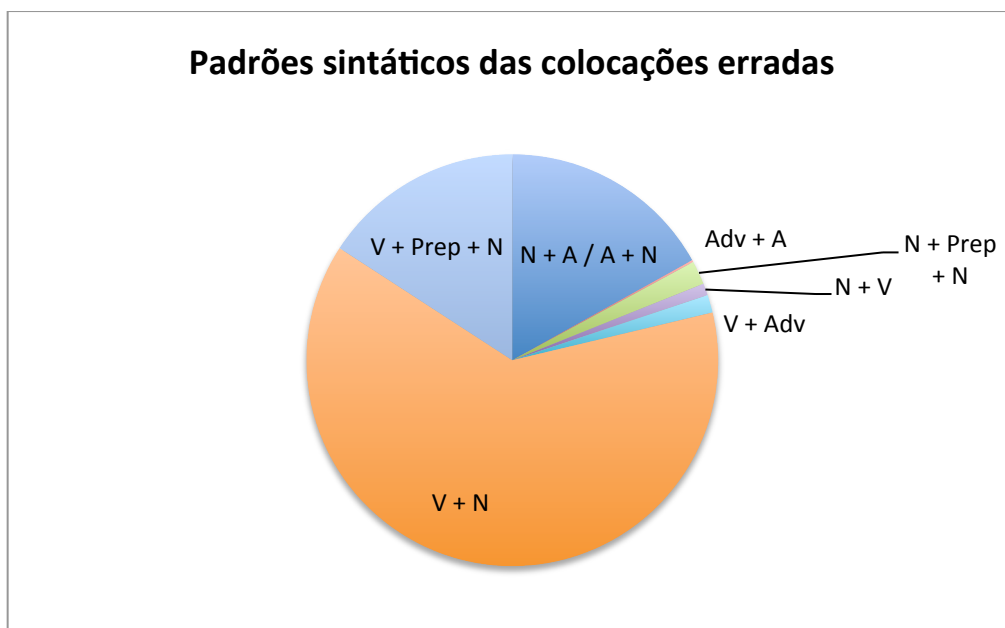


Gráfico 4 Padrões sintáticos das colocações erradas usadas pelos alunos.

Segundo Wanner *et al.* (2011), os alunos fazem erros em qualquer padrão sintático, no entanto, entre as colocações desviantes dos nossos dados, a forma sintática mais usada pelos alunos foi V + N. De fato, essa é a sequência mais estudada nas investigações com *corpora* de produções de aprendentes de uma língua estrangeira, pois os alunos apresentam dificuldades em selecionar o verbo correto não apenas dentro de uma colocação, mas também em sequências livres de V + N.

No estudo de Nesselhauf (2005) com estudantes alemães de inglês, um terço das combinações V + N analisadas não eram aceitáveis, principalmente devido a uma escolha errada do verbo, o que também está de acordo com o que observamos nos nossos dados.

Colocações que incluem adjetivos e advérbios parecem ser menos frequentes. Uma possível explicação encontra-se no facto de os alunos dominarem substantivos e verbos antes de chegarem ao domínio dos adjetivos e advérbios cuja presença parece aumentar em níveis de proficiência mais elevados (cf. Palapanidi *et al.*, 2014).

Segundo Philip G. (2007), muitos trabalhos de análise de produções de colocações apenas se centram nas colocações erradas, baseando os seus cálculos no número total de erros e não no número total de colocações (corretas e incorretas).

Tentando colmatar esta critica, vamos olhar, em seguida, para as colocações corretas. Como podemos comprovar através da observação do gráfico 5, tal como nas colocações erradas, o padrão sintático mais utilizado é V + N, sendo o segundo mais utilizado N + A/ A + N. Os alunos franceses e alemães usam quase exclusivamente a construção V + N.

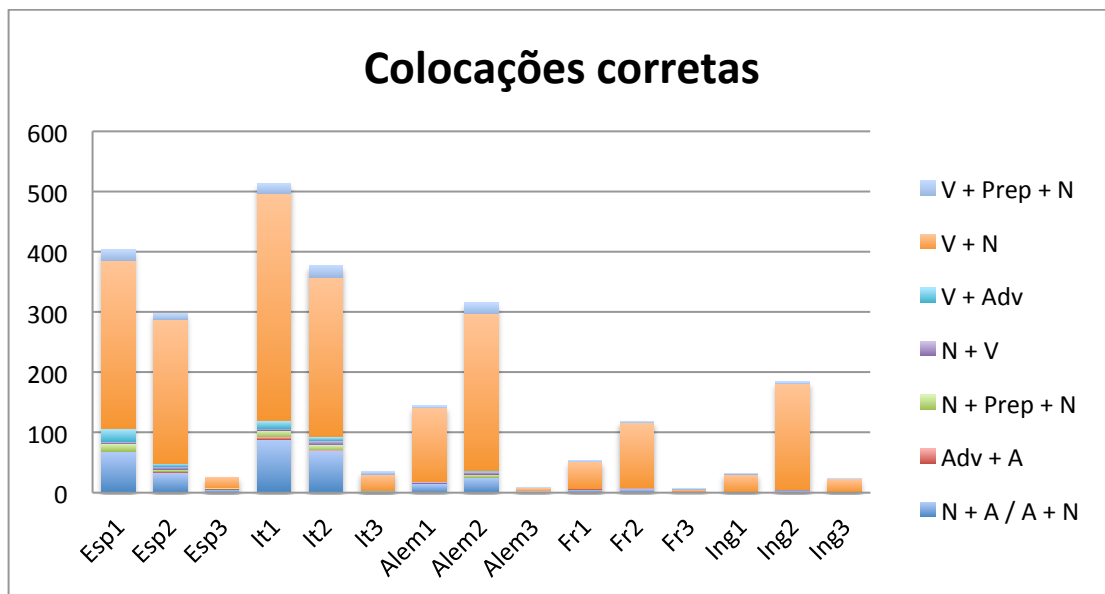


Gráfico 5 Padrões sintáticos das colocações corretas usadas pelos alunos.

Depois de analisarmos a produção não-nativa, vamos, no gráfico 6, observar os padrões sintáticos mais usados pelos nativos.

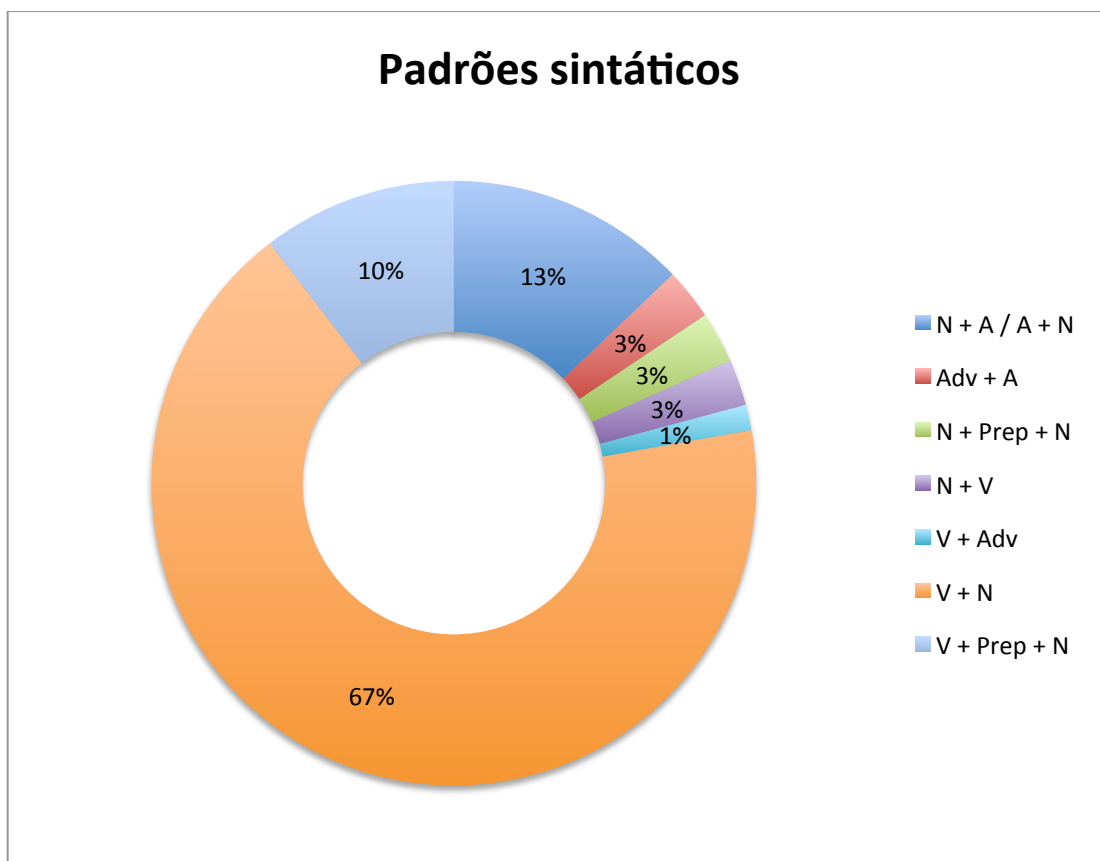


Gráfico 6 Padrões sintáticos das colocações usados pelos nativos.

Relembrando o gráfico 3 com o mesmo tipo de análise para os alunos estrangeiros, podemos ver que o padrão mais usado pelos nativos e não-nativos é V + N. Em seguida, os padrões mais usados são N + A / A + N e V + Prep + N, também em ambos os grupos. As combinações de N + Prep + N e N + V demonstram proporções de utilização semelhantes. A diferença mais significativa prende-se com as colocações compostas por Adv + A, sendo uma combinação bastante usada pelos nativos e a menos usada pelos não-nativos. A pouca utilização desta estrutura por parte dos alunos estrangeiros, pode dever-se ao facto de em combinações V + Adv, o advérbio normalmente não ser essencial para a correção da frase. Por exemplo, na frase *As minhas férias foram inteiramente dedicadas a caminhar*, o advérbio tem um valor enfático importante. No entanto, se o aluno não souber qual o advérbio mais apropriado, a frase faria sentido sem o mesmo.

Uma vez que as construções de V + N e N + A / A + N são as mais comuns nos dados recolhidos, na subsecção seguinte, observamos quais os verbos e os nomes mais utilizados pelos alunos e pelos nativos.

6.1.3. Verbos e substantivos

A tabela 17 mostra-nos os dez verbos mais usados na construção de colocações por parte dos alunos estrangeiros.

Entre os verbos mais frequentes, encontramos sobretudo verbos leves⁷⁹, como *dar, ter, fazer, tomar*. Como podemos verificar, o verbo *ter* é o mais frequente, apresentando aliás um número de ocorrências bastante alto em relação aos outros verbos. Algumas das combinações mais frequentes deste verbo são *ter saudades, ter tempo* ou *ter uma oportunidade*. Entre as utilizações incorretas encontramos, por exemplo *ter uma festa*, em vez de **fazer uma festa* ou **ter fotografia*, em vez de **tirar uma fotografia*. Sendo o primeiro uma tradução direta do inglês *have a party*, o segundo poderá tratar-se de uma confusão por parte de um aluno francês entre *tirar* e *ter*.

Verbo	Número de ocorrências em 146.455 palavras
Ter	832
Fazer	399
Passar	106
Dar	89
Apanhar	77
Jogar	71
Tomar	69
Ver	59
Sair	48
Ouvir	48
Total	1798

Tabela 17 Os dez verbos mais frequentes e correspondente número de ocorrências

⁷⁹ Segundo a Gramática do Português da Fundação Calouste Gulbenkian, trata-se de verbos que, em determinadas construções, se encontram desprovidos de parte do seu sentido descritivo básico, tornando-se incapazes de funcionar por si só como único predicado da frase (2013: 1214).

A lista de verbos mais frequentes que encontramos é muito semelhante ao elenco dos verbos mais utilizados pelos aprendentes de espanhol reportado por Orsolya (2015). Os três verbos mais frequentes na lista desta autora (*tener, hacer e passar*) são, na verdade, os mesmos que nós também encontramos nas produções dos aprendentes de português (*ter, fazer e passar*). Fazem ainda parte das duas listas verbos como, *tomar, dar e ver*.

Segundo esta autora e autores como Nesselhauf (2005), os aprendentes de uma língua estrangeira tendem a usar combinações que são constituídas por itens lexicais de alta frequência, como os verbos *supra* mencionados, constituindo isto uma estratégia para contornar a utilização de itens de vocabulário mais sofisticados ou estruturas gramaticais mais complexas. Prieto *et al.* (2009) ilustra estes casos, dando o exemplo dos colocativos que os nativos e não-nativos produzem com a base *viaje* (viagem). Segundo esta autora, os nativos usam colocativos como, *realizar, hacer, disfrutar, organizar* ou *planejar*, enquanto que os não nativos limitam-se à utilização do verbo *hacer*.

No nosso *corpus* de aprendentes observámos que *ver* é o colocativo mais utilizado com a base *filme*. No entanto, quando validamos esta combinação no *Corpus de Português Contemporâneo*⁸⁰, constatamos que *assistir a um filme* tem mais frequências na língua, representando esta a escolha mais frequente entre os nativos.

Fazendo uma análise aos verbos mais utilizados nos três níveis de proficiência, constatamos que existe alguma constância de utilização, nos três níveis, dos verbos *ter, fazer e dar*. No entanto, parece existir uma coincidência maior entre os verbos do primeiro e do segundo nível do que entre estes e os verbos do terceiro nível. Neste último nível estudado, surgem verbos mais complexos, como *conservar, considerar, transmitir* que não aparecem no nível 1 e 2. Em relação, às diferentes línguas maternas, podemos observar que verbos, como *ter, fazer, passar, tomar, jogar, dar e apanhar* são os dez verbos mais utilizados pelos aprendentes de todas as LMs.

Fizemos também uma análise dos verbos que são apenas utilizados por uma

⁸⁰ <http://www.corpusdoportugues.org/>

das cinco línguas maternas. Nas produções dos aprendentes de LM francesa e inglesa detetámos apenas cerca de cinco verbos que não são usados nas outras línguas. Já nos alunos de LM alemã, espanhola e italiana encontrámos uma maior variedade na escolha verbal. Apesar de não termos quantificado, notámos que os alemães tendem a usar verbos que representam estados ou sensações mais abstratas, como *proteger*, *apoiar*, *alcançar* ou *desempenhar*, enquanto que os alunos italianos e espanhóis escolhem verbos que possuem um significado mais concreto, sobretudo verbos descritivos, como *dormir*, *arrendar*, *almoçar*, *reservar* ou *crescer*.

Atentemos, em seguida, na tabela 18, que representa os dez nomes mais utilizados na produção de colocações por parte dos aprendentes de português. A maioria dos nomes utilizados são substantivos concretos bastante representativos das temáticas sobre as quais os alunos escreveram, nomeadamente, as férias, os tempos livres ou os seus *hobbies*. No entanto podemos também identificar alguns nomes abstratos como *saudades*, *música* ou *tempo*. Também Prieto *et al.* (2009) encontra nas produções dos aprendentes de língua espanhola, várias colocações com o nome *tiempo*, nomeadamente com os verbos *perder*, *tener*, *aprovechar*. Segundo o autor, esta variedade de colocativos utilizados demonstra que esta expressão parecer ser dominada pelos alunos.

Nomes	Número de ocorrências em 146.455 palavras
Tempo	210
Anos	162
Cabelo	121
Desporto	74
Saudades	71
Dinheiro	64
Música	59

Aulas	53
Férias	50
Noite	49
Total	913

Tabela 18 Os dez nomes mais frequentes e correspondente número de ocorrências

Em relação aos substantivos mais usados nos três níveis de ensino, podemos observar que não existe uma grande variedade na escolha. Substantivos, como *tempo* ou *saudades* ocorrem em todos os níveis. Também a produção de substantivos parece ser mais semelhante entre o nível 2 e 3 do que entre o nível 1 e 2, demonstrando uma evolução no conhecimento lexical maior nos últimos dois níveis. Em relação às diferentes línguas maternas, também não distinguimos grandes diferenças nos substantivos mais usados. Os substantivos são em regra geral concretos e estão bastante relacionados com as temáticas sobre as quais os alunos escreveram (*aulas, música, filme, amigo, curso*).

Analisando os substantivos que apenas surgem no primeiro nível, encontramos palavras bastante simples, como *avião, hotel, óculos, aniversário, prenda*. No segundo nível, o vocabulário muda um pouco e os substantivos usados são *notícias, vontade, acesso, desculpa, stress*. No terceiro nível, encontramos substantivos, como *direito, consideração, progresso, economia*. É claro que as temáticas sobre as quais os alunos escreveram são um pouco diferentes entre os níveis e isso influencia o vocabulário utilizado. Não obstante essas diferenças, parece-nos também que existe alguma evolução ao longo dos níveis, visível na escolha lexical.

Em relação às variações lexicais entre as várias línguas maternas, notamos, por exemplo, que os alunos de LM inglesa e alemã aproveitam a influência dos anglicismos e usam palavras como *laptop, ski, jogging* e *download*. No entanto, os alunos alemães demonstram uma maior variedade na produção lexical em relação aos alunos anglófonos. Os alunos franceses demonstram pouca variedade lexical, produzindo apenas cerca de dez substantivos que não são usados pelos alunos das outras LMs. No entanto, usam substantivos como *vindima (fazer a vindima)* ou *mimo*

(*dar mimo*) que não têm, em português, uma tradução muito semelhante ao francês. Os aprendentes de língua espanhola usam uma grande variedade de substantivos, tratando-se na sua maioria de substantivos concretos para designar nomes de coisas, como por exemplo *fogueira, velas, violino, montanha*. No entanto, encontramos também alguns substantivos abstratos, como *paz, ritmo, culpa*. Tal como os alunos espanhóis, também os alunos italianos apresentam uma grande variedade no uso de substantivos. Não querendo cair em estereótipos, são bastantes os substantivos relacionados com a culinária, nomeadamente *aperitivo, forno, massa, pimenta, refeição*. Em relação às outras línguas, não nos foi permitido encontrar uma temática tão recorrente.

Tal como na secção anterior, vamos também olhar para os verbos e substantivos mais usados nas colocações produzidas pelos nativos. A tabela 19, mostra-nos os verbos mais utilizados pelos nativos.

Verbos	Número de ocorrências em 11.667 palavras
Ter	51
Dar	26
Fazer	23
Passar	13
Receber	8
Pedir	8
Tomar	7
Tirar	7
Chegar	6
Responder	5

Tabela 19 Verbos mais usados na construção de colocações pelos nativos.

Como podemos ver na tabela, o verbo *ter* é o mais usado nas colocações dos nativos, como, por exemplo, em *ter uma oportunidade, ter sorte* ou *ter um*

problema. A alta frequência deste verbo é também registada nas produções dos aprendentes (tabela 17). Aliás, também os verbos *fazer, passar, dar* e *tomar* são tão frequentes nas produção dos nativos como nas produções dos alunos de PL2. Os restante verbos da tabela, apesar de não estarem entre os mais usados pelos alunos, registam também ocorrências. Existe uma grande coincidência dos verbos usados tanto por nativos como não-nativos, havendo apenas cerca de 10 verbos presentes nas colocações produzidas pelos nativos que não surgem nos textos dos aprendentes. Na verdade, os alunos estrangeiros demonstram uma grande variabilidade na escolha do colocativo em relação aos nativos. Por exemplo, o substantivo *tempo*, nas produções dos alunos, surge com os verbos *desperdiçar, passar, perde, poupar, ter, aproveitar, empregar*. Enquanto que nas produções dos nativos surge apenas com *aproveitar, perder, ter* e *passar*.

Atentemos em seguida nos substantivos mais frequentes na construção das colocações.

A tabela 20, em baixo, mostra-nos os substantivos mais frequentes nas colocações produzidas pelos nativos.

Substantivos	Número de ocorrências em 11.667 palavras
Curso	9
Plano	9
Aparência	7
Importância	6
Anos	6
Tempo	5
Desculpa	5
Notícias	4
Férias	4
Destino	4

Tabela 20 Substantivos mais usados na construção de colocações pelos nativos.

Relembrando a tabela equivalente para as produções dos alunos estrangeiros

(tabela 18), concluímos que, apesar das composições dos nativos e não-nativos terem tópicos muito semelhantes, existem poucos substantivos em comum. Entre os substantivos mais usados, encontramos apenas *tempo* e *anos*. Os nativos demonstram usar mais substantivos abstratos, como *aparência* ou *importância*, enquanto que os alunos tendem a usar substantivos mais concretos, como *cabelo* ou *dinheiro*. Assim, com o verbo *tirar*, o alunos produzem colocações, como *tirar um curso* ou *tirar uma foto*, enquanto que os nativos, para além destas, usam *tirar proveito* ou *tirar prazer*.

Para além desta diferença de uso de substantivos mais concretos e mais abstratos, parece-nos que os nativos demonstram uma maior variabilidade na seleção dos substantivos. Enquanto que os aprendentes usam a colocação *responder a uma pergunta*, os nativos produzem *responder a uma pergunta/ uma questão/ a um anúncio/ a um convite*.

6.1.4. Colocações mais frequentes

Depois de analisarmos alguns dos constituintes mais comuns das colocações, nesta secção vamos olhar para as colocações como um todo. A tabela 21 mostra-nos as colocações mais frequentes, tanto corretas como incorretas, presentes no *corpus*.

Colocação	Número de ocorrências em 146.455 palavras
Ter x anos	157
Ter saudades	69
Passar o tempo	56
Ter tempo livre	51
Ouvir música	48
Ter tempo	47
Fazer desporto	47
Ter aulas	42
Fazer compras	37
Ter dinheiro	32

Total	586
-------	-----

Tabela 21 As dez colocações mais frequentes no *corpus* dos aprendentes de português.

Como podemos concluir, as colocações mais frequentes são constituídas por V + N. Aliás, como tinha já sido dito, esta é a construção de eleição dos aprendentes de uma língua estrangeira. Tratam-se de colocações compostas por verbos e substantivos de alta frequência na língua e que fazem parte do vocabulário ensinado nas primeiras aulas dos níveis iniciais. Comparando, por exemplo *ter saudades* e *ouvir música*, que ocorrem nas produções dos alunos de todas as LMs, com o *Corpus de Português Contemporâneo*⁸¹, constatamos que *ter saudades*, é a colocação mais frequente para a base *saudades*, tal como *ouvir* é a segunda palavra mais combinada com a base *música*.

Sendo as temáticas das produções idênticas, também as colocações produzidas pelos alunos das cinco LMs são semelhantes assim, as colocações mais frequentes estão relacionadas, por exemplo, com *hobbies* (*ouvir música*, *ver um filme*), descrição física (*cabelo curto*) ou rotina diária (*tomar o pequeno-almoço*, *praticar desporto*).

Em relação às colocações utilizadas nos três níveis, podemos observar que as colocações do primeiro e do segundo nível são bastante semelhantes (*festejar o aniversário*, *cabelo louro*, *usar óculos*, *inscrever-se num curso*, *mudar de vida*). No terceiro nível, começa a surgir uma maior variedade do ponto de vista verbal, nominal e adjetival. Assim, observamos colocações como *ter consideração*, *terra árida*, *vasto leque*, *vento forte*, *vital importância*.

Observamos também que os colocativos usados para a mesma base mudam. Por exemplo, no segundo nível os alunos utilizam a colocação *ter atenção*, mas no terceiro nível surge o sinónimo, um pouco mais formal, *tomar atenção*. Também *fazer desporto* é mais usado no nível 1 no entanto, no nível 2, registamos várias ocorrências de *praticar desporto*. Na verdade, este último é o verbo que regista mais co-ocorrências com o nome *desporto* no *Corpus de Português Contemporâneo*⁸² e, portanto, será a opção mais semelhante à de um nativo.

⁸¹ <http://www.corpusdoportugues.org/>

⁸² <http://www.corpusdoportugues.org/>

Finalmente, vamos analisar as colocações mais usadas nos textos produzidos pelos nativos de português. A tabela 22 mostra-nos as colocações e as respetivas frequências.

Colocação	Número de ocorrências em 11.667 palavras
Fazer planos	9
Ter x anos	6
Dar importância	6
Pedir desculpa	5
Tirar um curso	4
Receber uma carta	4
Boa aparência	4
Ter férias	3
Passar tempo	3
Tomar o lugar	3

Tabela 22 Colocações mais usadas pelos nativos.

Relembrando novamente a tabela equivalente para os alunos estrangeiros (tabela 21), apenas as colocações *ter x anos* e *passar o tempo* se encontram, em ambos os *corpora*, entre as colocações mais frequentes. Apesar das colocações mais frequentes não serem as mesmas, existem também semelhanças na produção colocativa de nativos e não-nativos. Nomeadamente, as colocações mais frequentes entre os alunos de PL2 e os nativos são compostas por V + N. Aliás, como tínhamos já visto na secção 6.1.1.2, esta é a construção mais utilizada por ambos os grupos em estudo.

Sendo as temáticas dos textos dos alunos estrangeiros e dos nativos semelhantes, as colocações produzidas pertencem às mesmas áreas semânticas, como, por exemplo, os hobbies (*fazer desporto* (alunos), *tocar guitarra* (nativos)), descrição física (*cabelo loiro* (alunos), *cabelo ondulado* (nativos)) ou a rotina diária (*tomar banho* (alunos), *apanhar o comboio* (nativos)). No entanto, encontrámos

algumas colocações que embora semelhantes, demonstram que os nativos tendem a usar o colocativo de uso mais restrito em confronto com um de uso mais aberto. Por exemplo, os nativos produzem mais *assistir a aulas* e os aprendentes *ter aulas*. Os nativos também dizem mais *tirar um curso* enquanto que os alunos dizem mais *fazer um curso*.

Concluindo, a seleção dos colocativos, por parte dos alunos estrangeiros, é bastante coincidente com a dos nativos, sendo que, em torno de algumas bases, como *tempo*, demonstram mais variabilidade do que os próprios nativos. No entanto, demonstram uma menor variação na seleção dos substantivos e tendem a seleccionar nomes do tipo mais concreto e menos abstrato. Os nativos seleccionam o colocativo de uso mais restrito, enquanto que os aprendentes optam por usar colocativos com maior potencial de combinação, como por exemplo os verbos *fazer* ou *ter*. Sendo que este comportamento demonstra alguma prudência por parte dos alunos, que pode de facto evitar alguns erros.

6.2. Tipos de erros identificados nos corpora

Nesta secção, vamos apresentar os resultados da análise das colocações erradas, usando a tipologia descrita na secção 5.4.2. Na secção 6.2.1, começamos por analisar a localização dos erros. Em 6.2.2, fazemos a descrição dos erros e, finalmente, na secção 6.2.3, apresentamos uma possível explicação para os erros. Os valores apresentados são calculados em proporção, tendo em consideração a totalidade dos dados.

6.2.1. Localização dos erros

Vamos começar a análise dos problemas que encontrámos nas colocações pela localização do erro. A localização do erro pode variar entre a base, o colocativo (pode também ocorrer nos dois) ou afetar a colocação como um todo. Como podemos ver no gráfico 7, 67% da totalidade dos erros encontra-se no colocativo, 24% na base, 6% afecta a colocação como um todo e 3% dos erros afetam, numa mesma colocação, a base e o colocativo. Tal como nós, Nesselhauf (2005), Alonso

Ramos *et al.* (2010), Vincze *et al.* (2011) e Wanner *et al.* (2011) nos seus trabalhos, encontraram proporções muito semelhantes na distribuição dos erros, encontrando-se a maioria deles no colocativo. No entanto, Wanner *et al.* (2011) salientam que, contrariamente a outros estudos, os alunos podem dar erros não apenas no colocativo, mas também na base ou na totalidade da colocação.

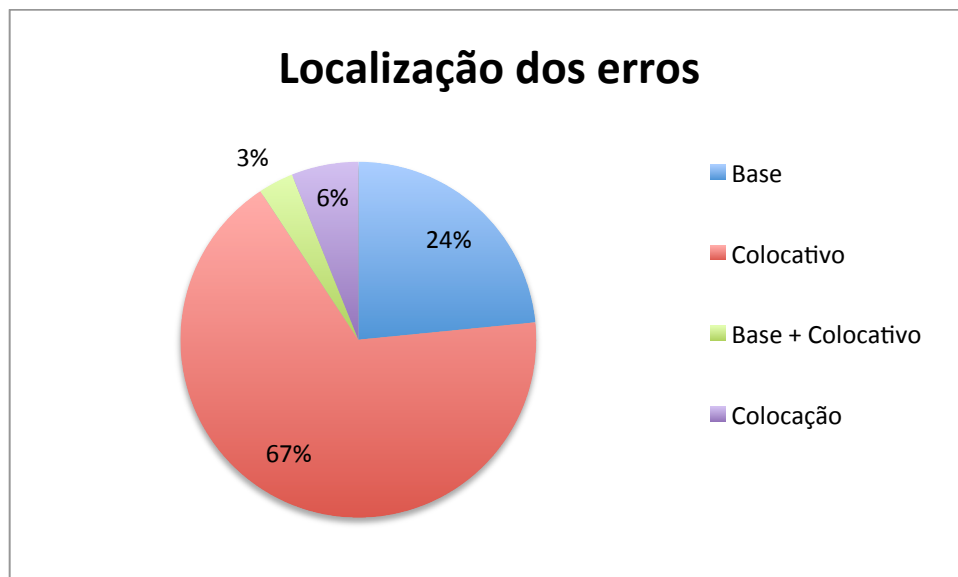


Gráfico 7 Localização dos erros na totalidade do *corpus*.

LM	nível	base	colocativo	base + colocativo	colocação
Espanhol	1	35	70	8	5
	2	17	50	0	7
	3	0	4	0	0
Inglês	1	5	19	13	1
	2	18	31	1	6
	3	1	3	0	0
Francês	1	3	12	1	2
	2	5	5	0	2
	3	0	2	0	0

Italiano	1	30	109	5	9
	2	18	60	2	3
	3	0	3	0	0
Alemão	1	13	26	1	2
	2	9	48	0	3
	3	0	0	0	0

Tabela 23 Localização dos erros nos três níveis e nas cinco línguas maternas.

Como podemos ver na tabela 23, e, mais especificamente, no gráfico 8, a distribuição dos erros que encontrámos na totalidade dos dados é também semelhante nos três níveis de proficiência. No entanto, no nível 3 não encontramos erros na base e colocativo ou na colocação como um todo. Centrando-se a totalidade dos erros somente na base ou no colocativo.

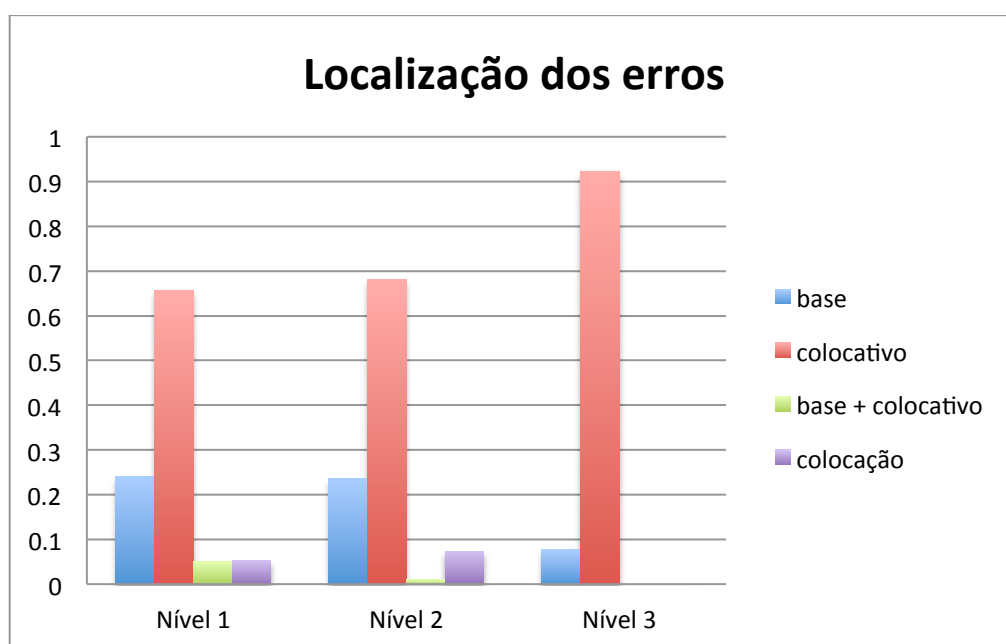


Gráfico 8 Localização dos erros, tendo em conta o número de colocações, nos três níveis de proficiência.

Tendo em consideração o número total de erros nas colocações produzidas, podemos ver, no gráfico 9, que os alunos de LM italiana e alemã são os que apresentam menos erros na base e mais erros no colocativo (**fazer uma volta* em vez de *dar uma volta* (LM alemã), **cabelo vermelho* em vez de *cabelo ruivo* (LM italiana)).

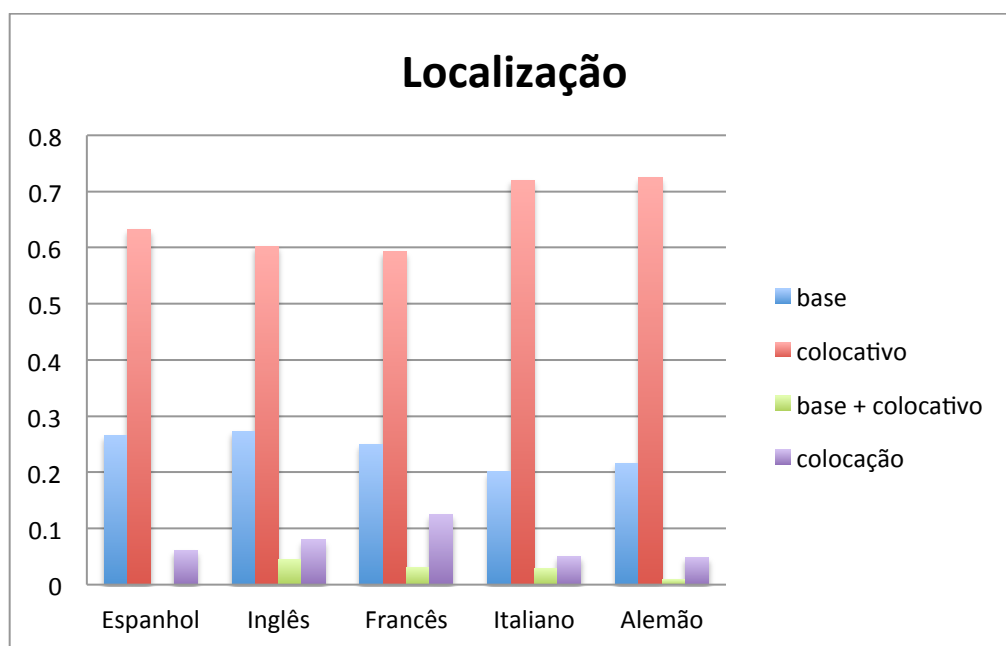


Gráfico 9 Localização dos erros, tendo em conta o número de colocações, nas cinco LMs.

Os alunos de LM inglesa e francesa, apesar de apresentarem mais erros no colocativo, demonstram um número bastante alto de erros na base (**alimentacion* muito *delicioso* em vez de *comida deliciosa* (LM inglesa), **ter fortuna* em vez de *ter sorte* (LM francesa)) (cf. gráfico 9).

Como foi já discutido, a maior dificuldade na produção de colocações está relacionada com a seleção do colocativo (cf. Hausmann, 1979). Neste sentido, seria possível considerar que os erros encontrados na base, como *mandar *recordas* em vez de *mandar lembranças*, são erros semelhantes aos que afetam as unidades lexicais e os verdadeiros erros colocacionais seriam apenas os erros no colocativo, como **tomar sol* em vez de *apanhar sol*, dada a sua restrição de seleção. No entanto, como a base faz parte da colocação, optámos por tratar os erros na base também como erros de colocação.

Os alunos de LM inglesa são os que mais erros apresentam na base e colocativo (**fazer autostop* em vez de *apanhar boleia*), tal como na colocação como um todo (**download* em vez de *fazer download*). Como podemos reparar através destes exemplos, os erros demonstram uma grande influência da língua materna, ou seja, são importações diretas de itens lexicais do inglês. Uma vez que o Inglês é uma língua que tem “emprestado” muito vocabulário a outras línguas (*fare l'autostop*

(it.)), incluindo ao português (*fazer download*), não é surpreendente que os alunos ingleses optem pela estratégia da importação direta perante uma dificuldade. Na secção 6.2.3, sobre a explicação dos erros, iremos novamente abordar a questão da influência da importação.

6.2.2. Descrição dos erros

Passemos em seguida à descrição dos erros encontrados, representados na sua totalidade no gráfico 10.

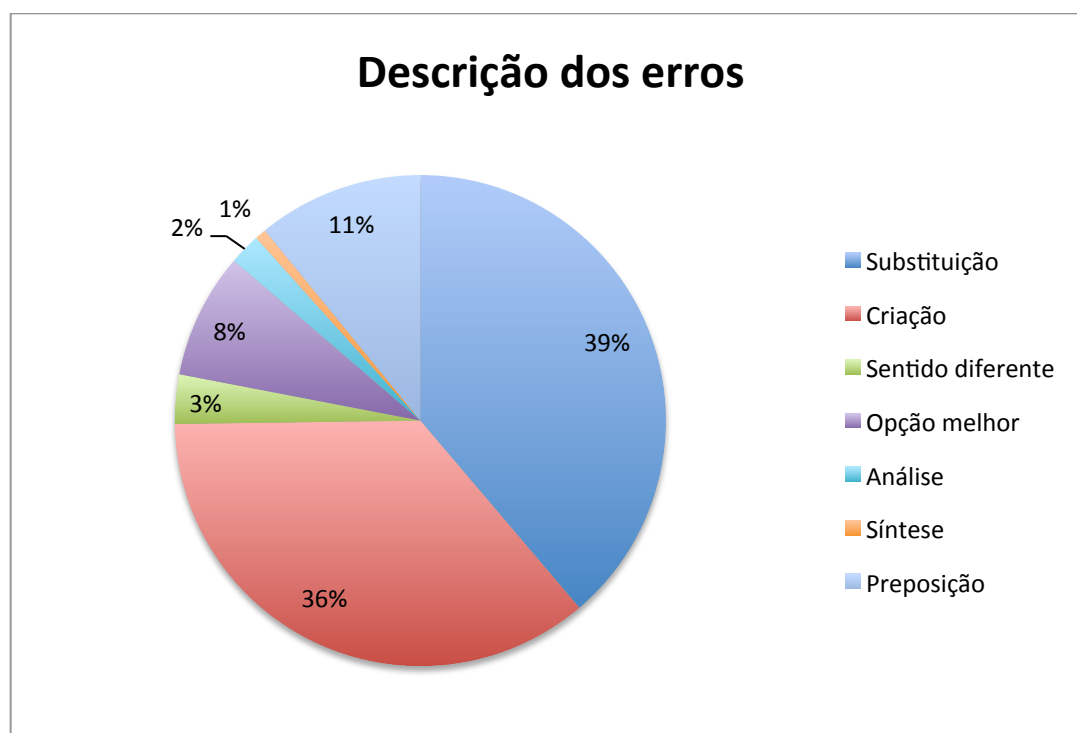


Gráfico 10 Descrição dos erros na totalidade do *corpus*.

Como podemos ver, 39% da totalidade dos erros são de substituição. Trata-se de casos em que os alunos substituem o elemento afetado da colocação por uma palavra inadequada mas que existe na LM ou na L2, é o caso de **jogar música* em vez de *tocar música*. *Jogar* existe na L2 e é uma tradução possível para *play* (inglês) ou *jouer* (francês). Nos casos de **música forte* em vez de *música alta* e **cabelo largo* em vez de *cabelo comprido*, temos dois exemplos de palavras (*forte* e *largo*) que, apesar de existirem na L2, neste caso foram seleccionadas pela influência da LM espanhola (*música fuerte* e *pelo largo*).

Criação é o segundo tipo de erro mais usado (36%). Neste caso, os alunos

inventam uma palavra que não existe na L2, sendo o mais provável tratar-se de uma importação direta de uma palavra da LM. No seguinte exemplo, **coger um táxi* (em vez de *apanhar um táxi*), a palavra *coger* foi criada, uma vez que não existe em português. No entanto existe em espanhol, podendo então trata-se de uma importação direta.

Os problemas nas preposições representam 11% dos erros. E um dos exemplos é **voltar em realidade* em vez de *voltar à realidade*, tratando-se de um problema de substituição da preposição *a* por *em*. As preposições são uma área crítica na aprendizagem do Português L2 e para a questão colocacional representam também uma dificuldade.

Podemos também observar que em 8% das ocorrências existiria uma opção melhor. Não se trata precisamente de erros, mas são casos em que existe uma expressão mais natural ou apropriada na língua. Por exemplo, **pele branca* em vez do adjetivo *clara* para descrever o tom da tez.

Existem ainda 3% de colocações que foram utilizadas com um sentido diferente. Por exemplo *fazer uma foto* tem o sentido de produzir uma foto, nomeadamente revelar uma foto. É um sentido aceitável na língua, no entanto, o aluno pretendia produzir o sentido *usar uma máquina para fotografar algo ou alguém* que é veiculado pela colocação *tirar uma foto*.

Por último, com 1% e 2% respetivamente, temos os casos de síntese e análise. No caso da síntese, uma unidade lexical é usada em vez de uma colocação, por exemplo **confinam* em vez de *fazer fronteira*. Neste caso, em particular, trata-se de uma importação do italiano, uma vez que nessa língua *confinare* significa *fazer fronteira*. No caso da análise, dá-se o processo inverso, o aluno cria uma expressão com a estrutura de uma colocação mas, na realidade, deveria ter usado uma unidade lexical. Atentemos no exemplo, **tomar o almoço* em vez de *almoçar*. Tal como no erro anterior, é também possível que a origem do erro se encontre numa transferência da LM ou numa generalização da L2, uma vez que em português se diz *tomar o pequeno-almoço*.

A tabela 24 mostra-nos os números totais por nível e por língua materna dos vários tipos de erros.

LM	Nível	Subs	Criação	Sentido diferente	Opção melhor	Análise	Síntese	Prep
Espanhol	1	46	54	2	7	4	0	13
	2	26	28	5	10	1	1	3
	3	1	1	0	2	0	0	0
Inglês	1	8	17	1	0	0	0	3
	2	20	19	1	4	4	1	8
	3	3	1	0	0	0	0	0
Francês	1	8	6	3	0	0	0	2
	2	1	8	2	1	0	0	0
	3	2	0	0	0	0	0	0
Italiano	1	50	59	2	20	5	2	20
	2	39	23	3	9	0	1	10
	3	3	0	0	0	0	0	0
Alemão	1	25	17	0	0	0	0	3
	2	31	11	3	3	0	0	12
	3	0	0	0	0	0	0	0

Tabela 24 Descrição dos erros nos três níveis e nas cinco línguas maternas.

Como podemos observar através da leitura da tabela anterior e do gráfico 11, no nível 1 encontramos sobretudo erros de criação. Face à falta de conhecimento, o aluno opta produzir itens lexicais que demonstram uma forte influência da LM ou de outra L2. Encontramos, no nível 1, por exemplo, produções como *cabelo *lauro* (*cabelo louro*) ou **tener dificuldade* (*ter dificuldade*).

No segundo nível, as substituições surgem com um valor ligeiramente mais alto do que as criações. Fazer uma substituição requer já um nível de conhecimento da língua mais avançado. Nesse sentido, é natural que este tipo de erro surja com valores superiores no segundo nível e ainda mais alto no terceiro. Alguns exemplos de erros, no segundo nível, são **tomar o sol* (*apanhar sol*), *esperar em *linha* (*esperar na fila*) e *momento *gracioso* (*momento mágico*) e, no nível 3, **atirar a atenção* (*chama a atenção*) e **tomar uma pausa* (*fazer uma pausa*). Estes exemplos, apesar de conterem um erro, são bastante perceptíveis e não causam grandes problemas de comunicação.

A utilização de uma colocação, mas com um sentido diferente do desejado (sentido diferente) aumenta também ligeiramente no segundo nível, desaparecendo totalmente no terceiro nível. Mais uma vez trata-se de uma estratégia um pouco mais sofisticada e que faz sentido ser mais usada no segundo nível do que no primeiro. No entanto, tratando-se de um erro terá que ser corrigido e aprendido

pelo aluno. Neste exemplo, *ter escolha* e *fazer uma escolha*, o aluno confunde o colocativo o que dá origem a dois significados bastante diferentes.

O erro opção melhor demonstra também um aumento ao longo dos três níveis. Alguns exemplos que encontramos são *pôr ao forno* (em vez de *levar ao forno*) (primeiro nível), *cabeça aberta* (em vez de *mente aberta*) (segundo nível) e *gastar o tempo livre* (em vez de *passar o tempo livre*) (terceiro nível). Como foi já dito na secção 5.4.2.2, estes casos não são necessariamente erros, mas antes usos que seriam mais adequados ou precisos, como nos casos de *mente aberta* e *levar ao forno*. No entanto, alguns podem causar problemas de compreensão, por exemplo *gastar tempo* tem uma conotação negativa que não está presente na colocação *passar tempo* e que poderá ser mal interpretada pelo interlocutor. É esperado, no entanto, que a adequação discursiva melhore ao longo da aprendizagem.



Gráfico 11 Descrição dos erros, tendo em conta o número de colocações, nos três níveis de proficiência.

No gráfico 12, podemos observar os tipos de erros divididos pelas línguas maternas. Os erros mais comuns em todas as línguas maternas são a substituição e a criação. No entanto, também os problemas na utilização das preposições parecem

ser transversais a todas as línguas.

Enquanto que os alunos de LM espanhola, inglesa e francesa usam mais a criação, os alunos italianos e sobretudo os alunos alemães optam por recorrer à substituição. Alguns exemplos de substituição, por parte dos alunos italianos, são **crescer uma família (constituir família)* ou **fazer prendas (dar prendas)*. Ambos os casos, tal como em muitos outros, trata-se de transferências do italiano *crescere una famiglia* e *fare regali*. Entre as produções dos alunos de LM alemã, encontramos **tomar o avião (das Flugzeug nehmen)*, **cabelos vermelhos (Rote Haare)*. Estes dois exemplos são traduções diretas do alemão, mas encontramos também exemplos que não são explicados por esse mesmo mecanismo, como **jogar desporto* (em vez de *fazer/praticar desporto*) em alemão *Sport treiben* (fazer desporto).

Como tínhamos dito, os alunos de LM espanhola, inglesa e francesa optam por criar mais palavras. Atentemos em alguns exemplos: **saiera o sol* (em vez de *nascer o sol*) produzido por um aluno de LM espanhola, **coger um táxi* (em vez de *apanhar um táxi*) presente na composição de um aluno de LM inglesa e, por último, **tenemos sortia* (em vez de *ter sorte*) foi escrito por um aluno de LM francesa. Como podemos concluir por estes exemplos, apesar de só um deles ser de um aluno de LM espanhola, os outros demonstram também uma grande influência dessa língua, uma vez que em espanhol se diz *salir el sol*, *coger un táxi* e *tener suerte*

Em relação aos erros nas preposições, são os alunos de LM alemã e inglesa que demonstram mais problemas e os alunos de LM francesa os que menos erros fazem. Entre as produções dos alunos de LM francesa, encontramos *sair de noite* e *aproveitar da vida*. Apesar de serem expressões corretas, no contexto em que foram usadas deveriam ser *sair à noite* e *aproveitar a vida*. Entre as produções dos alunos de LM alemã encontramos erros, como **fazer de surf (fazer surf)* e **assistir os jogos de futebol (assistir aos jogos de futebol)*. Os alunos de LM inglesa cometem erros, como **fazer de desporto (fazer desporto)* e **ter de saudades (ter saudades)*. Estas dificuldades podem sugerir que para alunos de LM inglesa e alemã será mais difícil aprender o sistema proposicional do português, enquanto que os alunos de LM francesa conseguem aproveitar maiores semelhanças entre as línguas.



Gráfico 12 Descrição dos erros, tendo em conta o número de colocações, nas cinco LMs.

6.2.3. Explicação dos erros

Nesta fase de análise, vamos prestar atenção à explicação dos erros, representados, na sua totalidade, no gráfico 13.

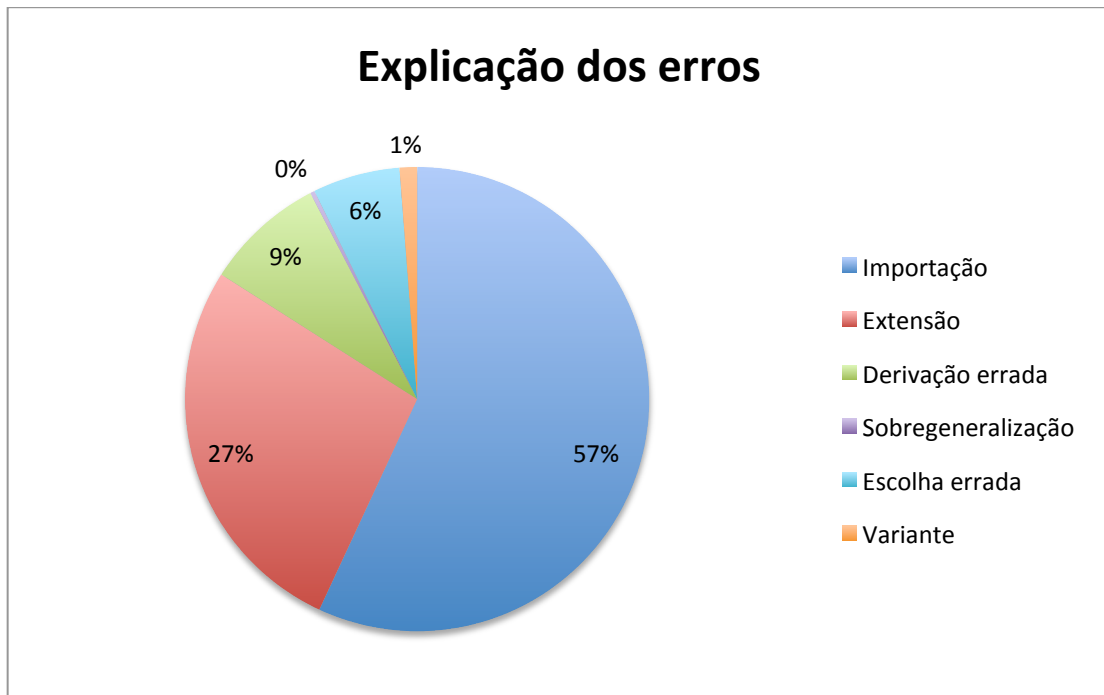


Gráfico 13 Explicação dos erros na totalidade do *corpus*.

Como nos mostra o gráfico, um pouco mais de metade dos erros devem-se a uma transferência da LM ou de outra L2. Autores como Vincze *et al.* (2011), na análise dos seus dados de aprendentes de espanhol, encontram também uma maioria de erros de transferência (67%). Trata-se de erros que demonstram uma clara influência de uma outra língua. Por exemplo, no caso de **cabelo corto* (em vez de *cabelo curto*) existe uma importação direta do espanhol na seleção do colocativo *corto*. No caso da colocação **curar o orto* (*tratar da horta*), o aluno de LM italiana importa o verbo italiano *curare* adaptando-o à grafia portuguesa dos verbos da primeira conjugação no infinitivo. Em outros casos, como **buscar uma casa* em vez de *procurar uma casa*, trata-se da utilização de um sinónimo que não é o mais correto no contexto da procura imobiliária.

A extensão é o segundo tipo de erro mais comum nas produções dos alunos (27%). Estes são os casos nos quais os alunos estendem o significado de uma palavra que existe em português. Em **faz chuva* em vez de *chover* ou **tomar o almoço* em vez de *almoçar*, os alunos criam expressões que assumem a configuração de uma colocação, usando os verbos *fazer* e *tomar*. Tratando-se de verbos que não têm um valor lexical muito marcado, tornam-se os “alvos” ideais para serem utilizados quando os alunos não têm a certeza qual o colocativo correto que a base deve seleccionar.

De seguida, 9% dos erros são de derivação errada. Nestes casos o aluno produz uma forma inexistente na L2, podendo esta resultar de um processo de derivação errado partindo de uma analogia com uma outra forma na L2 ou de uma adaptação ortográfica da língua alvo. No entanto, apesar de em alguns casos ser notável a semelhança a uma palavra na L2, nem sempre é possível identificar o processo que gerou a influência. Por exemplo, **engravar música* em vez de *gravar música* ou **cabelo caracolado* em vez de *cabelo encaracolado*.

Os problemas de escolha errada são 6% da totalidade dos erros. Nestes casos, um erro foi identificado, no entanto não conseguimos encontrar uma justificação óbvia nem uma clara intervenção da língua materna ou de outras L2. Por exemplo, **ar linda* em vez de *ar puro* ou *olhar música* em vez de *ouvir música*.

Representando 1% dos erros, encontramos os erros de variante. Este foi um tipo de erro que optámos por distinguir e que não existia em outras taxonomias,

nomeadamente na de Alonso Ramos *et al.* (2010) ou na de Philip (2007). No nosso caso específico, este tipo de erro dá unicamente conta dos erros de uso da variante do Português do Brasil. Alguns exemplos são, **baixar* em vez de *fazer download*, **tomar um vinho* em vez de *beber um (copo) de vinho* ou **fazer regime* em vez de *fazer dieta*. No primeiro caso, trata-se do uso de uma unidade lexical, usando-se, em Português Europeu, uma colocação. No segundo exemplo, existe uma variação no uso do colocativo (*tomar/beber*) e, no último caso, uma variação na base (*regime/dieta*).

Por último, não foi encontrado um número muito significativo de sobregeneralizações (apenas 2). Estes são os casos nos quais os alunos escolhem uma palavra mais vaga do que seria necessário, nomeadamente *fazer arte* em vez de *criar arte*. *Fazer* é um verbo mais vago do que *criar* e este último parece ser mais apropriado a um contexto artístico, no entanto não estamos perante propriamente um erro.

A tabela 25 mostra-nos os números totais por nível e por língua materna das várias explicações que encontrámos para os erros.

LM	Nível	Impor	Extensão	Derivação errada	Sobreg	Escolha errada	Variante
Espanhol	1	84	25	4	0	5	0
	2	42	18	9	0	3	2
	3	2	1	1	0	0	0
Inglês	1	23	2	1	0	2	0
	2	29	20	3	0	3	1
	3	2	2	0	0	0	0
Francês	1	9	6	3	0	0	0
	2	4	4	4	0	0	0
	3	0	2	0	0	0	0
Italiano	1	83	41	18	1	10	0
	2	48	24	2	1	3	3
	3	3	0	0	0	0	0
Alemão	1	22	7	6	0	5	2
	2	21	26	4	0	9	0
	3	0	0	0	0	0	0

Tabela 25 Explicação dos erros nos três níveis e nas cinco línguas maternas.

Observando a tabela anterior (25) e o gráfico 14, verificamos que a importação da LM ou de uma L2 é o recurso mais utilizado por todos os alunos,

independentemente do nível, embora tenda a baixar um pouco no nível 2 e 3 em relação ao primeiro nível. Thewissen (2008), Vincze *et al.* (2011), Wanner *et al.* (2011) e Alonso Ramos *et al.* (2010) encontraram, nos seus *corpora*, proporções semelhantes de importações (60% - 70%).

Segundo Biskup (1992), a transferência depende da LM e da distância que o aluno entende existir entre ela e a língua alvo. Assim, a não existência de uma tradução direta entre a L2 e a língua materna pode resultar num desempenho fraco por parte do aluno. Por exemplo, a colocação incorreta **jogar o saxofone* pode ser explicada pela falta de congruência entre *tocar saxofone*, em português, e *play the saxophone*, em inglês, ou *joue du saxophone*, em francês. Nesselhauf (2005), através da observação das produções de alunos de LM alemã a aprenderem inglês, conclui que 50% dos casos onde não existe congruência entre a colocação em alemão e inglês resultam em erro.

Olhando para o gráfico 14, notamos que o número de extensões parece aumentar ao longo dos três níveis de proficiência. Isto pode demonstrar que o aluno do primeiro nível está mais dependente do conhecimento que possui de outras línguas (LM ou L2), enquanto que alunos de níveis superiores começam a usar mecanismos mais sofisticados, como a extensão do significado de uma palavra. Como exemplos, encontramos a extensão do significado verbo deslexicalizado *fazer* em **fazer uma photo* em vez de *tirar uma foto* ou a extensão do significado do adjetivo *meio*, no caso de **fogo meio* em vez de *fogo médio*.

Por último, os erros de derivação errada apresentam uma mesma proporção constante nos três níveis. Muitos destes erros derivam de pequenas diferenças ortográficas, especialmente entre as línguas românicas. Por exemplo, **fazer sporto* em vez de *fazer desporto* ou **fazer viajens* em vez de *fazer viagens*.



Gráfico 14 Explicação dos erros, tendo em conta o número de colocações, nos três níveis de proficiência.

Atentemos, em seguida, no gráfico 15 que nos mostra a explicação dos erros por língua materna dos aprendentes.



Gráfico 15 Explicação dos erros, tendo em conta o número de colocações, nas cinco LMs.

Para os alunos de LM espanhola, a importação parece ser o mecanismo mais utilizado face a um conhecimento incompleto da L2. Este comportamento é detetável tanto no nível 1 como no nível 2. A escolha deste mecanismo pode ser justificada através das semelhanças que existem entre o português e o espanhol, podendo estas levar os alunos a acreditar que a importação tem uma alta probabilidade de funcionar a seu favor. De facto, existem muitos casos em que existe congruência na tradução das duas línguas, por exemplo *café fuerte* e *café forte* ou *hacer una pausa* e *fazer uma pausa*.

Os alunos de LM francesa, tal como os alunos das restantes LMs, usam sobretudo as importações, no entanto são também os alunos que apresentam maior número de extensões e de derivações erradas. Sendo exemplos da primeira, **cabelo moreno* em vez de *cabelo escuro* e **tempo de exames* em vez de *época de exames* e da última, **cabelo lauro* em vez de *cabelo louro* e **situação inconfortável* em vez de *situação desconfortável*. Não sendo o francês tão semelhante ao português, como o italiano ou o espanhol, é natural que os alunos optem por usar outros mecanismos, além da transferência.

Por último, foi entre as produções dos alunos de LM alemã que encontramos mais erros para os quais não conseguimos encontrar uma explicação (escolha errada). Sendo esta a língua etimologicamente mais distante do português, torna-se menos produtivo o uso de mecanismos como a importação ou a extensão. Assim os alunos podem usar outros raciocínios e generalizações cuja fonte não conseguimos identificar, como por exemplo **securitar o clima* em vez de *proteger o ambiente* ou *fazer vistas* em *Eu mia tempo livre gosto muito de fazer vistas!* (para a qual não encontramos uma correção).

Prestemos, agora, atenção à tabela 26, que nos mostra o número de importações organizados pelas línguas alvo de importação por língua materna e nível.

LM	Nível	Espanhol	Inglês	Francês	Italiano	Alemão
Espanhol	1	82	1	0	1	0
	2	36	1	1	4	0

	3	0	2	0	0	0
	1	15	5	3	0	0
Inglês	2	14	14	0	1	0
	3	1	1	0	0	0
	1	2	0	7	0	0
Francês	2	2	0	1	1	0
	3	0	0	0	0	0
	1	16	0	0	67	0
Italiano	2	11	0	0	39	0
	3	0	0	0	3	0
	1	9	1	1	0	11
Alemão	2	4	0	1	0	16
	3	0	0	0	0	0

Tabela 26 Erros de transferência nos três níveis e nas cinco línguas maternas.

De acordo com Leiria (2006), no que diz respeito às combinatórias de palavras, a maioria dos alunos utilizam a sua língua materna quando desconhecem a expressão correta. Através da análise dos nossos dados (cf. tabela 26), reiteramos a mesma conclusão da autora *supra* citada. Na verdade, a LM de cada um dos alunos é o principal apoio para a transferência, sendo os alunos de LM espanhola os que mais o fazem e os alunos de LM inglesa os únicos que recorrem mais a outra língua (espanhol) do que à própria língua materna. Para os alunos de LM francesa, italiana e alemã, o espanhol é a segunda língua à qual mais recorrem, depois da própria língua materna. Os alunos de LM espanhola também recorrem ao italiano e ao inglês. Somente os alunos de LM alemã recorrem ao alemão como apoio linguístico, sendo uma língua etimologicamente mais distante do português, nenhuma dos alunos das outras LMs a usa como apoio. Destes comportamentos podemos concluir que quanto mais próxima do português for a língua materna do aprendente, mais esta será usada como suporte e que os alunos estão cientes desta distância. Gabrys-Biskup (1992) também conclui, nos seus estudos com alunos polacos e alemães, que os aprendentes estão bastante inclinados para fazer transferências de LM quando desconhecem a expressão correta, mas, segundo a autora, esta está sempre dependente da distância entre a LM e a L2 ou, pelo menos, da concepção que os

alunos têm desta distância. Nas suas observações os alunos polacos recorrem mais à LM e dão menos erros do que os alunos alemães que recorrem a estratégias mais criativas que deram origem a mais erros. Uma vez que a língua alvo era o inglês, seria de esperar que os alunos alemães beneficiassem mais das semelhanças entre línguas, no entanto os alunos alemães não reconheceram esse potencial de transferência. Tanto através da análise dos nossos dados, como das observações por Gabrys-Biskup fica bastante claro que o recurso à transferência está dependente da distância ou proximidade que o aluno entende que existe entre a sua LM e a língua alvo.

6.3. Medidas de avaliação da competência colocacional

Ao longo desta tese, foi já amplamente discutido que a competência colocacional é de extrema importância para o domínio de uma língua estrangeira. Tratando-se de estruturas que representam uma área crítica da aprendizagem⁸³, acreditamos que para além de uma descrição dos erros, é também necessário estimarmos a competência colocacional dos alunos e percebermos de que é que esta depende. Para tal, nesta secção, propomos métricas de avaliação da competência colocacional.

O conceito de competência colocacional surge na literatura através de Lewis (2000: 49). Para este autor, uma colocação corresponde a uma co-ocorrência de palavras que os alunos não esperam encontrar juntas, o domínio destas idiossincrasias combinatórias chama-se competência colocacional (“collocational competence”). Um outro autor que também define competência colocacional é Sanromán Vilas. Em concordância com o autor citado, para Sanromán Vilas (2006: 1038), o termo competência colocacional “dá conta do domínio que os falantes têm da produção e compreensão de colocações”.

Em relação à sua avaliação, esta competência tem normalmente sido medida através da realização de testes com aprendentes de uma língua estrangeira. Estes

⁸³ Segundo Gabrys-Biskup a maior parte dos erros cometidos por alunos de L2, são erros de colocação (1992:87).

testes podem ser compostos por vários tipos de exercícios, nomeadamente frases na língua alvo que apenas contém a base a negrito e é necessário preencher com o colocativo (Sanróman Vilas, 2006), frases que contém um espaço em branco para colocar a base e o colocativo, apresentando a primeira letra de cada um deles (Fernandez e Schmitt, 2015). Para além destes tipos de testes de resposta fechada, existem trabalhos que usam técnicas de elicitación. Granger (1998) utiliza um teste de escolha múltipla em alunos franceses aprendentes de inglês. O teste consistia em dar aos alunos uma lista de advérbios amplificadores. Os alunos tinham que escolher, de entre os adjetivos também disponibilizados, quais se combinavam com cada advérbio. Outros investigadores optam por realizar exercícios de tradução, nomeadamente de colocações de V + N e de A + N, podendo esta tradução ser feita apenas da colocação, sem contexto, ou de uma frase onde esta se encontra (Biskup, 1992; Bahns e Eldaw, 1993).

No caso da nossa investigação não foram realizados testes fechados, mas levou-se a cabo uma análise das colocações que os alunos produziram num contexto livre de escrita⁸⁴, o que, no nosso entender, reflete mais fidedignamente a competência colocacional adquirida pelo aluno. Como não realizámos um teste, não tínhamos um valor que pudesse ser considerado o máximo de competência colocacional a atribuir a um aluno, no entanto definimos duas métricas para medir alguns aspetos desta competência.

6.3.1. Proliferação

A primeira métrica que propomos tem como objetivo medir a proliferação de colocações nos textos que os alunos produziram, ou seja, quantas colocações o aluno produz, independentemente de estas estarem corretas ou incorretas.

Para obter esse resultado multiplicamos o número total de colocações produzidas pelo aluno por dois, uma vez que prototipicamente as colocações são compostas por dois elementos, e dividimos este resultado pelo número total de palavras escritas pela aluno. A fórmula, em baixo, representa o cálculo que propomos.

⁸⁴ Apesar de esta ser orientada através de um tópico.

$$Proliferação = \frac{(total\ colocações \times 2)}{total\ de\ palavras}$$

6.3.2. Precisão

A segunda medida que propomos é a precisão, que mede a exatidão com a qual os alunos produzem as colocações, ou seja, da quantidade de colocações que os alunos tentam produzir, quantas destas estão realmente corretas. Assim, dividimos o número total de colocações corretas pelo número total de colocações produzidas.

$$Precisão = \frac{total\ de\ colocações\ corretas}{total\ de\ colocações}$$

O valor de precisão atribuído a cada aluno varia numa escala de 0 a 1, sendo o valor 1 o máximo que o aluno pode obter.

6.4. Competência colocacional nativa e não-nativa

Na presente secção, vamos fazer uma comparação entre a produção de colocações no *corpus* dos aprendentes de português e a produção dos nativos que compõe o grupo de controlo, tendo em consideração as métricas previamente definidas, ou seja, a proliferação e a precisão. Para obtermos a proliferação e precisão de colocações usadas por um nativo num texto semelhante aos que os aprendentes produziram, anotámos igualmente as colocações presentes nos textos produzidos pelos nativos do grupo de controlo descrito na secção 4.1.5. Como foi já reportado, este grupo de nativos produziu 51 textos, com um total de 11.667 palavras. Tal como os textos produzidos pelos não-nativos, identificámos as

colocações produzidas nestes textos e encontramos 3.197.

Os valores que reportamos nesta secção e na secção 6.5 são calculados tendo em consideração a proliferação e a precisão específica de cada aluno e de cada nativo. Relembro que até aqui, os valores apresentados foram computados tendo em consideração a totalidade dos alunos.

6.4.1. Proliferação nativa e não-nativa

A tabela 27 mostra-nos o resultado da média e respectivo desvio padrão da proliferação dos alunos não-nativos e dos nativos. Como esperado, o valor médio da proliferação nos nativos é mais elevado em comparação com os não-nativos.

Proliferação		
	Não-nativos	Nativos
Média	0.044 ± 0.028	0.068 ± 0.034

Tabela 27 Proliferação dos não-nativos e nativos.

O gráfico 16 mostra dois diagramas de caixa, construídos a partir dos resultados da medida de proliferação dos nativos e não-nativos.

Dado que a média e mesmo o desvio padrão podem não ser muito informativos em relação à distribuição dos dados, optou-se por apresentar os diagramas de caixa associados aos dados. Estes têm a seguinte leitura:

- Os traços horizontais dos extremos mostram os valores máximos e mínimos obtidos em termos de proliferação dos nativos e não-nativos (por exemplo, os nativos obtiveram a proliferação máxima de 0.15 e os não-nativos uma mínima de 0).
- As bolas indicam os valores atípicos (*outliers*).
- A linha grossa dentro de cada caixa representa a mediana (isto é, 50% dos dados estarão acima da linha e os outros 50% abaixo).
- A linha superior de cada caixa indica a separação entre o quartil 4 e o

quartil 3 (ou seja, 25% dos dados estão acima dessa linha e 25% dos dados estão entre essa linha e a mediana). Do mesmo modo, a linha inferior indica a separação entre o quartil 1 e o quartil 2 (ou seja, 25% dos dados estão abaixo dessa linha e 25% dos dados estão entre essa linha e a mediana).

Posto isto, o gráfico 16 reforça a ideia de que a produção de colocações entre os nativos é mais alta do que nos aprendentes, o que se encontra na linha da hipótese de Granger (1998) que refere que os aprendentes usam menos construções pré-fabricadas do que os nativos, ou seja, usam mais o *open-choice principle*, por oposição ao *idiom principle* (Sinclair, 1991: 109). Howarth (1998) também comprova que os aprendentes evitam usar colocações e encontra menos colocações nas produções de não-nativos do que de nativos.

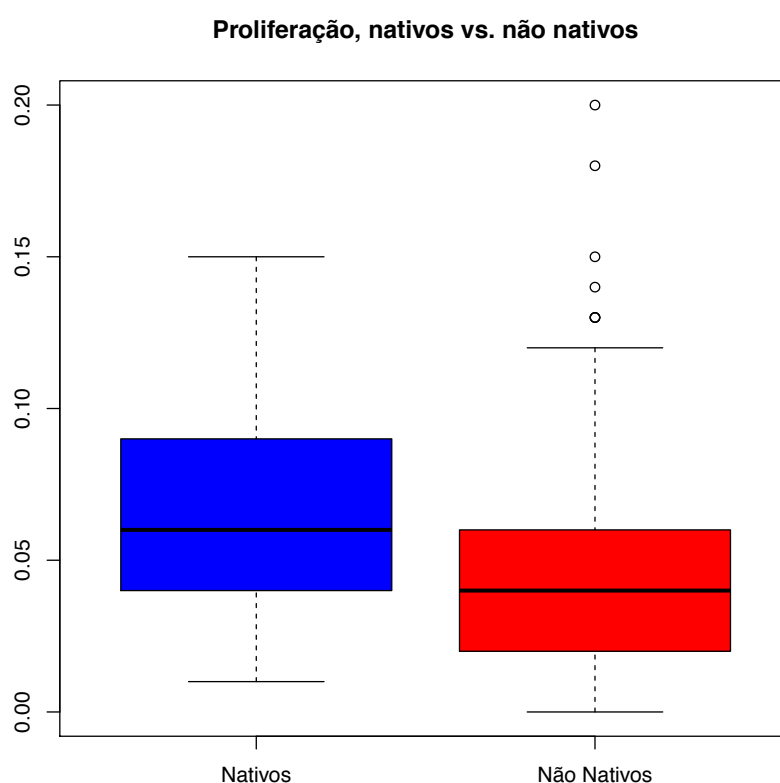


Gráfico 16 Proliferação dos nativos e não-nativos.

6.4.2. Precisão nativa e não-nativa

Como podemos ver na tabela 28, a média da precisão dos não-nativos é mais baixa do que a dos nativos. No caso dos nativos esta medida equivale a um, uma vez que as produções dos nativos foram consideradas todas corretas. Dado que a precisão dos nativos é um, não apresentaremos, nesta secção, o correspondente gráfico de caixa.

Precisão		
	Não-nativos	Nativos
Média	0.727±0.311	1.0±0.0

Tabela 28 Precisão dos não-nativos e nativos.

6.5. Impacto das variáveis individuais

Tendo definido, na secção 6.3, duas métricas para a avaliação da competência colocacional dos alunos, nesta secção, vamos contrastá-las com as informações que possuímos sobre os alunos e assim perceber quais delas estarão relacionadas com o domínio das colocações. Relembro que os valores são apresentados tendo em consideração a proliferação e a precisão calculadas para cada aluno. A tabela em baixa sumariza os dados que os vários *corpora* disponibilizam sobre os alunos e que iremos utilizar na nossa análise.

Informação	PEAPL2	CLUL	FCSH	Ciberescola
Nível	✓	✓	✓	✓
LM	✓	✓	✓	✓
Anos de estudo da L2	✓	✓	✓	✓
Imersão	✓	✓	✓	✓
Uso da L2	✓	✓	x	✓
Nº de L2	✓	✓	✓	✓

Idade	✓	✓	✓	✓
Formação acadêmica	x	✓	✓	✓

Tabela 29 Informação disponível nos quatro *corpora* sobre os alunos.

Nos seguintes pontos, iremos usar cada uma das informações que possuímos para percebermos o seu impacto na proliferação e na precisão.

6.5.1. Nível

Nesta secção analisamos a influência do nível dos alunos na proliferação e na precisão. Apresentamos, seguidamente, a média e o desvio padrão destas medidas e mostramos os respetivos diagramas de caixa.

	Proliferação			Precisão		
	Nível 1	Nível 2	Nível 3	Nível 1	Nível 2	Nível 3
Média	0.050± 0.032	0.039± 0.022	0.031 ±0.015	0.690± 0.329	0.781± 0.260	0.831± 0.248

Tabela 30 Média e desvio padrão da proliferação e precisão dos não-nativos e nativos.

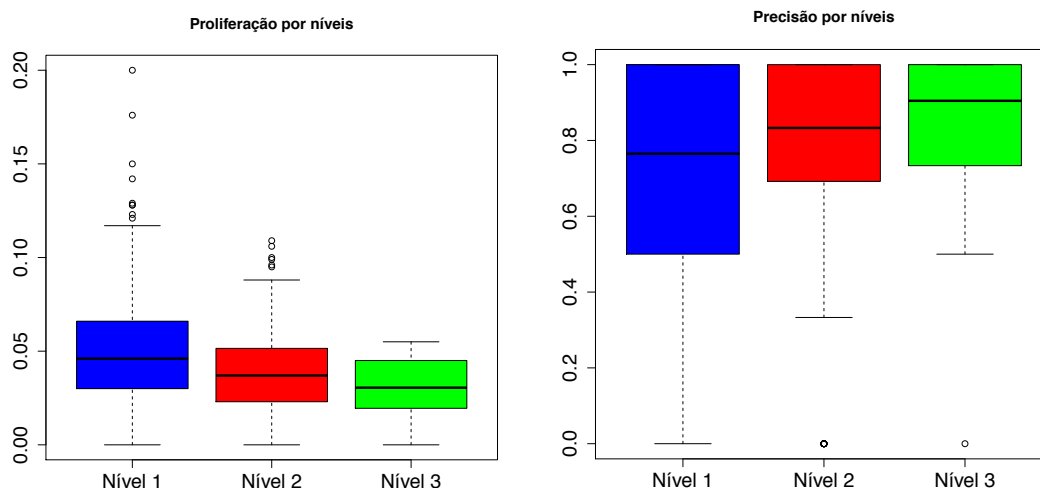


Gráfico 17 Proliferação e precisão tendo em consideração os três níveis de proficiência.

Apesar de na literatura não ser conhecida uma clara relação entre a produção de colocações e o nível de proficiência (Howarth, 1998), através do cálculo da média e da proliferação dos três níveis de aprendizagem (tabela 30 e gráfico 17), podemos

notar que existe uma diminuição da proliferação ao longo dos níveis, enquanto que a precisão parece aumentar.

Estes dados parecem confirmar a hipótese de Prieto *et al.* (2009) de que os aprendentes de níveis mais elevados, tendem a evitar o uso de colocações. Possuindo um conhecimento mais profundo da língua alvo, podem mais facilmente evitar estas estruturas e optar por uma reformulação com a qual se sintam mais confiantes.

Por sua vez, a precisão, parece aumentar, mostrando que na totalidade de colocações produzidas, o número de colocações corretas aumenta. Estes dois resultados podem demonstrar que os alunos produzem menos colocações, no entanto optam, cautelosamente, por utilizar as colocações que já dominam.

6.5.2. Língua Materna

Nesta secção analisamos a influência da língua materna dos alunos na proliferação e na precisão. Apresentamos, seguidamente, a média e o desvio padrão destas medidas e mostramos os respetivos diagramas de caixa.

	Proliferação					Precisão				
	Esp	It	Fr	Ing	Alem	Esp	It	Fr	Ing	Alem
Média	0.047±	0.050±	0.037±	0.030±	0.042±	0.755±	0.724±	0.811±	0.608±	0.785±
	0.030	0.029	0.020	0.022	0.026	0.271	0.306	0.242	0.379	0.266

Tabela 31 Média e desvio padrão da proliferação e precisão para as cinco LMs.

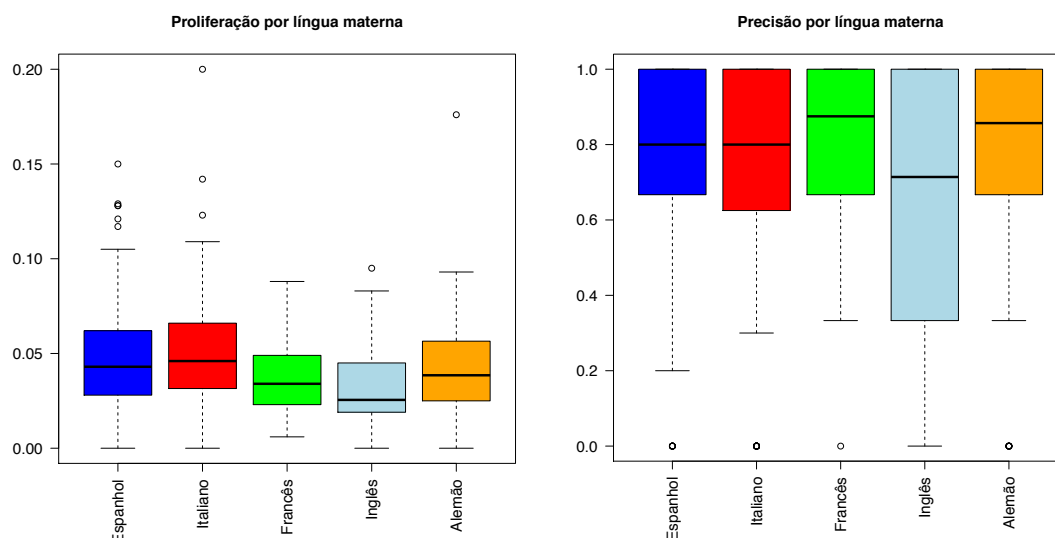


Gráfico 18 Proliferação e precisão tendo em consideração as cinco LMs dos alunos.

Atentando no desempenho dos alunos das várias línguas maternas, no gráfico 18, podemos observar que a proliferação e a proficiência não ocorrem nas mesmas proporções. Os alunos de LM italiana e espanhola são os alunos que apresentam uma maior proliferação de colocações nas suas produções. Já a maior precisão é reportada pelos alunos de LM francesa e alemã.

Apesar da língua italiana e espanhola serem mais próximas do português em relação ao alemão, são os alunos alemães que produzem mais colocações corretas. Este comportamento poderá dever-se ao facto de muitos alunos alemães dominarem o espanhol, uma língua etimologicamente mais próxima do português do que o alemão e que poderá servir como língua *pivot* de apoio.

6.5.3. Anos de estudo da L2

Nesta secção analisamos a influência dos anos de estudo da língua portuguesa por parte dos alunos na proliferação e na precisão. Apresentamos, seguidamente, a média e o desvio padrão destas medidas e mostramos os respetivos diagramas de caixa.

	Proliferação				Precisão			
	1 ano	2 anos	3 anos	+ 4 anos	1 ano	2 anos	3 anos	+ 4 anos
Média	0.040±	0.044±	0.036±	0.037±	0.708±	0.732±	0.777±	0.813±
	0.025	0.028	0.021	0.023	0.341	0.280	0.301	0.258

Tabela 32 Média e desvio padrão da proliferação e precisão para os anos de aprendizagem.

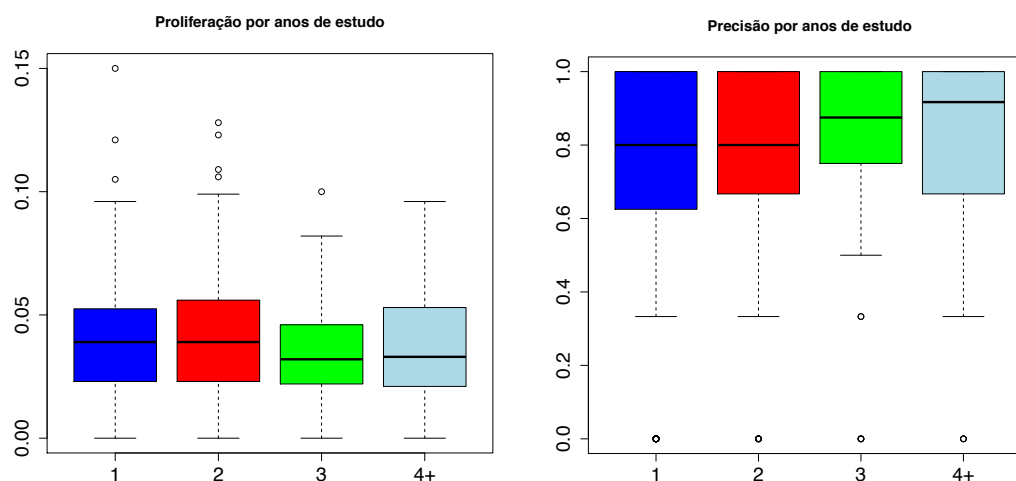


Gráfico 19 Proliferação e precisão tendo em consideração os anos de estudo de português dos alunos.

Como podemos ver, através do gráfico 19, recolhemos dados de alunos que estudavam português há um ano, até a alunos que aprendiam português há 20 anos. Optámos por deixar de fora os dados dos alunos que estavam a começar a aprender e que ainda tinham pouca experiência prévia com a língua, uma vez que o nosso objetivo era avaliar o progresso ao longo dos anos de ensino.

Investigações sobre o uso de colocações em aprendentes de L2 tem demonstrado que independentemente do número de anos de estudo, até mesmo os alunos de níveis mais avançados apresentam dificuldades na produção de colocações (cf. Nesselhauf, 2003) e tendem a usar combinatórias livres em vez destas. Nesselhauf (2005) é mesmo da opinião de que o número de anos de estudo formal não tem impacto na produção de colocações.

No entanto, alguns investigadores, como Gitsaki (1999) reportam um aumento do conhecimento colocacional ao longo do processo de aprendizagem. De facto, nos nossos dados, a precisão aumenta e demonstra o seu valor mais alto no

grupo de alunos que tem mais anos de experiência com a língua. Já a proliferação não demonstra um padrão claro.

6.5.4. Imersão

Nesta secção analisamos a influência da imersão ou não imersão no país da língua alvo, neste caso Portugal, na proliferação e na precisão. Apresentamos, seguidamente, a média e o desvio padrão destas medidas e mostramos os respetivos diagramas de caixa.

	Proliferação		Precisão	
	Imersão	Sem imersão	Imersão	Sem imersão
Média	0.045±0.026	0.043±0.030	0.778±0.244	0.686±0.351

Tabela 33 Média e desvio padrão da proliferação e precisão para a imersão ou não.

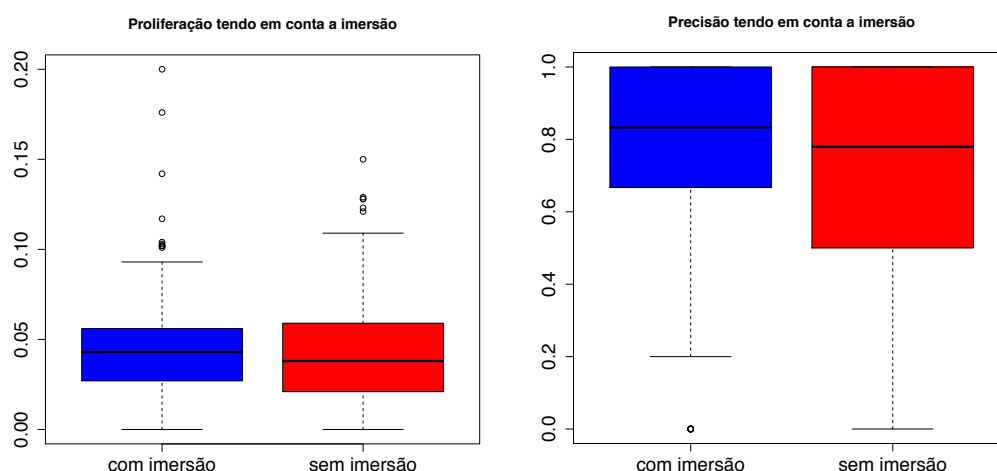


Gráfico 20 Proliferação e precisão tendo em consideração os dados de imersão dos alunos.

Através do gráfico 20 podemos ver que os valores da proliferação e da precisão aumentam ligeiramente nos alunos que estão aprender português em contexto de imersão. Fernández e Schmitt (2015), no seu trabalho com alunos espanhóis aprendentes de inglês, encontram uma correlação entre a imersão e o aumento do conhecimento colocacional. De facto é amplamente sabido que a imersão no ambiente da L2 tende a melhorar o domínio da língua e facilita o processo de aprendizagem (cf. Pawley e Syder, 1983). Nesselhauf (2005) tem uma

opinião ainda mais radical. Segundo o autor a exposição no país parece ter mais impacto na produção de colocação do que anos de estudo formal. Segundo o mesmo, isto pode levar a concluir que o seu ensino não está a ser conduzido da melhor forma, enquanto que a mera exposição parece ajudar mais.

6.5.5. Uso da língua alvo

Nesta secção analisamos a influência do uso ou não uso da língua alvo fora do contexto escolar por parte dos alunos na proliferação e na precisão. Apresentamos, seguidamente, a média e o desvio padrão destas medidas e mostramos os respetivos diagramas de caixa.

	Proliferação		Precisão	
	Uso	Não uso	Uso	Não uso
Média	0.043±0.023	0.046±0.033	0.738±0.243	0.702±0.335

Tabela 34 Média e desvio padrão da proliferação e precisão para o uso ou não uso da L2.

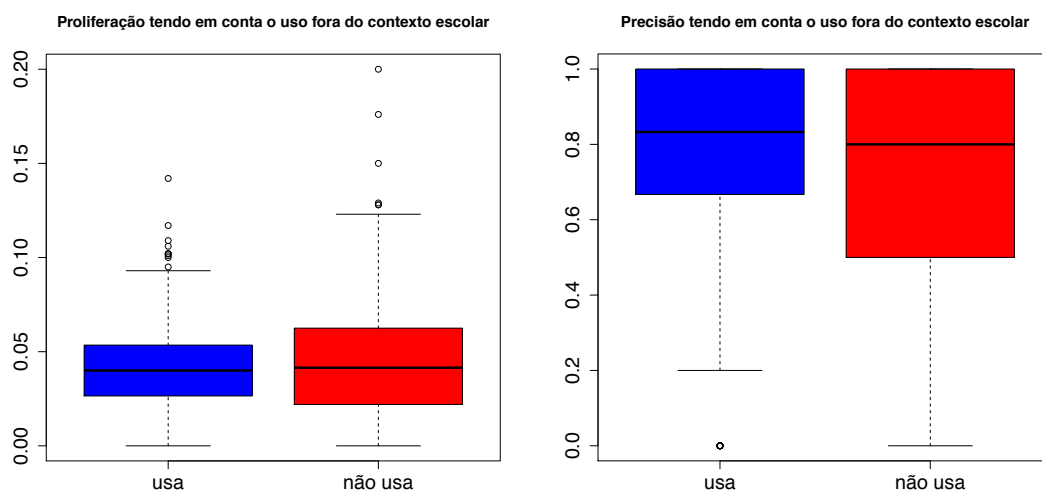


Gráfico 21 Proliferação e precisão tendo em consideração os dados sobre o uso da língua alvo fora do contexto escolar.

Apesar de os alunos aprenderem a língua no contexto de imersão, isto não quer necessariamente dizer que o aluno esteja a utilizar a língua regularmente fora

do contexto escolar. Para comunicar no dia-a-dia, muitas vezes os alunos acabam por recorrer a outra língua, normalmente o inglês. Acontece também, sobretudo com os alunos Erasmus, a formação de grupos por nacionalidades ou línguas maternas e o português acaba por não ser utilizado no convívio fora da universidade. Por estes motivos, vamos também analisar o desempenho dos alunos que usam e dos que não usam o português fora da sala de aula.

O gráfico 21 mostra-nos que a proliferação é ligeiramente mais alta nos alunos que admitem não usar a língua fora da sala de aula. Mas apesar deste valor ser mais alto nos alunos que não usam a L2, a precisão é mais alta nos alunos que usam. Também Fernández e Schmitt (2015) encontraram, nos seus dados, uma forte correlação entre o aumento do conhecimento colocacional e a utilização da língua alvo fora do contexto escolar, nomeadamente nos que usavam as redes sociais, iam ao cinema ou viam televisão na língua alvo.

6.5.6. Número de L2

Nesta secção analisamos a influência do número de L2 que os alunos conhecem na proliferação e na precisão. Apresentamos, seguidamente, a média e o desvio padrão destas medidas e mostramos os respetivos diagramas de caixa.

	Proliferação				Precisão			
	1 L2	2 L2	3 L2	+ 4 L2	1 L2	2 L2	3 L2	+ 4 L2
Média	0.046±	0.046±	0.039±	0.041±	0.734±	0.750±	0.735±	0.739±
	0.032	0.028	0.021	0.022	0.294	0.281	0.329	0.283

Tabela 35 Média e desvio padrão da proliferação e precisão para o número de L2.

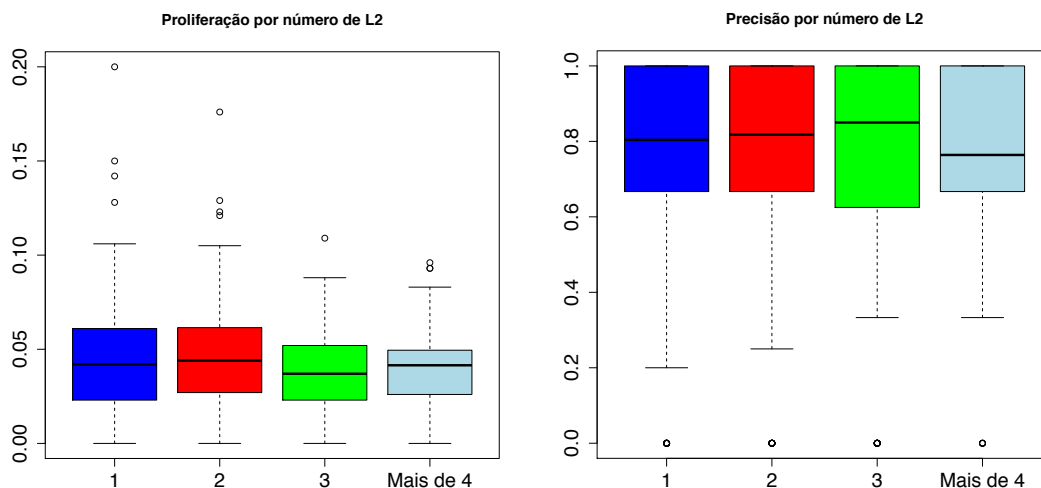


Gráfico 22 Proliferação e precisão tendo em consideração os dados sobre o nº de L2 que os alunos dominam.

Dada a importância que a importação de outras línguas assume na produção de colocações (secção 6.2.3.), outro fator que pode pesar no domínio das colocações será o número de línguas que o aluno domina, sem contar com a língua alvo, neste caso o português.

Como podemos ver no gráfico 22, a proliferação diminui com o domínio de três línguas, depois aumenta nos alunos que sabem quatro línguas. Não sendo muito visível um padrão, é provável que a proliferação não seja muito afetada pelo número de línguas. A precisão parece aumentar quando o aluno domina até três línguas, depois começa a diminuir. Isto pode demonstrar que o conhecimento de uma ou duas línguas é mais favorável à aprendizagem das colocações. No entanto, quando mais línguas são estudadas, podem também surgir mais dúvidas e problemas, resultando num desempenho mais fraco.

6.5.7. Idade

Apesar de a idade não ser uma variável muito relevante, nesta secção analisamos a influência da idade dos alunos na proliferação e na precisão. Apresentamos, seguidamente, a média e o desvio padrão destas medidas e mostramos os respetivos diagramas de caixa.

	Proliferação					Precisão				
	<19	20-29	30-39	40-49	>49	<19	20-29	30-39	40-49	>49
Média	0.0533	0.045±	0.039±	0.028±	0.038±	0.675±	0.738±	0.709±	0.810±	0.779±
	±0.017	0.028	0.021	0.017	0.028	0.179	0.298	0.288	0.240	0.367

Tabela 36 Média e desvio padrão da proliferação e precisão para as idades dos alunos.

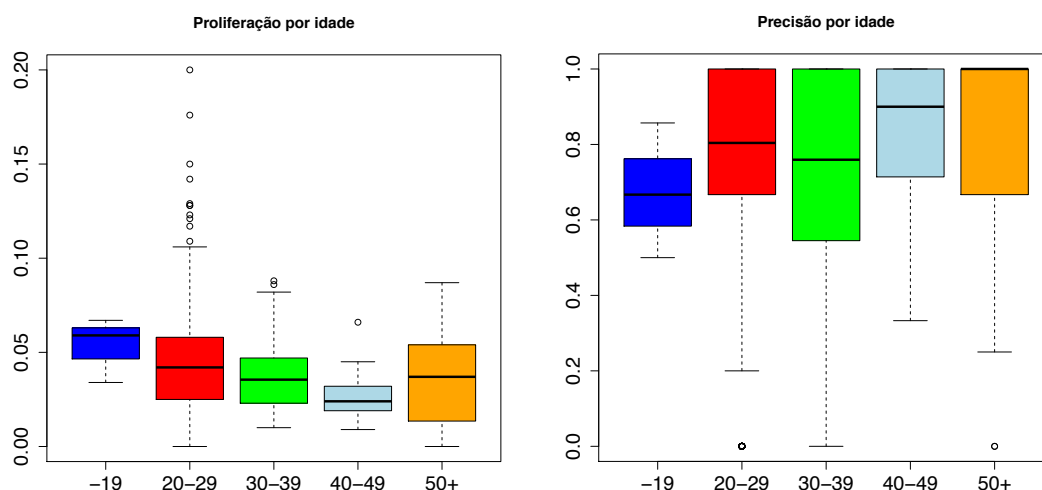


Gráfico 23 Proliferação e precisão tendo em consideração a idade dos alunos.

Como podemos ver no gráfico 23, a produção de colocações parece ser mais prolífica entre os alunos mais novos (até aos 19 anos). A partir dos 20 até aos 49 parece diminuir e volta a aumentar ligeiramente nos alunos com mais de 50 anos. A precisão apresenta os seus valores mais altos entre os alunos com mais de 50 anos. Isto pode demonstrar que os alunos mais velho arriscam menos na produção, usando apenas aquelas colocações que realmente dominam enquanto que os mais novos arriscam mais.

6.5.8. Formação académica

Por último, nesta secção analisamos a possível influência da formação académica dos alunos na proliferação e na precisão. Para não enumerarmos todos os cursos dos alunos, optámos por agrupá-los em dois domínios mais genéricos: Ciências e Humanidades. Apresentamos, seguidamente, a média e o desvio padrão da proliferação e da precisão e mostramos os respetivos diagramas de caixa.

	Proliferação		Precisão	
	Ciências	Humanidades	Ciências	Humanidades
Média	0.046±0.026	0.044±0.030	0.618±0.382	0.660±0.375

Tabela 37 Média e desvio padrão da proliferação e precisão para a formação acadêmica dos alunos.

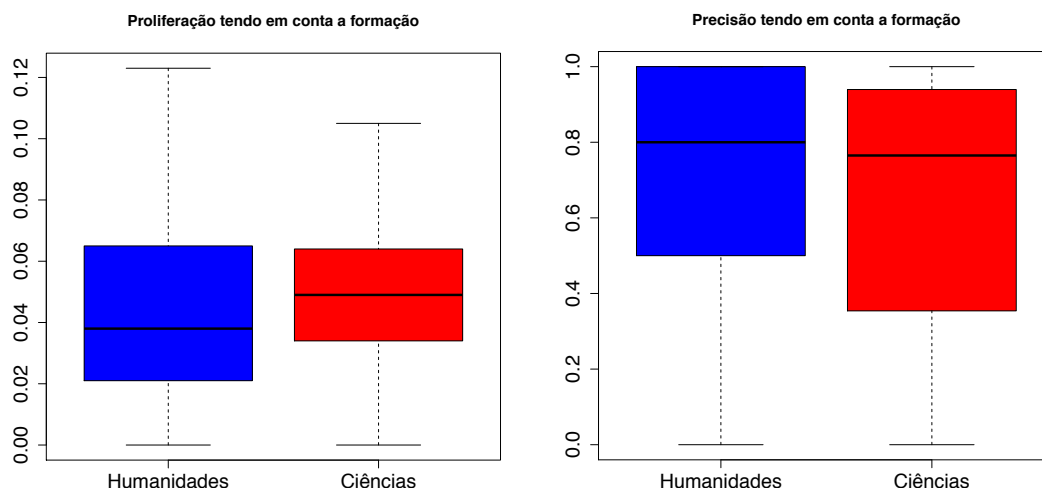


Gráfico 24 Proliferação e precisão tendo em consideração a formação académica dos alunos.

Como podemos ver no gráfico 24, os alunos de Ciências são os que apresentam o maior valor de proliferação. Porém, a precisão mais alta é obtida pelos alunos de Humanidades. Isto pode demonstrar que o conhecimento mais aprofundado de linguística e de várias línguas, que se espera que alguns alunos de Humanidades tenham, representa uma clara vantagem para o domínio das colocações numa L2.

6.6. Erros colocacionais em PL2 e em MT

Nesta última secção, vamos estabelecer uma comparação entre os erros colocacionais produzidos pelos alunos aprendentes de Português L2 e os erros em colocações gerados automaticamente por sistemas de tradução automática (MT).

A motivação para esta comparação reside no argumento, discutível, de que um sistema automático de tradução pode ser visto como um aprendente não-nativo

de uma língua que seria, neste caso, a língua fonte da tradução. Este argumento está relacionado com a facto de que nem um falante não-nativo, nem um sistema de tradução possuem ainda um conhecimento perfeito da língua alvo, podendo então existir transferência de algumas propriedades linguísticas da língua materna para a L2 ou, no caso da tradução automática, da língua fonte para a língua alvo. Hashemi e Hwa (2014) argumentam mesmo que as semelhanças entre estes dois tipos de produções podem ajudar a prever padrões de erros e, consequentemente, melhorar os modelos de tradução automática. O nosso objetivo será, no entanto, o inverso, ou seja, perceber se os erros em colocações produzidas por sistemas de tradução automática demonstram semelhanças com os erros produzidos pelos alunos de PL2 e se poderão ser usados para prever erros humanos.

Para podermos realizar esta comparação utilizámos os dados da análise de erros em colocações em tradução automática. Essa recolha e análise de dados foi levada a cabo no âmbito de um trabalho mais extenso para a criação de uma taxonomia de erros para tradução automática realizado durante a elaboração deste doutoramento. Os resultados deste trabalho encontram-se publicados em Costa *et al.* (2015). Nesse trabalho analisámos as colocações presentes num *corpus* gerado por quatro sistemas de tradução: Google Translate⁸⁵, Systran⁸⁶ e dois sistemas treinados com a tecnologia Moses, usando os dois modelos mais populares: *phrase based model* (Koehn et al., 2005) e *hierarchical phrase-based model* (Chiang, 2007). Usaram-se três textos diferentes para realizar a tradução de inglês para português: 250 frases retiradas de uma TED talk⁸⁷ (3.346 palavras), 250 frases da revista da TAP UP⁸⁸ (3.346) e 250 frases de um *corpus* de questões que foi traduzido para português (1.846) (Costa et al., 2012). No total foram encontradas 172 colocações: 47 no *corpus* da Ted talk, 84 no da TAP e 41 no *corpus* das questões. Como referimos, os três *corpora* foram traduzidos por quatro motores de tradução, assim, no total, avaliámos 188 colocações no *corpus* da Ted talk, 336 no *corpus* da TAP e 164 no *corpus* das questões. Apesar de termos usado *corpora* de natureza diferente e de termos usado quatro sistemas de tradução, a quantidade de dados não era muito

⁸⁵ <http://translate.google.com>

⁸⁶ <http://www.systranet.com/translate>

⁸⁷ <http://www.ted.com/>

⁸⁸ <http://upmagazine-tap.com/>

grande. Assim sendo, optámos por fazer uma análise global dos dados.

Após a identificação das colocações, avaliámos quais estavam corretas e quais estavam incorretas. O gráfico, em baixo, mostra-nos a proporção das colocações corretas e incorretas na totalidade do *corpus* de tradução.

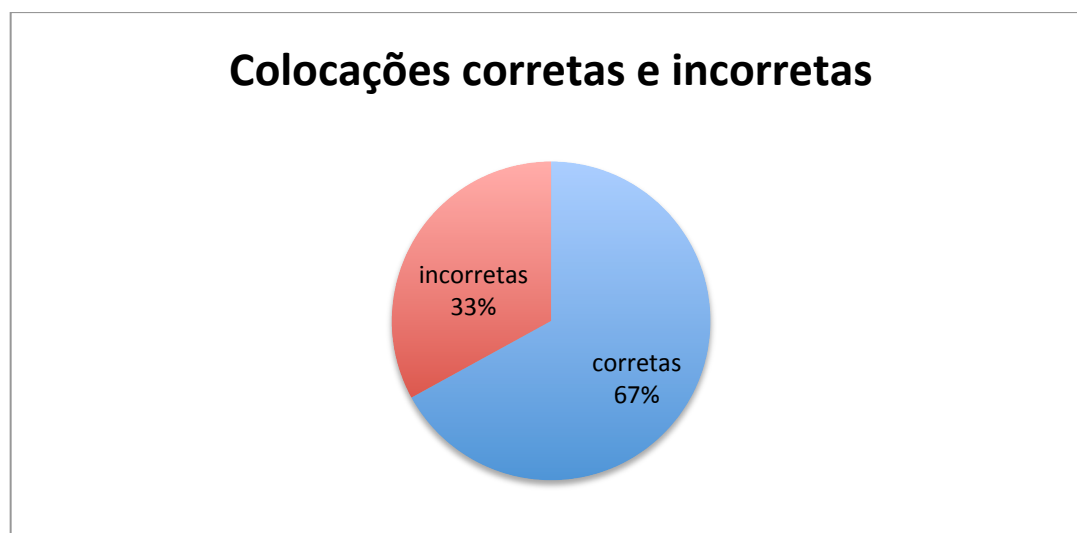


Gráfico 25 Proporção de colocações corretas e incorretas.

Como podemos ver através do gráfico 25, tal como acontecia no *corpus* dos aprendentes, a maioria das colocações encontra-se correta. O Google é o sistema que produz mais colocações corretas e menos erradas, e é o Systran o que produz mais incorretas e menos corretas.

Em seguida, as colocações que apresentavam algum erro foram classificadas seguindo os dois níveis de análise – localização e descrição – que descrevemos em mais detalhe na secção 5.4.2. Foram também marcados os erros gramaticais presentes nas colocações. No entanto, estes não iremos usar nesta análise, uma vez que não fizemos esse tipo de análise no *corpus* de PL2.

6.6.1. Localização dos erros colocacionais

Atentemos agora na localização dos erros. Como podemos ver, no gráfico 26, a maioria dos erros encontra-se no colocativo. Alguns erros que encontrámos no colocativo são, por exemplo, traduzir *take a bath* por **ter um banho* em vez de *tomar um banho* ou traduzir *take a nap* por **tomar uma sesta* em vez de

fazer/dormir uma sesta. Apesar de se tratar de erros automáticos, demonstram algumas semelhanças com erros que encontrámos nas produções do alunos, nomeadamente no uso problemático de *fazer* e *tomar*: **fazer um aperitivo/tomar um aperitivo*, **fazer cuidado/ tomar cuidado*, **fazer um banho/tomar um banho*.

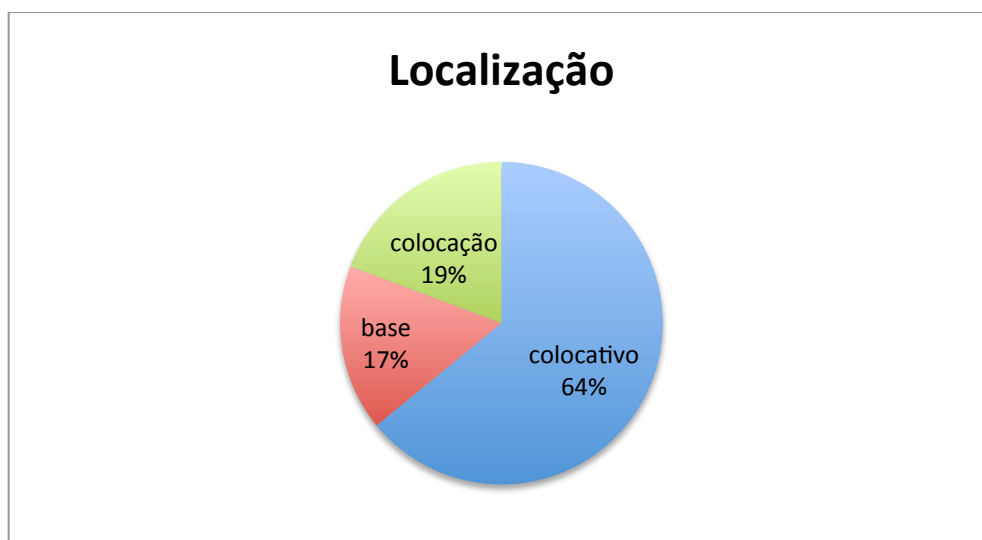


Gráfico 26 Localização dos erros de colocações em MT.

Apesar dos problemas na escolha do colocativo serem os mais comuns, foram também identificados casos de tradução errada da base. Por exemplo, *permanent collection* traduzido por *recolha permanente* em vez de *exposição permanente*, *plan a trip* traduzido por *planear deslocações* em vez de *planear uma viagem* e *act in a play* traduzido por *atuar num jogo* em vez de *atuar numa peça*. Trata-se de erros em que a tradução usada é uma possível aceção da palavra na língua fonte. No entanto, naquele contexto, não é a opção correta. A mesma lógica faz com que os alunos de PL2 criem erros semelhantes, como *despertar cedo* em vez de *acordar cedo* ou *alimentação deliciosa* em vez de *comida deliciosa*. Nestas casos, *despertar* é um sinónimo possível de *acordar* e *alimentação* também é um sinónimo de *comida*. Todavia, não são a opção apropriada, tendo em conta o colocativo selecionado.

Existem ainda 19% de erros que afetam a colocação globalmente, como no caso da tradução de *have supper* por **ter ceia* em vez de *jantar*. Nas produções dos alunos de PL2 encontrámos um erro muito semelhante: *tomar o jantar* em vez de *jantar*. Contudo, no caso da tradução automática trata-se de um erro de tradução palavra por palavra, enquanto que, no caso dos alunos, pode tratar-se de uma

sobregeneralização do verbo *tomar*, uma vez que se diz *tomar o pequeno-almoço*.

Podemos então concluir que estas percentagens não são muito diferentes da distribuição dos erros nos alunos de PL2. Os alunos de PL2 também apresentam uma grande dificuldade na seleção do colocativo, apesar de ocorrerem igualmente alguns erros na base e na colocação como um todo.

6.6.2. Descrição dos erros colocacionais

Atentemos, em seguida, no gráfico 27, que nos mostra a descrição dos erros nas colocações que encontrámos no *corpus* de tradução automática.

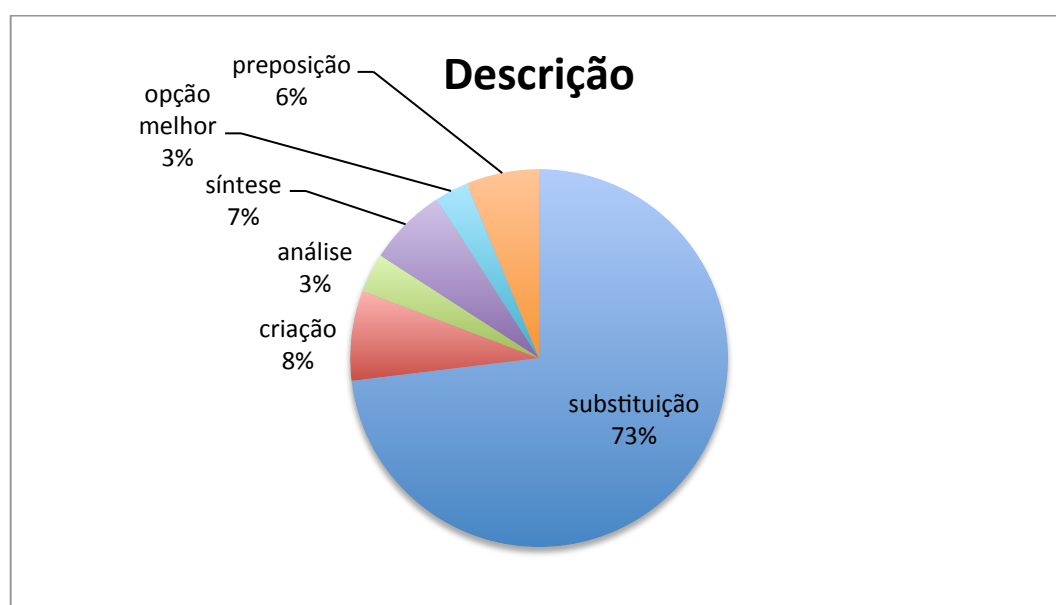


Gráfico 27 Descrição dos erros de tradução automática.

Se relembarmos o gráfico equivalente da descrição dos erros dos alunos de PL2 (secção 6.2.2.), verificamos que os erros de substituição são os mais comuns tanto nas produções dos alunos como nas traduções automáticas. Trata-se de erros em que um elemento da colocação é substituído por uma palavra que não se adequa, mas que existe na L2 ou na língua alvo, no caso da tradução. Alguns exemplos que encontrámos são **dizer a história* em vez de *contar uma história* como tradução de *tell a story*, **libertar um álbum* em vez de *lançar um álbum* como tradução de *release an album* ou **terra nativa* em vez de *terra natal* para traduzir *native land*. É fácil explicar este tipo de erro através da influência da língua de

partida, o que os torna também erros passíveis de encontrar nas produções dos alunos de PL2. No entanto, existem também algumas produções para as quais não conseguimos encontrar uma explicação, como a tradução escolhida de *wind blows* ter sido **vento funde* em vez de *o vento sopra*.

Em relação ao erro de criação, este é bastante mais comum em produções humanas do que em produções automáticas. Um sistema automático contém apenas as palavras que existem no léxico, sendo difícil produzir palavras com erros ortográficos. Já os alunos tendem a procurar semelhanças entre as palavras e a fazer raciocínios de generalização que podem resultar em erros ortográficos. Em todo o caso, optámos por incluir no erro de criação as palavras que não foram traduzidas e foram deixadas, neste caso, em inglês. Por exemplo, *mop the floor* foi traduzido para **mop o chão* em vez de *lavar o chão* e *gray hair* foi traduzido para *gray cabelo* em vez de *cabelo grisalho*. No entanto, este tipo de importação direta também não é incomum nos alunos estrangeiros, como no caso de **coger um táxi* (em vez de *apanhar um táxi*) produzido por um aluno de LM espanhola.

Foi também possível encontrar alguns erros de análise e de síntese (3% e 7%, respetivamente) entre as produções dos sistemas automáticos. No caso de erro de síntese, uma só palavra é produzida quando deveria ter sido gerada uma colocação, ou seja, *early evening* deveria ter sido traduzida por *início da noite*, no entanto foi traduzida por *noite*. Um exemplo contrário, de análise, é a suposta colocação já mencionada *ter ceia* (*have supper*) que deveria ter sido traduzida por um verbo apenas: *jantar*.

Representando 3% dos erros, temos os problemas de opção melhor. São os casos em que haveria uma opção melhor, mais próxima do que um nativo diria, mas que não se encontra completamente errada. Por exemplo, a tradução de *ask a question* por *perguntar uma pergunta* em vez de *fazer uma pergunta* ou o caso de *pluck fruit* traduzido por *arrancar fruta (da árvore)* em vez de *apanhar fruta*.

Por último, os casos de utilização incorreta de preposição representam 6% dos erros. Tal como acontece com os erros humanos, os problemas podem dever-se a uma troca de preposição, uso de preposição quando não é necessário e o inverso. Por exemplo, no caso de *watch a game* traduzido por **assistir um jogo* em vez de *assistir a um jogo*, falta a preposição. Para traduzir *shoot a pistol*, o tradutor

automático produz **dispara na pistola*, acrescentando a preposição *em*, em vez de *disparar a pistola*. No caso de *this for knowledge*, a tradução produzida foi *sede para o conhecimento* em vez de *sede de conhecimento*, trocando a preposição *de* por *para*.

Esta comparação entre erros colocacionais dos alunos de PL2 e das produções dos sistemas de tradução automática, apesar de pequena, permitiu-nos tirar algumas conclusões. Primeiro, comprovámos que é possível classificar os erros automáticos com uma taxonomia desenhada para erros de alunos. Segundo, esta classificação permitiu-nos perceber que os erros colocacionais, tanto humanos como automáticos, tendem a localizar-se no colocativo e que o erro de substituição é também o mais comum em ambos os casos. Por último, observámos que a maioria dos erros se deve à influência da língua fonte, o que equivale à influência da LM no caso dos alunos. Tais observações levam-nos a crer que as colocações produzidas automaticamente conseguem mimetizar os erros dos alunos estrangeiros. Assim, podem não só ser usadas para prever dificuldades, mas também ajudar na criação de material didático direcionado a alunos de uma determinada língua materna.

Conclusão

A presente dissertação tem como objetivo geral contribuir para a área de investigação da aprendizagem do léxico em Português L2, mais concretamente perceber como um particular tipo de estruturas fraseológicas – as colocações – são utilizadas pelos alunos de PL2. Procurou-se, primeiramente, fazer uma descrição das colocações usadas em contexto para, depois, nos centrarmos no uso incorreto destas e nos tipos de erros encontrados, sendo esta análise central para avaliarmos a competência colocacional dos alunos.

Neste capítulo conclusivo, resumo o trabalho realizado nos capítulos anteriores, faço um balanço dos principais resultados do trabalho apresentado na tese e elenco as nossas contribuições, apresento as limitações da nossa investigação e considero possíveis direções para pesquisas futuras

Resumo do trabalho realizado

Na parte inicial da dissertação, dedicada à dilucidação de questões de âmbito teórico, delimitou-se, conceptualmente, o tipo de estruturas linguísticas que se tomaria por objecto de estudo. Do elevado número de expressões, de natureza muito heterogénea, que integram a classe das fraseologias, optou-se por seleccionar as colocações. Como foi oportunamente referido, a escolha recaiu sobre estas estruturas porque, não causando problemas do ponto de vista da descodificação, elas provocam, contudo, particulares dificuldades em momentos de codificação, já que nem sempre são entendidas pelos alunos como unidades fraseológicas. Assim,

apenas em momentos de tradução no sentido da LM para a L2 o aluno se dará conta de que nem sempre existe equivalência total entre as colocações da sua LM e as da língua-alvo de aprendizagem, neste caso o português. Um não nativo dificilmente consegue prever estas restrições. Com foi dito, sendo as colocações transparentes, não existe um esforço consciente para as decodificar, porém esse esforço é necessário para a codificação. Não sendo necessárias operações mentais conscientes na sua compreensão, não deixam rasto na memória, daí a necessidade do seu ensino/aprendizagem.

Depois de ter sido esclarecida a natureza das colocações, no terceiro capítulo desta dissertação propomos uma reflexão sobre o papel das colocações no ensino de L2, nomeadamente, se devem ser abordadas em contexto escolar e quais os materiais mais apropriados para o seu ensino. Seguidamente, apresentamos os trabalhos mais pertinentes sobre o uso de colocações em *corpora* de aprendizagem de L2. Centramo-nos na definição de colocação e nas taxonomias de erro que usam para classificar os problemas encontrados nas colocações erradamente produzidas. Segundo Howarth (1998b: 5), trabalhos desta natureza são essenciais, uma vez que existe mais literatura sobre os idiomatismos, apesar de serem menos frequentes na língua e de também causarem menos problemas aos alunos do que as colocações.

Para levarmos a cabo a nossa investigação, a constituição do *corpus* é um passo fundamental. Assim, a primeira parte do capítulo 4 é dedicada à descrição do processo de recolha de *corpora* produzidos por alunos de PL2. Na segunda parte desse mesmo capítulo, apresentamos algumas ferramentas de extração de colocações em Português Europeu e estabelecemos uma comparação entre elas. No capítulo cinco, estabelecemos os nossos critérios definitórios de colocação, descrevemos o processo de identificação manual de colocações que foi levado a cabo e apresentamos a taxonomia de erros que usámos.

No capítulo central das nossa investigação, propomo-nos averiguar quais os padrões sintáticos mais utilizados pelos alunos, os verbos e os substantivos mais usados na construção de colocações e, por último, quais as colocações mais comuns. Comparamos também estes aspetos com a produção colocacional de nativos. No que concerne às colocações produzidas pelos aprendentes que revelaram erros na sua construção, propomo-nos compreender, através de uma taxonomia apropriada,

a localização, a descrição e a explicação desses erros. Para podermos avaliar quantitativamente a competência colocacional dos alunos, sugerimos métricas para medir a proliferação e a precisão do uso das colocações. Depois de compararmos a proliferação e precisão dos aprendentes com os nativos, estas métricas foram depois relacionadas com variáveis individuais dos alunos, como anos de estudo de português ou número de línguas que o aluno conhece. Por último, comparamos os erros colocacionais produzidos pelos alunos aprendentes de Português L2 e os erros em colocações geradas automaticamente por sistemas de tradução automática.

Principais conclusões

As principais conclusões que podemos tirar deste trabalho de investigação estão relacionadas com os seis pontos que, em seguida, elencamos:

- **Avaliação dos motores de extração**

Para esta avaliação utilizámos quatro sistemas de extração de colocações para o Português Europeu: *DeepDict*, *Sketch Engine*, *CRPC* e *Wortschatz*. As pesquisas foram feitas, usando como base das potências colocações, dez substantivos.

Como foi já dito, o *Sketch Engine*, apesar de não ser um sistema gratuito, encontra mais colocações do que os restantes sistemas. Esta ferramenta obtém melhores resultados do que as outras em vários aspetos da nossa avaliação, nomeadamente permitindo um ajuste bastante fino dos parâmetros de pesquisa (limiar de frequência, procura por janela de palavras, procura por lema ou padrão) e permitindo a visualização dos resultados de acordo com diferentes métricas. No entanto, estas ferramentas demonstram também limitações, pois a maioria delas apenas permite fazer pesquisas partindo da base, o que inviabiliza fazer procuras se não possuímos esta informação. Apesar do *Sketch Engine* permitir a pesquisa da colocação inteira, esta hipótese também não se verificou viável para os nossos dados, uma vez que os erros ortográficos presentes nos mesmos fazem com que a análise automática se revele problemática. Esta mesma questão também se torna

problemática se optarmos por fazer uma pesquisa por padrão sintático, uma vez que os erros ortográficos podem gerar uma etiquetagem sintática errada, enviesando os resultados da procura. A resolução destas duas últimas limitações pode passar pela utilização de textos anteriormente corrigidos. A extração prévia das bases para as quais queremos encontrar os colocativos, será uma solução para os sistemas que apenas permitem fazer pesquisas partindo de uma determinada palavra.

- **Caracterização da produção colocacional dos alunos de PL2**

A análise das colocações presentes em *corpora* de PL2 permitiu-nos observar que o número de colocações (em relação ao número de palavras) parece decrescer nos dois últimos níveis em relação ao primeiro. Este comportamento pode ser explicado, na nossa opinião, através de uma estratégia de evitação. Este mecanismo é também observado por Philip (2007) que o reconhece como sendo uma das reações possíveis perante a falta de conhecimento colocacional.

Em relação ao número de colocações corretas e incorretas produzidos pelos alunos de PL2, observámos um aumento de produções corretas e uma diminuição das colocações erradas nos níveis mais proficientes. Este resultado demonstra que alunos de níveis superiores revelam um domínio maior das colocações e, corroborando as conclusões anteriores, alunos mais proficientes correm menos riscos e, portanto cometem menos erros.

No que concerne aos padrões sintáticos, o mais utilizado na construção de colocação, tanto corretas como erradas, é V + N. Colocações que incluem adjetivos e advérbios parecem ser menos frequentes. Uma possível explicação encontra-se no facto de os alunos dominarem substantivos e verbos antes de chegarem ao domínio dos adjetivos e advérbios (cf. Palapanidi *et al.*, 2014).

Observando os verbos mais utilizados na construção de colocações, concluímos que os verbos leves são os mais utilizados, sendo o mais comum o verbo *ter*. Os aprendentes de uma língua estrangeira tendem a usar combinações que são constituídas por itens lexicais de alta frequência, como o verbo *supra* mencionado, constituindo isto uma estratégia para contornar a utilização de itens de vocabulário mais sofisticados. Em relação aos substantivos, a maioria dos nomes utilizados são

substantivos concretos bastante representativos das temáticas sobre as quais os alunos escreveram, nomeadamente, as férias, os tempos livres ou os seus *hobbies*. Sendo as temáticas um pouco diferentes entre os níveis, a escolha de vocabulário demonstra essas diferenças. No entanto, parece-nos que existe alguma evolução ao longo dos níveis visível na escolha lexical. No primeiro nível, os alunos usam substantivos mais simples, como *avião* ou *hotel*, enquanto que no terceiro nível surge vocabulário mais complexo como *consideração* ou *progresso*. Ainda em relação ao terceiro nível, registámos uma maior variedade do ponto de vista verbal, nominal e adjetival, assim observamos colocações como *ter consideração*, *terra árida*, *vasto leque*, *vento forte*, *vital importância*.

Estas observações levam-nos a concluir que existe progresso na aprendizagem das colocações, aumentando o número de estruturas corretas e diminuindo o número de erros ao longo dos três níveis. No entanto, o número de colocações diminui no segundo e terceiro nível, mostrando assim a importância das estratégias de evitação.

- **Taxonomia de erros em colocações**

A primeira etapa de análise da taxonomia de erros centra-se na sua localização, assim observámos que a maioria dos erros se encontram no colocativo. É inevitável, segundo Palmer (1933: 8 *apud* Howarth, 1998) que na ausência de informação colocacional, o aluno tenha que formar a combinatória adivinhando qual será o colocativo correto ou através de uma analogia da língua materna, criando estruturas pouco comuns como, **make a question*, **perform a favour*, **to do trouble*, **to keep patience*. Apesar da maioria dos erros que a literatura documenta ser no colocativo (Hausmann, 1979; Nesselhauf, 2005; Alonso Ramos *et al.*, 2010), salientamos que, contrariamente ao que alguns estudos sugeriam, erros na base ou na totalidade da colocação são também comuns e, portanto, devem também ser tratados em qualquer análise aos erros colocacionais.

No que concerne à descrição dos erros, concluímos que a maioria deles se deve a uma substituição ou a uma criação de uma palavra. No primeiro nível, observámos mais criações, que demonstram uma forte influência da LM ou de outra

L2. No segundo e terceiro nível, as substituições vão aumentando. Fazer uma substituição requer já um nível de conhecimento da língua mais avançado, nesse sentido, é natural que este tipo de erro surja com valores superiores no segundo nível e ainda mais altos no terceiro. A utilização de uma colocação mas com um sentido diferente do desejado (sentido diferente) aumenta também ligeiramente no segundo nível. Mais uma vez, trata-se de uma estratégia um pouco mais sofisticada que é mais usada no segundo nível do que no primeiro.

A explicação do erro mais comum entre os nossos dados é a transferência da LM ou de outra L2, independentemente do nível do aluno. Como foi já dito, os alunos não podem ser considerados *phraseologically virgin territory*, uma vez que possuem já conhecimento fraseológico da LM ou de outra L2 que, inevitavelmente, assume um papel na aquisição de novas construções. Sobretudo nas fases iniciais do ensino de uma língua, os alunos são mais expostos a uma metodologia pedagógica que favorece a equivalência, tornando-os inclinados para optarem por uma tradução palavra por palavra. Não sendo dada muita importância ao significado fraseológico e à equivalência fraseológica no processo de aprendizagem, é preciso coragem para um aluno deixar de produzir padrões familiares e decalcados da LM, e experimentar padrões lexicais desconhecidos.

Apesar da prevalência das importações, a extensão é o segundo tipo de erro mais usado e parece aumentar ao longo dos três níveis de proficiência. Isto pode demonstrar que o aluno do primeiro nível está mais dependente do conhecimento que possui de outras línguas (LM ou L2), enquanto que alunos de níveis superiores começam a usar mecanismos mais sofisticados, como a extensão do significado de uma palavra.

Para os alunos de LM espanhola, a importação parece ser o mecanismo mais utilizado face a um conhecimento incompleto da L2. Este comportamento é detetável tanto no nível 1 como no nível 2. A escolha deste mecanismo pode ser justificada através das semelhanças que existem entre o português e o espanhol, podendo estas levar os alunos a acreditar que a importação tem uma alta probabilidade de funcionar a seu favor. Os alunos de LM francesa, tal como os alunos das restantes LMs, são os alunos que apresentam maior número de extensões e de derivações erradas. Não sendo o francês tão semelhante ao

português, como o italiano ou o espanhol, é natural que os alunos optem por usar outros mecanismos, além da transferência.

Olhando para as línguas das quais os alunos mais importam, concluímos que a LM de cada um dos alunos é o principal apoio para a transferência, sendo os alunos de LM espanhola os que mais o fazem e os alunos de LM inglesa os únicos que recorrem mais a outra língua (espanhol) do que à própria língua materna. Destes comportamentos podemos concluir que quanto mais próxima do português for a língua materna do aprendente, mais esta será usada como suporte e que os alunos estão cientes desta distância.

- **Métricas e impacto de variáveis individuais**

Como foi explicado em mais detalhe na secção 6.3, estabelecemos duas métricas que têm como objetivo medir a competência colocacional, ou seja, a proliferação e a precisão na produção de colocações. Essas métricas foram, depois, contrastadas com as informações que possuímos sobre os alunos, como idade, anos de estudo de português, para percebermos quais delas são preponderantes na aprendizagem de colocações.

Através da análise destes dados concluímos que a proliferação parece baixar ao longo dos três níveis. Estes dados confirmam a hipótese de que os aprendentes de níveis mais elevados, tendem a evitar o uso de colocações. Possuindo um conhecimento mais profundo da língua alvo, podem mais facilmente evitar estas estruturas e optar por uma reformulação com a qual se sintam mais confiantes. Por sua vez, a precisão parece aumentar, mostrando que na totalidade de colocações produzidas, o número de colocações corretas aumenta.

Em relação à LM, apesar do italiano e espanhol serem línguas mais próximas do português em relação ao alemão, são os alunos alemães que produzem mais colocações corretas. Normalmente, alunos de línguas maternas mais próximas do português são favorecidos pelas semelhanças entre as línguas. No entanto, será necessário que os alunos tenham consciência dessa proximidade para poderem tirar partido dela.

No que concerne ao número de anos de aprendizagem do português, não foi

encontrada nenhuma relação, já a aprendizagem em imersão parecer ter um papel positivo na aprendizagem de colocações. Em relação, à utilização da língua alvo fora do contexto escolar, os alunos que disseram usar o português fora da sala de aula demonstram uma maior precisão no uso das colocações.

Tendo em consideração o número de L2 que o aprendente domina, a precisão parece aumentar quando o aluno fala até três línguas; depois começa a diminuir. Isto pode demonstrar que o conhecimento de uma ou duas línguas ajuda na aprendizagem das colocações, no entanto se o número de línguas for superior a quatro, podem também surgir mais erros, resultando num desempenho mais fraco.

Em relação à idades dos alunos, concluímos que a produção de colocações para ser mais prolífica entre os alunos mais novos (até aos 19 anos). A precisão apresenta os seus valores mais altos entre os alunos com mais de 50 anos. Isto pode demonstrar que os alunos mais velho arriscam menos na produção, usando apenas as colocações que realmente dominam enquanto que os mais novos arriscam mais.

O percurso académico parece ter algum impacto na aprendizagem das colocações. São os alunos de Humanidades, em comparação com os alunos de Ciências, os que mostram ter a maior precisão colocacional. Isto demonstra que o conhecimento mais aprofundado de linguística ou de várias línguas, que os alunos de Humanidades podem ter, representa uma vantagem para o domínio das colocações numa L2.

- **Comparação entre a competência colocacional nativa e não-nativa**

Usando as métricas já definidas, comparamos a proliferação e precisão dos nativos e não-nativos. Olhando para a proliferação de colocações dos aprendentes e dos falantes nativos, constata-se uma óbvio “subutilização” de colocações por parte dos primeiros. Em relação à precisão, esse valor é um, no caso dos nativos, uma vez que as produções destes foram consideradas todas corretas.

Atentando agora, especificamente, nos padrões sintáticos, o mais usado pelos nativos e não-nativos é V + N. A diferença mais significativa prende-se com as colocações compostas por Adv + Adj, sendo uma combinação bastante usada pelos nativos e a menos usada pelos não-nativos. A pouca utilização desta estrutura por

parte dos alunos estrangeiros, pode, mais uma vez e como foi já dito, dever-se ao facto de os alunos dominarem substantivos e verbos antes de chegarem ao domínio dos adjetivos e advérbios, cuja presença, segundo a literatura, parece aumentar em níveis de proficiência mais elevados.

Em relação aos verbos utilizados na construção das colocações, existe uma grande coincidência entre os verbos usados tanto por nativos como não-nativos. No entanto, em torno de algumas bases, como *tempo*, os aprendentes demonstram mais variabilidade na escolha dos colocativos do que os próprios nativos. Já em relação aos substantivos, demonstram uma menor variação e tendem a seleccionar nomes do tipo mais concreto e menos abstrato. No que concerne às colocações usadas, sendo as temáticas dos textos dos alunos estrangeiros e dos nativos semelhantes, as colocações produzidas pertencem às mesmas áreas semânticas. No entanto, existem diferenças no tipo de colocações usadas: os nativos tendem a seleccionar colocações mais restritas, enquanto que os aprendentes usam colocações que demonstram pouca coesão, usando colocativos com maior potencial de combinação, como os verbos *fazer* ou *ter*. Em comparação com os nativos, os alunos tendem também a usar repetidamente algumas colocações, em vez de usar outras opções. Estas construções preferidas pelos alunos são chamadas as “collocational teddy bears” por Hasselgreen (1994) porque os alunos se sentem seguros a usá-las, evitando assim alguns potenciais erros.

Concluo, dizendo que estes resultados não são particularmente surpreendentes e vão ao encontro dos resultados de outras áreas de investigação da aprendizagem de L2, nomeadamente, do vocabulário monolexémico

- **Comparação entre os erros colocacionais em PL2 e em MT**

A comparação entre os erros colocacionais encontrados nos *corpora* de PL2 e nos textos traduzidos automaticamente, permitiu-nos tirar algumas conclusões que podem ter algum impacto no ensino/aprendizagem das colocações.

Em primeiro lugar, provámos que uma taxonomia desenhada para classificar erros humanos em colocações serve também para classificar erros automáticos, demonstrando que existem semelhanças entre eles. Consequentemente, esta

classificação permitiu-nos perceber que os erros colocacionais, tanto humanos como automáticos, tendem a localizar-se no colocativo e que o erro de substituição é também o mais comuns em ambos os casos. Finalmente, concluímos que a grande maioria dos erros de colocações em tradução automática se deve à influência da língua de partida, o que será o equivalente à influência da LM, no casos dos aprendentes. Tais conclusões, levam-nos a crer que as colocações produzidas automaticamente conseguem mimetizar os erros dos alunos estrangeiros, sendo que podem ser usadas para prever dificuldades, mas também para ajudar na criação de material didático direcionado a alunos de uma determinada língua materna.

Contribuições

Em forma de resumo, nesta secção, elencamos as contribuições que esperamos dar através da elaboração desta dissertação, assim:

- Estabelecemos critérios definitórios de colocação, pois, no nosso entender, uma categorização do fenómeno é fundamental para melhorar o seu ensino/aprendizagem;
- Contribuímos com dados empíricos provenientes de uma recolha de *corpora* de produções de alunos de PL2. Trata-se de material real que apresenta mais vantagens para avaliarmos os problemas dos alunos do que o uso de testes de licitação ou exercícios de tradução;
- Identificámos as colocações presentes no *corpus*, o que permite reconhecer as principais necessidades dos alunos. A observação dos dados permite-nos dispor de informação não só sobre os erros colocacionais dos estrangeiros mas também sobre a sua destreza na hora de produzir colocações e assim detetar as carências no seu uso. Esta informação pode não só contribuir para o trabalho do professor, oferecendo-lhe uma boa referência para elaborar material didático adequado às necessidades reais do estudante, mas pode também servir de base para a criação de um dicionário;

- Avaliámos vários motores de extração de colocações: a análise das suas funcionalidades e limitações podem servir de base para a criação de novas ferramentas nesta área;
- Provando que os erros colocacionais não são homogéneos, definimos uma taxonomia especializada em erros colocacionais. No nosso entender, para obter as verdadeiras dificuldades experienciadas pelos alunos e para se oferecer ajuda eficiente e orientada às dificuldades reais, é necessário uma análise linguística fina dos erros em colocações. Muitos dos estudos centram-se num tipo particular de erro e não permitem uma visão geral do fenómeno, mais detalhes sobre os erros podem ajudar na correção do erro mais concreta e providenciar material didático direccionado ao tipo de dificuldade. Consequentemente este tipo de informação pode facilitar o desenvolvimento de técnicas para a correção automática de erros colocacionais nas produções dos estrangeiros. Por último, sendo uma taxonomia independente da língua, poderá servir à comunidade.
- Definimos métricas que pretendem avaliar os vários aspetos da competência colocacional dos alunos de PL2;
- Demonstramos quais são as variáveis individuais dos alunos que de facto impactam o processo de aprendizagem das colocações;
- Contribuímos também com uma caracterização da produção colocacional dos nativos do português, estabelecendo assim uma comparação entre estes e os aprendentes;
- Comprovámos que os erros em colocações gerados automaticamente podem ser comparados com erros humanos e, como tal, podem servir de base para a criação de exercícios ou outro tipo de material didático.
- Finalmente, no contexto desta tese foram publicados seis artigos, um em revista e seis em conferências, encontrando-se ainda em preparação um artigo a ser submetido à revista *Studies in Second Language Acquisition*. Segue a lista dos artigos:

2016

Ângela Costa, Rui Correia, Luísa Coheur, **Building a corpus of errors and quality in Machine Translation: experiments on error impact**, *In* 10th edition of the Language Resources and Evaluation Conference, Slovenia, May

Ângela Costa and Luísa Coheur, **Towards a Statistical-enriched Corpus Containing Portuguese Collocations in Use: reviewing possible extraction tools**, *In* Computational Processing of the Portuguese Language. 12th International Conference PROPOR2016, Springer-Verlag, Lecture Notes in Artificial Intelligence Series, Tomar, July 2016

Ângela Costa, Luísa Coheur, Teresa Lino, **A study on the production of collocations by European Portuguese learners**, *In* 12th Workshop on Multiword Expressions (MWE 2016) ACL 2016, Berlin, Germany, August 11, 2016

2015

Ângela Costa, Wang Ling, Tiago Luís, Rui Correia, Luísa Coheur, **A linguistically motivated taxonomy for Machine Translation error analysis**, Machine Translation, Springer, vol. 29, pages 127-161, doi: 10.1007/s10590-015-9169-0, June 2015

Ângela Costa, Luísa Coheur, Teresa Lino, **Challenges on the translation of collocations**, *In* Workshop on Multi-word Units in Machine Translation and Translation Technology, Malaga, Spain, June 2015

2014

Ângela Costa, Tiago Luís, Luísa Coheur, **Translation Errors from English to Portuguese: an Annotated Corpus**, *In* LREC 2014, Ninth International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), Iceland, May 2014

Limitações e sugestões para trabalho futuro

Como nota final, gostaria de salientar que, não obstante este trabalho se tenha proposto a lançar alguma luz sobre alguns aspetos da produção de colocações em contextos de português como L2, outras questões, a equacionar em futuros

trabalhos, ficaram ainda por explorar. Refiro-me, nomeadamente, a um alargamento do estudo a alunos de línguas maternas mais distantes do português ou, ainda, a uma investigação mais exaustiva sobre colocações construídas em torno da mesma base ou do mesmo colocativo. Apesar de nos termos centrado na descrição da utilização das colocações por parte dos aprendentes, são também necessárias descrições contrastivas através de *corpora* bilingue. Estabelecendo comparações entre a língua materna do aprendente e a L2, será mais fácil avaliar possíveis áreas de dificuldade e a influência da LM e, conseqüentemente, produzir material pedagógico apropriado para ajudar grupos de LMs. Seria também importante perceber se o aluno aprendeu de facto as colocações porque, se a forma correta for igual à língua materna, o aluno pode estar simplesmente a decalcar.

Nesta dissertação, recolheram-se vários dados sobre as colocações utilizadas pelos alunos, nomeadamente os padrões sintáticos destas e as bases das colocações usadas. Como tarefa futura, esta informação poderá contribuir para facilitar o processo de identificação de colocações. Através da utilização dos padrões sintáticos das colocações recolhidas manualmente, podem-se realizar pesquisas em *corpora* informatizado e morfologicamente etiquetado, nomeadamente no COPLE2⁸⁹. Assim, poderemos obter uma lista de candidatos a colocação. Utilizando as bases, poderá ser feita uma pesquisa no *Corpus* de Referência do Português Contemporâneo e verificar se os colocativos selecionados pelos aprendentes são os mesmo que os falantes nativos usam. Certas combinações que são pouco frequentes na língua podem justificar um uso incorreto por parte dos alunos.

Como foi já dito, as colocações demonstram um grande peso na língua e uma elevada frequência na linguagem quotidiana e como tal, a sua aprendizagem correta deve considerar-se um objetivo primordial no processo de aquisição de uma segunda língua. Esperamos que este trabalho possa evidenciar a importância destas estruturas serem explicitamente ensinadas no contexto de sala de aula, a falta de material didático apropriado para o seu ensino e, sobretudo, a necessidade da

⁸⁹ <http://alfclul.clul.ul.pt/teitok/learnercorpus/index.php?action=home>

criação de um dicionário que apresente uma descrição mais detalhada deste tipo de construções, nomeadamente mostrando opções de colocativos para cada palavra, informação que normalmente não encontramos num dicionário monolíngue.

Bibliografía

- Ágel, Vilmos (2004), "Phraseologismus als (valenz)syntaktischer Normalfall", in: Steyer, Kathrin (Hrsg.), 65-86.
- Aisenstadt, Esther (1981), "Restricted collocations in English lexicology and lexicography", *ITL Review of Applied Linguistics* 53, 53–61.
- Aisenstadt, Esther (1979), "Collocability restrictions in dictionaries", in: R.R.K. Hartmann (ed.) *Dictionaries and their Users*, Exeter Linguistic Studies, 71-74.
- Alonso Ramos, Margarita (1993), *Las funciones léxicas en el modelo lexicográfico de I. Mel'čuk*, Tese de Doutoramento Inédita. Departamento de Lengua Española, Madrid, UNED.
- Alonso Ramos, Margarita (1994-1995), "Hacia una definición del concepto de colocación: de J. R. Firth a I. A. Mel'čuk", *Revista de Lexicografía*. 1, 9-28.
- Alonso Ramos, Margarita (2003), "Hacia un Diccionario de colocaciones del español y su codificación", in: Martí, M. A. Fernández, A. Vázquez, G. (eds.), *Lexicografía computacional y semántica*, Barcelona, Edicions de la Universitat de Barcelona, 11-34.
- Alonso Ramos, Margarita (2006b), "Entón, ¿é unha colocación ou non?: análise contrastiva das colocacións", *Cadernos de Fraseología Galega*, 8, 29-43.
- Alonso Ramos, Margarita (2010), "No importa si la llamas o no colocación, descríbela". Mellado, C. et al. (eds.), *Nuevas propuestas para el español y el alemán*, Frank & Timme, Berlin, 55-80.
- Alonso Ramos, Margarita/Wanner, Leo/Vincze, Orsolya/Casamayor del Bosque, Gerard/Vázquez Veuga, Nancy/ Mosqueira Suárez, Estela/Prieto González,

- Sabela (2010), "Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora", in: N. Calzolari et al. (eds.) *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Paris: ELRA, 3209-3214.
- Alonso Ramos, Margarita/Wanner, Leo/Vázquez Veira, Nanzy/Vincze, Orsolya, Mosqueira/Suárez, Estela Mosqueira/Prieto González, Sabela (2010^a), "Tagging collocations for learners", in: S. Granger and M. Paquot, editors, *eLexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLex 2009, Cahiers du Cental*, volume 7, Louvain-la-Neuve.
- Alonso Ramos, Margarita (2013), "Colocaciones, diccionario y corpus de aprendices", in: Casas Gómez, M. (dir.), Vela Sánchez, R. (ed.), *Eugenio Coseriu, in memoriam, XIV Jornadas de Lingüística*. Cádiz: Universidad de Cádiz, 57-71.
- Altenberg, B. (1993). "Recurrent verb-complement constructions in the London-Lund corpus", in: J. Aarts, P. de Haan, & N. Oostdijk, *English language corpora: design, analysis and exploitation*. Amsterdam: Rodopi, 227±245
- Altenberg, Bengt (2001 [1998]), "On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations", in: Cowie, A. P. (ed.), 101-122.
- Altenberg, Bengt/Granger, Sylviane (2001), "The grammatical and lexical patterning of make in native and non-native student writing", *Applied Linguistics* 22 (2), 173-194.
- Antunes, Sandra/Mendes, Amália (2015), "Portuguese multiword expressions: Data from a learner corpus", in: *Third Learner Corpus Research Conference*, Radboud University Nijmegen, September.
- Athayde, Maria Francisca (1999), "Construções com verbo-suporte e fraseologias no Alemão e no Português", *Runa. Revista Portuguesa de Estudos Germanísticos*, 27, 1999 (1997-1998), 285-304.
- Athayde, Maria Francisca (2000), *A estrutura semântica das construções com verbo-suporte preposicionadas do Português e do Alemão*. Dissertação de Doutoramento em Linguística Contrastiva apresentada à Faculdade de Letras da Universidade de Coimbra.
- Athayde, Maria Francisca (2001), *Construções com verbo-suporte (Funktionsverbgefüge) do Português e do Alemão*, Centro Interuniversitário de Estudos Germanísticos, Coimbra: Minerva Coimbra (= cadernos do ciegl, n.º 1).
- Athayde, Maria Francisca (2007), *Relatório de Programa, Conteúdos e Métodos de*

Ensino Teórico e Prático das Matérias da Disciplina de “Seminário de Linguística”. Relatório apresentado ao concurso para provimento de uma vaga de Professor Associado do 4º Grupo (Estudos Germanísticos) da Faculdade de Letras da Universidade de Coimbra.

Athayde, Maria Francisca (2008), “Modelos fraseológicos” em português e em alemão”, *Lusorama, Zeitschrift für Lusitanistik*, Nr. 73-74 (Maio de 2008), 94-116.

Athayde, Maria Francisca (2009), “Fazer uma pergunta, hacer una pregunta, poser une question, eine Frage stellen, to ask a question, fare una domanda – colocações em línguas europeias: a ilusão da semelhança?”. Volume de Homenagem a Maria Manuela Gouveia Delille, Vol. II, Instituto de Estudos Alemães, Faculdade de Letras, Universidade de Coimbra.

Bahns, Jens (1993), “Lexical collocations: a contrastive view”, *ELT Journal*. Volume 47/1, 56-63.

Bahns, Jens/Eldaw, M. (1993), “Should we teach EFL students collocations?” *System* 21 (1), 101-114.

Bally, Charles (1909), *Traite de stylistique française*. Paris.

Benson, Morton/Benson, Evelyn/Ilson, Robert (1986), *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*, Amsterdam / Philadelphia: John Benjamins.

Biskup, Danuta (1992), “L1 influence on learners’ renderings of English collocations. A Polish/German empirical study”, in: Arnaud, P.J.L. e Béjoint, H. (eds.), *Vocabulary and Applied Linguistics*, 85-93. London: Macmillan.

Blanco, Xavier (2006), “Significaciones gramaticales y sentidos colocacionales. ¿Más que una mera coincidencia?”, in: *Cadernos de Fraseoloxía Galega*, 8. Centro Ramón Piñeiro para a Investigación en Humanidades. Santiago de Compostela. Xunta de Galicia.

Branco, António/Silva, João (2004), “Evaluating solutions for the rapid development of state-of-the-art pos taggers for Portuguese”, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*. European Language Resources Association (ELRA), Lisbon, Portugal

Burger, Harald/Häcki Buhofer, Annelies/Gréciano, Gertrud (Hrsg.) (2003), *Flut von Texten – Vielfalt der Kulturen. Ascona 2001 zur Methodologie und Kulturspezifität der Phraseologie*, Baltmannsweiler: Schneider Verlag Hohengehren.

- Burger, Harald (²2003), *Phraseologie. Eine Einführung am Beispiel des Deutschen*. 2., überarbeitete Auflage, Berlin: Erich Schmidt Verlag.
- Burger, Harald (2004), "Phraseologie – Kräuter und Rüben? Traditionen und Perspektiven der Forschung", in: Steyer, Kathrin (Hrsg.), 19-40.
- Cardoso, Carla (2012), *Aprendizagem de vocabulário na aula de L2 – das palavras isoladas às colocações*. Faculdade de Letras da Universidade do Porto.
- Carlini Roberto/Codina-Filba Joan/Wanner Leo (2014), "Improving Collocation Correction by Ranking Suggestions Using Linguistic Knowledge", *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*, Uppsala, Sweden.
- Chang, Yu-Chia/Chang, Jason/Chen, Hao-Jan/Liou, Hsien-Chin (2008), "Automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology", *Computer Assisted Language Learning*, 21(3), 283–299.
- Chiang, David (2007), "Hierarchical Phrase-Based Translation", *Computational Linguistics*, 33(2) June, 201–228
- Clas, André (1994), "Collocations et langues de spécialité", in: *Meta*, XXXIX, 4, 576-579.
- Corbin, Danielle (1997), "Locutions, composés, unités polylexématiques: lexicalisation et mode de construction", in: *La locution entre langue et usages*, ed. Michel Martins-Baltar, 53-101.
- Corder, S. Pit (1967), "The significance of learners' errors", in: Richards, Jack (ed.) (1974), *Error Analysis. Perspectives on Second Language Acquisition*, New York: Longman, 19-27.
- Corder, S. Pit (1974), "Error Analysis", in: Allen J. P. B. and Pit Corder (1974, editors). *Techniques in Applied Linguistics* (The Edinburgh Course in Applied Linguistics) London: Oxford University Press.
- Corpas Pastor, Gloria (1996), *Manual de fraseología española*, Madrid: Gredos (= Biblioteca Románica Hispánica).
- Corpas Pastor, Glória (2001), "En torno al concepto de colocación", *EUSKERA - XLVI*, I, 89- 108.
- Coseriu, Eugeniu (1977), *Principios de Semántica Estructural*. Madrid: Gredos.

- Costa, Angela/Luís, Tiago/Ribeiro João/Mendes Ana/Coheur, Luísa (2012), "An English-Portuguese parallel corpus of questions: translation guidelines and application in SMT", *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12)*, European Language Resources Association (ELRA), May 2012.
- Costa, Angela/Coheur, Luísa/Lino, Teresa (2015), "Challenges on the translation of collocations", in: *Workshop on Multi-word Units in Machine Translation and Translation Technology*, Malaga, Spain, June 2015.
- Costa, Angela/Coheur, Luísa (2016), "Towards a Statistical-Enriched Corpus Containing Portuguese Collocations in Use: Reviewing Possible Extraction Tools", in: *12th International Conference, Computational Processing of the Portuguese Language (PROPOR)*, Lecture Notes in Computer Science, vol. 9727, Tomar, Portugal, July 2016.
- Costa, Angela/Ling, Wang/Luís, Tiago/Correia, Rui/ Coheur, Luísa (2015), "A linguistically motivated taxonomy for Machine Translation error analysis", *Machine Translation*, Springer, vol. 29, 127-161
- Costa, Angela/Coheur, Luísa/Lino, Teresa (2016), "A study on the production of collocations by European Portuguese learners", *Proceedings of the 12th Workshop on Multiword Expressions (MWE' 2016)*, Association for Computational Linguistics (ACL), Berlin, Germany, August 2016.
- Cowie, Anthony. P. (1981), "The treatment of collocations and idioms in learners' dictionaries" *Applied Linguistics* 2 (3), 223–235.
- Cowie, Anthony P. (1986), "Collocational Dictionaries – A comparative view", in: Murphy (ed.) *Fourth Joint Anglo-Soviet Seminar*, London: British Council, 61-69.
- Cowie, Anthony P. (1991), "Multiword Units in Newspaper Language", in: Granger, S (ed.), *Perspectives on the English Lexicon: A tribute to Jaques van Roey*, 101-116. CILL 17, 1-3.
- Cowie, Anthony P. (1992), "Multiword Lexical Units and Communicative Language Teaching", in: Arnaud, P./Béjoint, H. (eds.), *Vocabulary and Applied Linguistics*, London: Macmillan, 1-12.
- Cowie, Anthony P. (ed.) (2001[1998]), *Phraseology: Theory, Analysis and Applications*, Oxford: Clarendon Press.
- Crowther, Jonathan/Dignen, Sheila/ Lea, Diana (eds.), (2002), *Oxford Collocations Dictionary for Students of English*, Oxford: Oxford University Press.

- Cruse, David Alan (1986). *Lexical Semantics*. Cambridge, U.K.: Cambridge University Press.
- Cruse, David Alan (2000), *Meaning in Language. An Introduction to Semantics and Pragmatics*, Oxford: Oxford University Press.
- Cruse, David Alan/Hundsnurscher, Franz/Job, Michael/Lutzeier, Peter Rolf (Hrsg.) (2002), *Lexikologie/Lexikology. Ein international Handbuch zur Natur und Struktur von Wörtern und Wortschätzen*, 1. Halbband/Volume 1, Berlin/New York: Walter de Gruyter (= Handbücher zur Sprach- und Kommunikationswissenschaft; Bd. 21).
- Daelemans, Walter/Zavrel, Jakub/Berck, Peter/Gillis, Steven (1996), "Mbt: A memory-based part of speech tagger-generator", *Proceedings of fourth workshop on very large corpora*. ACL SIGDAT (1996), 14-27.
- Dagneaux, Estelle/Denness, Sharon/Granger, Sylviane (1998), "Computer-aided error analysis" *System*, 26, 163–174.
- Dechamps, Christina (2013), "Les Collocation de la Langue Juridique Française: Problématiques de l'enseignement/apprentissage à des apprenants lusophones" Tese de Doutoramento. Universidade Nova de Lisboa – FCSH.
- Ellis, Rod (1994), *The Study of Second Language Acquisition*, Oxford: Oxford University Press.
- Evert, Stefan (2005), *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.
- Feilke, Helmuth (1998), "Idiomatische Prägung", in: Barz, Irmhild/Öhlschläger, Günther (Hrsg.), *Zwischen Grammatik und Lexikon*, Tübingen: Max Niemeyer Verlag (= Linguistische Arbeiten, 390), 69-80.
- Fernandez Beatriz González/Schmitt, Norbert (2015), "How much collocation knowledge do L2 learners have", *International Journal of Applied Linguistics* 166 (1), Johns Benjamins Publishing Company, 194 - 126.
- Firth, J.R. (1957) *Papers in linguistics 1934–51*. Oxford University Press.
- Fleischer, Wolfgang (²1997), *Phraseologie der deutschen Gegenwartssprache*, Tübingen: Max Niemeyer Verlag.
- Fontenelle, Thierry (1992^a), "Collocation acquisition from a corpus or from a

dictionary: a comparison", *Proceedings Euralex-92*, 221-228.

Fontenelle, Thierry (1994), "What on the earth are collocations: An assessment of the ways in which certain words co-occur and others do not", *English Today* 10(4), 42–48.

Futagi, Yoko/Deane, Paul/Chodorow, Martin/Tetreault. Joel (2008), "A Computational Approach to Detecting Collocation Errors in the Writing of Non-Native Speakers of English", *Computer Assisted Language Learning* ,21(1), 353–367.

Gabrys-Biskup, Danuta (1992), "L1 influence on learners' renderings of English collocations: a Polish/German empirical study", in: P. J. L. Arnaud and H. Bějont (eds.): *Vocabulary and Applied Linguistics*. Basingstoke: Macmillan. 85-93

Gilquin, Gaëtanelle/Granger, Sylviane/Paquot, Magali (2007), "Learner corpora: The missing link in EAP pedagogy", in: Thompson. P. (ed.) *Corpus-based EAP Pedagogy*. Special issue of Journal of English for Academic Purposes, 6, (4), 319-335

Gitsaki, Christina (1999), *Second language lexical acquisition: A study of the development of collocational knowledge*, San Francisco, CA: International Scholars Publications.

Granger, Sylviane (2001 [1998]), "Prefabricated patterns in advanced EFL writing: collocations and formulae. In A.P. Cowie (ed.) *Phraseology: theory, analysis and applications*, Oxford: Oxford University Press, 145-160.

Granger, Sylviane (2007), "Corpus d'apprenants, annotation d'erreurs et ALAO: une synergie prometteuse", *Cahiers de lexicologie*, 91/2, 465-480.

Gries, Stefan T. (2008), "Phraseology and linguistic theory: A brief survey", in: Granger, Sylviane/Meunier, Fanny (eds.), *Phraseology: an interdisciplinary perspective*. Amsterdam, the Netherlands: John Benjamins.

Granger, Sylviane/Paquot Magali (2008), "Disentangling the phraseological web", in: Granger, Sylviane/Meunier, Fanny (eds.), *Phraseology: an interdisciplinary perspective*. Amsterdam, the Netherlands: John Benjamins.

Gross, Maurice (1981), "Les bases empiriques de la notion de prédicat sémantique", *Langages*, 63, 7-51.

Grossmann, Francis/Tutin, Agnès (2002), "Collocations régulières et irrégulières: esquisse de typologie du phénomène collocatif" , *Revue Française de*

- Halliday, M.A.K./Hasan, Ruqaiya (1976), *Cohesion in English*. London: Longman.
- Hashemi, Homa B./Hwa, Rebecca, (2014) "A Comparison of MT Errors and ESL Errors". *9th Language Resources and Evaluation Conference (LREC)*
- Hasselgren, A. (1994) "Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary", *International Journal of Applied Linguistics* 4/2: 237-2
- Hausmann, Franz Josef (1979), "Un dictionnaire des collocations est-il possible?", in: *Travaux de linguistique et de littérature* 17/1, 187-195.
- Hausmann, Franz Josef (1984), "Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen", *Praxis des neusprachlichen Unterrichts* 31, 395-406.
- Hausmann, Franz Josef (1985), "Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels", in: Bergenholtz, Henning./Mugdan, Joachim (Hrsg.), *Lexikographie und Grammatik*, Tübingen: Max Niemeyer Verlag, 118-129.
- Hausmann, Franz Josef (1989), "Le dictionnaire de collocations" (Artikel 95) in: Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (1989), *Wörterbücher – Dictionaries – Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Erster Teilband, Berlin/New York: Walter de Gruyter, 1010-1019.
- Hausmann, Franz Josef (1998), "O dicionario de colocacións. Criterio de organización", in: Ferro Ruibal, Xesús (ed.), *Actas do 1º Coloquio Galego de Fraseoloxía*, Vigo: Xunta de Galicia, 63–81.
- Hausmann, Franz Josef (2004), "Was sind eigentlich Kollokationen?", in: Steyer, Kathrin (Hrsg.), 309-334.
- Hausmann, Franz Josef (2007), "Lexicographie française et phraséologie", in *Collocations, phraséologie, lexicographie*. Aachen, Shaker Verlag, 121-153.
- Heid, Ulrich (1994), "On Ways Words Work Together -- Topics in Lexical Combinatorics" in: Willy Martin *et al.*, editor, *Proceedings of the VIth Euralex International Congress*, Amsterdam eingeladener Hauptvortrag, 226-257.
- Heid, Ulrich (1998), "Towards a corpus-based dictionary of German noun-verb

- collocations", *Proceedings of the Euralex International Congress 1998*, Liège, 301-312.
- Heid, Ulrich/Gouws, Rufus H. (2006), "A model for a multifunctional electronic dictionary of collocations", *Proceedings of the XIIth Euralex International Congress*, Euralex Torino, 979-988.
- Herbst, Thomas (1996), "What are collocations: sandy beaches or false teeth?", *English Studies* 77, 379-393.
- Higueras García, Marta (2004), "Claves prácticas para la enseñanza del léxico", in: *Carabela* 56. Madrid. SGEL, 5-25.
- Hill, Jimmie/Lewis, Michael (eds.) (1997), *LTP Dictionary of Selected Collocations*, Londres: LTP.
- Hill, Jimmie (2000), "Revising priorities: From grammatical failure to collocational success", in: M. Lewis (Ed.), *Teaching collocation. further developments in the lexical approach* (pp. 47-69). Hove: LTP.
- Howarth, Peter (1993), "A Phraseology Approach to Academic Writing", in: G. Blue (ed.), *Language Learning and Success: Studying through English*. London: Macmillan, 58-60.
- Howarth, Peter (1998b), "Phraseology and second language proficiency", *Applied Linguistics* 19(1), 24-44.
- Howarth, Peter (2001 [1998]), "The phraseology of learners 'academic writing", in: Anthony P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press, 161-186.
- Iriarte Sanromán, Álvaro (2001), *A Unidade Lexicográfica. Palavras, Colocações, Frasesmas, Pragmatemas*. Braga: Centro de Estudos Humanísticos-Universidade do Minho.
- Irsula Peña, Jesús (1994), *Substantiv-Verb-Kollokationen: kontrastive Untersuchungen Deutsch-Spanisch*, Frankfurt am Main/Berlin/Bern/New York/Paris/Wien: Peter Lang.
- Jakobson, Roman/Halle, Moris (1956), *Fundamentals of Language*, The Hague: Mouton.
- James, Carl (1998), *Errors in Language Learning and Use*. London and New York: Longman.

- Jones, S./Sinclair, J. McH. (1974), "English Lexical Collocations", *Cahiers de Lexicologie* 24, 15–61.
- Kilgarriff, Adam/Rychly, Pavel/Smrz, Pavel/Tugwell, David (2004), "The sketch engine", *Proceedings of EURALEX*.
- Koehn, Philipp (2005), "Europarl: A Parallel Corpus for Statistical Machine Translation", *Conference Proceedings: the tenth Machine Translation Summit, AAMT*, 79-86.
- Leiria, Isabel (2006), *Léxico, Aquisição e Ensino do Português Europeu língua não materna*, Lisboa: Fundação Calouste Gulbenkian/ Fundação para a Ciência e a Tecnologia.
- Leiria, Isabel (2004), "Português língua segunda e língua estrangeira: investigação e ensino", in: *Idiomático. Revista Digital de Didáctica de PLN*, nº 3. Centro Virtual Camões.
- Lewis, Michael (ed.) (2000), *Teaching Collocation. Further Developments in the Lexical Approach*, LTP, London.
- Liu, Liu-er (2002), *A corpus-based lexical semantic investigation of verb-noun miscollocations in Taiwan learners' English*. Master's thesis, Tamkang University, Taipei.
- Lozano, Cristóbal (2009), "CEDEL2: Corpus Escrito del Español L2", in: Callejas, Bretones/ Carmen, M. et al. (eds.) *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería, 197-212.
- Lüdeling, Anke/Walter, Maik/Kroymann, Emil/Adolphs, Peter (2005), "Multi-level error annotation in learner corpora", *Proceedings of the Corpus Linguistics Conference*. Birmingham, United Kingdom, 14-17.
- Makkai, Adam (1972), *Idiom Structure in English*, The Hague, Mouton.
- Martelli, Aurelia (2006), "A corpus-based description of English lexical collocations used by Italian advanced learners", in: Corino, Elisa/ Marelllo, Carla/Onesti, Cristina (eds.), *Proceedings XII EURALEX International Congress*, Torino, Italia, September 6th–9th 2006. Alessandria: Edizioni Dell'Orso, 1005–1012.
- Mel'čuk, Igor (1992), "Paraphrase et lexique: la Théorie Sens-Texte et le *Dictionnaire explicatif et combinatoire*", in: Mel'chuk et al. (1992), 9-58.

- Mel'čuk, Igor/Arbatchewsky-Jumarie, Nadia/Iordanskaja, Leo/Mantha, Suzanne (1992), *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III*, Montréal: Les Presses de l'Université de Montréal.
- Mel'čuk, Igor (1995), "Phrasemes in Language and Phraseology in Linguistics", in: Everaert (ed.) (1995), 167-232.
- Mel'čuk, Igor. (1997) *Vers une linguistique Sens-Texte*. Leçon inaugurale (10 de Janeiro de 1997), Collège de France, Chaire internationale.
- Mel'čuk, Igor (2001 [1998]), "Collocations and Lexical Functions", in: Cowie, Anthony P. (ed.), 23-53.
- Mel'čuk, Igor A. (2003), "Collocations: définition, rôle et utilité", in: Grossmann, Francis/Tutin, Agnès (eds.), *Les Collocations: analyse et traitement*, De Werelt, Amsterdam, 23–31.
- Mel'čuk, Igor/Polguère, Alian (2007): *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*, collection "Champs linguistiques", Louvain-la-Neuve: De Boeck Duculot.
- McEnery, Tony/Xiao, Richard (2006), *Corpus-based language studies: An advanced resource book*. Taylor & Francis.
- Muñiz, Eva-Maria (2004), *El concepto de 'colocación' en español*, Tesis de Doctorado, Santiago de Compostela, Universidad de Santiago de Compostela.
- Nesselhauf, Nadja (2003), "The use of collocations by advanced learners of English and some implications for teaching", *Applied Linguistics*, 24(2), 223-242.
- Nesselhauf, Nadja (2005). *Collocations in a learner corpus*. Amsterdam: Benjamins.
- Palapanidi, Kiriaki/María Pilar, Llach (2014), "Can lexical errors inform about word class acquisition in fl?: evidence from greek learners of spanish as a foreign language", *Revista de Lingüística y Lenguas Aplicadas*, 9(1): 67–78.
- Pawley, Andrew/Syder, Frances Hodgetts (1983) "Two puzzles for linguistic theory: nativelike selection and nativelike fluency", in: Richards, Jack/Schmidt, Richard (eds.) *Language and Communication*, London, Longman.

- Philip, Gill (2007), "Decomposition and delexicalisation in learners' collocational (mis)behaviour", in: Davies, Matthew/Rayson, Paul/Hunston, Susan/Danielsson, Pernilla (eds.), *Online Proceedings of Corpus Linguistics 2007*, Birmingham, United Kingdom, July 27th-30th 2007. Birmingham: University of Birmingham: 1-11.
- Pöll, Bernhard (1996), *Portugiesische Kollokationen im Wörterbuch: Ein Beitrag zur Lexikographie und Metalexikographie*, Bonn: Romanistischer Verlag.
- Prieto, Sabela/Mosqueira, Estela/Vázquez, Nancy (2009), "Córpora y enseñanza de lenguas: se buscan colocaciones", Cantos Gómez, Pascual/Sánchez Pérez, Aquilino (eds.): *A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus*. Murcia: AELINCO, 366-373.
- Rio-Torto, Graça (2006), "O Léxico: semântica e gramática das unidades lexicais", in: Athayde, Maria Francisca (ed.), 11-34.
- Raposo, Eduardo Bozaglo Paiva/do Nascimento, Maria Fernanda Bacelar/da Mota, Maria Antónia Coelho/Segura, Luísa/Mendes, Amália/Vicente, Graça/Veloso, Rita (eds.) (2013) *Gramática do Português I e II*. Gulbenkian
- Rodríguez-Fernández, Sara/Carlini Roberto/Wanner, Leo (2015), "Classification of Grammatical Collocation Errors in the Writings of Learners of Spanish", *Procesamiento del Lenguaje Natural* (55), 49-56.
- Rodríguez-Fernández, Sara/Carlini, Roberto,/Wanner, Leo (2015), "Classification of Lexical Collocation Errors in the Writings of Learners of Spanish", *Recent Advances in Natural Language Processing*, Hissar, Bulgaria
- Sanromán, Álvaro Iriarte (2001), *A unidade lexicográfica. Palavras, colocações, frasesmas, pragmatemas*, Braga, Universidade do Minho, Centro de Estudos Humanísticos.
- Sanromán Vilas, Begona (2006a), "El papel de la información semántica en el aprendizaje de las colocaciones léxicas", in: Corino, Elisa/Marello, Carla/Onesti, Cristina (eds.), *Proceedings XII Euralex International Congress*, Alessandria: Edizioni dell'Orso, 1037-1044.
- Sanromán Vilas, Begona (2009), "Towards a semantically oriented selection of the values of Oper1. The case of golpe 'blow' in Spanish", in: Beck, David *et alii* (eds.), *Proceedings of the Fourth International Conference on Meaning-Text Theory*. Montréal: OLST, 327-337.
- Santos, Diana, Rocha, Paulo (2001), "Evaluating CETEMPúblico, a Free Resource for Portuguese", *Proceedings of the 39th Annual Meeting on Association for*

- Computational Linguistics*, Association for Computational Linguistics, 450-457.
- Schindler, Wolfgang (2002), "Lexik, Lexikon, Wortschatz: Probleme der Abgrenzung", in: Cruse, David Alan/Hundsnurscher, Franz/Job, Michael/Lutzeier, Peter Rolf (Hrsg.), 31-44.
- Shei, Chiang/Pain, Helen (2000), "An ESL writer's collocational aid", *Computer Assisted Language Learning*, 13(2), 167–182.
- Sinclair, John/Jones, Susan/Daley, Robert/Krishnamurthy, Ramesh (2004 [1970]), *English Collocation Studies: The OSTI Report*. Continuum Books, London and New York.
- Sinclair, John (1991), *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.
- Steyer, Kathrin (Hrsg.) (2004), Wortverbindungen – mehr oder weniger fest, Institut für Deutsche Sprache, Jahrbuch 2003, Berlin/New York: Walter de Gruyter.
- Smadja, Frank/McKeown, Kathleen (1991), "Using collocations for language generation", *Computational Intelligence*, 7(4). Special issue on natural language generation.
- Thewissen, Jennifer (2008), "The phraseological errors of French-, German- and Spanish-speaking EFL learners: evidence from an error-tagged learner corpus", *Proceedings from the 8th Teaching and Language Corpora Conference (TaLC8)*. Lisbon: Associação de Estudos e de Investigação Científica do ISLA-Lisboa, 300-306.
- Tono, Yukio (2003), "Learner corpora: Design, development and applications", in: Archer, Dawn *et al.* (eds.), *Proceedings of the Corpus Linguistics 2003*, University Centre for Computer Corpus Research on Language, Lancaster, United Kingdom, March 28th–31th 2003, 323–343.
- Van den Bosch, Antal/Daelemans, Walter (1999), "Memory-based morphological analysis", *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 285-292
- Vincze, Orsolya/Alonso Ramos, Margarita/Mosqueira Suárez, Estela/Prieto González, Sabela (2011), "Exploiting a learner corpus for the development of a CALL environment for learning Spanish collocations", in: Kosem, Iztok/Kosem, Karmen (eds.), *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, 280-285.

- Wanner, Leo/Alonso Ramos, Margarita/Vincze, Orsolya/Nazar, R., Ferraro, G., Mosqueira, Estela/Prieto, Sabela (2011), *"Annotation of Collocations in a Learner Corpus for Building a Learning Environment"*, Learner Corpus Research conference 2011.
- Williams, Geoffrey (1998), "Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles", *International Journal of Corpus Linguistics* 3 (1), 151–171.
- Wittgenstein, Ludwig (1958), *Philosophical Investigations*. Trad. G. E. M. Anscombe. Oxford: Basil Blackwell.
- Zuluaga, Alberto (1980), *Introducción al estudio de las expresiones fijas*, Frankfurt am Main/Bern/Cirencester/U.K.: Peter Lang (= Studia Romanica et Linguistica, 10).
- Zuluaga, Alberto (2002), "Los "enlaces frecuentes" de María Moliner: Observaciones sobre las llamadas colocaciones", in: *PhiN* (Philologie im Netz), 22/2002. Herausgegeben von Paul Gévaudan, Hiltrud Lautenbach, Peter Schneck und Dietrich Scholler.

Listas

Lista de figuras

Figura 1 Classificação de combinação de palavras de Hausmann (1984).....	19
Figura 2 Classificação de combinação de palavras de Mel’cuk (1998)	21
Figura 3 Classificação de combinação de palavras de Cowie (2001[1998]).....	22
Figura 4 Classificação de combinação de palavras de Burger (² 2003).....	23
Figura 5 Resultados do Wortschatz para a palavra <i>momento</i>	96
Figura 6 Resultados do DeepDict para a palavra <i>momento</i>	97
Figura 7 Resultados do CRPC para o substantivo <i>momento</i>	98
Figura 8 Resultados do “word sketch” para a palavra <i>momento</i>	100
Figura 9 Resultados do “concordance” para a palavra <i>momento</i>	102
Figura 10 Resumo da categoria localização do erro.	127
Figura 11 Resumo da categoria descrição do erro.	133
Figura 12 Resumo da categoria explicação do erro.....	137

Lista de gráficos

Gráfico 1 Percentagem de colocações corretas e incorretas por nível.	145
Gráfico 2 Percentagem de colocações corretas e incorretas por nível e LM dos aprendentes.	146
Gráfico 3 Padrões sintáticos das colocações utilizadas pelos alunos.	147
Gráfico 4 Padrões sintáticos das colocações erradas usadas pelos alunos.	148

Gráfico 5 Padrões sintáticos das colocações corretas usadas pelos alunos.	149
Gráfico 6 Padrões sintáticos das colocações usados pelos nativos.	150
Gráfico 7 Localização dos erros na totalidade do <i>corpus</i>	161
Gráfico 8 Localização dos erros, tendo em conta o número de colocações, nos três níveis de proficiência.	162
Gráfico 9 Localização dos erros, tendo em conta o número de colocações, nas cinco LMs.	163
Gráfico 10 Descrição dos erros na totalidade do <i>corpus</i>	164
Gráfico 11 Descrição dos erros, tendo em conta o número de colocações, nos três níveis de proficiência.	167
Gráfico 12 Descrição dos erros, tendo em conta o número de colocações, nas cinco LMs.	169
Gráfico 13 Explicação dos erros na totalidade do <i>corpus</i>	169
Gráfico 14 Explicação dos erros, tendo em conta o número de colocações, nos três níveis de proficiência.	173
Gráfico 15 Explicação dos erros, tendo em conta o número de colocações, nas cinco LMs.	173
Gráfico 16 Proliferação dos nativos e não-nativos.	180
Gráfico 17 Proliferação e precisão tendo em consideração os três níveis de proficiência.	182
Gráfico 18 Proliferação e precisão tendo em consideração as cinco LMs dos alunos.	184
Gráfico 19 Proliferação e precisão tendo em consideração os anos de estudo de português dos alunos.	185
Gráfico 20 Proliferação e precisão tendo em consideração os dados de imersão dos alunos.	186
Gráfico 21 Proliferação e precisão tendo em consideração os dados sobre o uso da língua alvo fora do contexto escolar.	187
Gráfico 22 Proliferação e precisão tendo em consideração os dados sobre o nº de L2 que os alunos dominam.	189
Gráfico 23 Proliferação e precisão tendo em consideração a idade dos alunos.	190

Gráfico 24 Proliferação e precisão tendo em consideração a formação académica dos alunos.	191
Gráfico 25 Proporção de colocações corretas e incorretas.	193
Gráfico 26 Localização dos erros de colocações em MT.....	194
Gráfico 27 Descrição dos erros de tradução automática.	195

Lista de tabelas

Tabela 1 Exemplos de colocações em inglês e as suas traduções em alemão e francês.	36
Tabela 2 Número de alunos e número de textos por língua materna e nível usados no estudo.....	82
Tabela 3 Número de textos escritos por estímulo.....	83
Tabela 4 Número de alunos e número de textos por língua materna e nível usados no estudo.....	86
Tabela 5 Número de textos escritos por estímulo.....	87
Tabela 6 Número de alunos por língua materna.	88
Tabela 7 Tópicos das composições e número de composições.....	89
Tabela 8 Número de alunos por língua materna.	90
Tabela 9 Tópicos das composições e número de composições.....	90
Tabela 10 Número total de alunos, de textos e número de palavras.....	91
Tabela 11 Estímulos das composições produzidas pelos nativos.	92
Tabela 12 Comparação entre as extrações do Wortschatz (WoC), DeepDict (DD), CRPC, Word Sketch (SE-1) e Concordance (SE-2).	103
Tabela 13 Comparação entre as características dos quatro sistemas.	105
Tabela 14 Localização dos tipos de erros: substituição, criação, síntese, análise, sentido diferente, preposição e opção melhor.	133
Tabela 15 Relação entre a descrição dos erros e a sua explicação: importação, extensão, derivação, sobregeneralização, escolha errada, variedade e interlíngua e intralíngua.	137
Tabela 16 Número de alunos, palavras, textos, colocações corretas, incorretas e total de colocações por nível e LM.	144

Tabela 17 Os dez verbos mais frequentes e correspondente número de ocorrências	151
Tabela 18 Os dez nomes mais frequentes e correspondente número de ocorrências	154
Tabela 19 Verbos mais usados na construção de colocações pelos nativos.	155
Tabela 20 Substantivos mais usados na construção de colocações pelos nativos. ..	156
Tabela 21 As dez colocações mais frequentes no <i>corpus</i> dos aprendentes de português.	158
Tabela 22 Colocações mais usadas pelos nativos.	159
Tabela 23 Localização dos erros nos três níveis e nas cinco línguas maternas.	162
Tabela 24 Descrição dos erros nos três níveis e nas cinco línguas maternas.	166
Tabela 25 Explicação dos erros nos três níveis e nas cinco línguas maternas.	171
Tabela 26 Erros de transferência nos três níveis e nas cinco línguas maternas.	175
Tabela 27 Proliferação dos não-nativos e nativos.	179
Tabela 28 Precisão dos não-nativos e nativos.	181
Tabela 29 Informação disponível nos quatro <i>corpora</i> sobre os alunos.	182
Tabela 30 Proliferação dos não-nativos e nativos.	182
Tabela 31 Média e desvio padrão da proliferação e precisão para as cinco LMs.	183
Tabela 32 Média e desvio padrão da proliferação e precisão para os anos de aprendizagem.	185
Tabela 33 Média e desvio padrão da proliferação e precisão para a imersão ou não.	186
Tabela 34 Média e desvio padrão da proliferação e precisão para o uso ou não uso da L2.	187
Tabela 35 Média e desvio padrão da proliferação e precisão para o número de L2.	188
Tabela 36 Média e desvio padrão da proliferação e precisão para as idades dos alunos.	190
Tabela 37 Média e desvio padrão da proliferação e precisão para a formação académica dos alunos.	191

