

Collocation and Collocation Error Processing in the Context of Second Language Learning

Sara Rodríguez Fernández

TESI DOCTORAL UPF / 2018

Director de la tesi

Prof. Leo Wanner

Department of Information and Communication Technologies



By My Self and licensed under
Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported



You are free to Share – to copy, distribute and transmit the work Under the following conditions:

- **Attribution** – You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Noncommercial** – You may not use this work for commercial purposes.
- **No Derivative Works** – You may not alter, transform, or build upon this work.

With the understanding that:

Waiver – Any of the above conditions can be waived if you get permission from the copyright holder.

Public Domain – Where the work or any of its elements is in the public domain under applicable law, that status is in no way affected by the license.

Other Rights – In no way are any of the following rights affected by the license:

- Your fair dealing or fair use rights, or other applicable copyright exceptions and limitations;
- The author's moral rights;
- Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.

Notice – For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to this web page.

To my parents

—

—

|

|



Acknowledgements

First of all, I thank my PhD supervisor, Leo Wanner, for his kindness, patience and helpful guidance during all these years.

I am grateful to the members of the thesis committee, Eneko Agirre, Ulrich Heid and Alain Polguère, for their interest and their time for reading this thesis.

I am also thankful to the TALN members for their support. Special thanks must be given to Roberto Carlini, my faithful co-author, for the technical help, valuable advice and at-all-levels support that he has given to me during these years. I thank Luis Espinosa-Anke, also a co-author, for his work and enthusiasm, and Vanesa Vidal, for her help in the annotation process. Joan Codina, Joan Soler, Laura Pérez Mayos, Sergio Cajal, Francesco Barbieri and Miguel Ballesteros have also helped, in one way or another, in the mathematical/technical aspects of this thesis. Simon Mille has proven himself an inexhaustible source of motivation and good advice. To all of you, thank you.

Many thanks to my brother Pablo and to my friends, for being by my side during this period.

Finally, I thank my parents, to whom this thesis is dedicated, for their support, their raising me in the values of constancy and effort, their permanent help and advice, and for their love.

**This work has been funded by the Spanish Ministry of Economy and Competitiveness (MINECO), through a predoctoral grant (BES-2012-057036) in the framework of the project HARenES (FFI2011-30219-C02-02).*



Abstract

It is generally acknowledged that collocations in the sense of idiosyncratic word cooccurrences are a challenge in the context of second language learning. Learners often produce “ungrammatical” combinations such as **give a suggestion* or **make a walk*. Advanced computational tools able to aid L2 learners with collocations are thus highly desirable. However, state-of-the-art “collocation checkers” are merely able to detect a possible miscollocation and offer as correction suggestion a list of collocations of the given keyword whose semantics is often ignored. In order to address these shortcomings we propose techniques for collocation retrieval and semantic classification that retrieve, for a given base and the intended meaning, the actual collocate lexeme(s), and techniques for collocation error detection and classification. Given the small size of our learner corpora, we also provide an algorithm for the generation of an artificial collocation error corpus for Spanish.

Keywords: *Collocation, collocation errors, collocation (error) detection, collocation (error) classification, computer assisted language learning, artificial corpus generation*

Resumen

Suele admitirse que las colocaciones en el sentido de coocurrencias idiosincráticas de palabras son un reto en el aprendizaje de lenguas. Los estudiantes producen frecuentemente combinaciones “agramaticales” como **dar una sugerencia* o *hacer un paseo*. Herramientas computacionales avanzadas de ayuda al aprendizaje de colocaciones serían altamente deseables. Sin embargo, los correctores actuales solo detectan posibles errores y ofrecen como correcciones listas de colocaciones de la base cuya semántica suele ser ignorada. Para abordar estas limitaciones proponemos técnicas de extracción y clasificación semántica de colocaciones, que devuelven el(los) colocativo(s) para una base y significado dados y técnicas de detección y clasificación de errores colocacionales. Dado el pequeño tamaño de nuestro corpus de aprendices, también se proponen técnicas para generar un corpus artificial de errores colocacionales para el español.

Palabras clave: *Colocaciones, errores colocacionales, detección de colocaciones (y errores), clasificación de colocaciones (y errores), aprendizaje de lenguas asistido por ordenador, generación de corpus artificial*

Contents

Abstract	vii
Resumen	viii
List of Figures	xii
List of Tables	xiii
1 Introduction	1
1.1 Collocations and Foreign Language Learning	1
1.2 Computational aids to collocation learning	4
1.3 Our proposal	6
1.4 Outline of the thesis	6
2 Background on collocations	7
2.1 On the nature of collocations	7
2.1.1 The statistical approach: The Firthian tradition	8
2.1.2 The lexicological approach	9
2.1.3 Conclusion	10
2.2 Collocation typologies	15
2.2.1 Lexical Functions: a tool for the description of collocations	16
2.2.2 Conclusion	18
2.3 Collocation error typology	21
2.3.1 1 st dimension: Location of the error	22
2.3.2 2 nd dimension: Description of the error	22

2.3.3	3 rd dimension: Causes of the error	26
2.3.4	Conclusion	29
3	State of the art	31
3.1	Collocation extraction	32
3.1.1	The onset of collocation extraction	33
3.1.2	A step forward in collocation extraction	34
3.1.3	Conclusions	36
3.2	Semantic classification of collocations	36
3.2.1	Lexico-semantic resources as source of semantic information	37
3.2.2	Contextual features as source of semantic information	38
3.2.3	Conclusions	40
3.3	Collocation error detection and correction	40
3.3.1	Detection of collocation errors	42
3.3.2	Suggestion and ranking of potential corrections	45
3.3.3	Machine learning algorithms for collocation error correction	50
3.3.4	Conclusions	51
3.4	Grammatical error detection and correction	52
3.4.1	Determiner and preposition errors	53
3.4.2	Gender, number and word order errors	58
3.5	Artificial corpora for error detection/correction	60
3.5.1	Grammatical error generation	61
3.5.2	Lexical error generation	65
3.5.3	Conclusions	66
4	Semantics-driven recognition of collocations	67
4.1	Word embeddings based techniques	67
4.2	Example-based approach	69
4.2.1	Exploiting the analogy property	69
4.2.2	Experimental setup	71
4.2.3	Outcome of the experiments	72
4.2.4	Discussion	73
4.3	Weakly supervised approach	76
4.3.1	Methodology for the acquisition of collocation resources	77
4.3.2	Experimental setup	79
4.3.3	Outcome of the experiments	81
4.3.4	Discussion	82
4.4	Summary and conclusions	87

5 Collocation error classification	89
5.1 Collocation error classification	89
5.1.1 A hybrid approach to collocation error classification	90
5.1.2 Experimental setup	96
5.1.3 Outcome of the experiments	98
5.1.4 Discussion	99
5.2 Summary and conclusions	106
6 An artificial corpus for collocation error detection	107
6.1 Adapted collocation error typology	108
6.2 CEDEL2 Corpus analysis	110
6.3 Artificial corpus generation	111
6.3.1 General design	113
6.3.2 Error generators	114
6.3.3 Resources	119
6.4 Enriching the Spanish GigaWord with artificial errors	121
6.4.1 Collocation errors	121
6.4.2 Non-collocation errors	122
6.4.3 Sentence complexity	122
6.5 Evaluation of the generated corpus for collocation error de- tection	123
6.5.1 The collocation error marking model	123
6.5.2 Experimental setup	126
6.5.3 Outcome of the experiments	128
6.5.4 Discussion	129
6.6 Summary and conclusions	132
7 Conclusions and Future Work	135
7.1 Contributions of the thesis	135
7.2 Publications	136
7.3 Limitations of the Thesis	137
7.4 Future work	138
Bibliography	139
A Association Measures	155
B Artificial error examples	157

List of Figures

2.1	Types of syntagms (translated from Mel'čuk (2011))	12
2.2	Types of phrasemes (translated from Mel'čuk (2013))	13
2.3	Location dimension (from Alonso Ramos et al. (2010))	22
2.4	Descriptive dimension (from Alonso Ramos et al. (2010))	24
2.5	Explanatory dimension (from Alonso Ramos et al. (2010))	27
4.1	Examples of vector offsets for 'intense' collocations	70
4.2	Example of relations between bases and 'intense' collocates	78
6.1	Algorithm sequential reading	124
6.2	Example of the processing steps of the algorithm	125
6.3	Recurrent Neural Networks schema	126

List of Tables

2.1	LF / semantic gloss correspondence	19
4.1	Seed examples for each semantic gloss	72
4.2	Number of collocations found for each semantic gloss	73
4.3	Performance of the acquisition and classification of collocations with respect to semantic glosses	74
4.4	Examples of correctly retrieved collocates for each semantic gloss	75
4.5	Semantic glosses and size of training sets	80
4.6	Precision achieved by the system configurations (S3 – S4) and baselines (S1 –S2) tested on Spanish data	82
4.7	Mean Reciprocal Rank achieved by the system configurations (S3 – S4) and baselines (S1 – S2) tested on Spanish data	83
4.8	Precision achieved by the system configurations (S3 – S6) and baselines (S1 – S2) tested on English data	83
4.9	Mean Reciprocal Rank achieved by the system configurations (S3 – S6) and baselines (S1 – S2) tested on English data	84
4.10	Precision of the coarse-grained evaluation of the English S6 con- figuration	84
4.11	Examples of retrieved Spanish collocations	85
4.12	Examples of retrieved English collocations	86
5.1	Features for the classification of ‘Extended Substitution’ errors. .	92
5.2	Number of instances of the lexical error types and correct collo- cations in CEDEL2.	97
5.3	Number of instances of the grammatical error types in CEDEL2.	97
5.4	Accuracy of the lexical error detection systems.	98

5.5	Accuracy of the grammatical error detection functions	99
6.1	Lexical collocation error typology underlying our work	108
6.2	Grammatical collocation error typology underlying our work	109
6.3	Frequency of collocation errors in CEDEL2	110
6.4	Multiple error types	111
6.5	Confusion set GoCI	112
6.6	Confusion set GoCD	112
6.7	Confusion set GoBD	112
6.8	Confusion set GoCS	112
6.9	Confusion set GoBS	113
6.10	Possible error types for each collocation pattern	115
6.11	Error types created by each error generator	116
6.12	Non-collocation errors in the CEDEL2 sample	120
6.13	Syntactic complexity features in the GigaWord and CEDEL2 samples	123
6.14	Number of collocation errors in training, development and test corpora	127
6.15	Number of multiple collocation errors in artificial training, de- velopment and test corpora	127
6.16	Performance of the system when trained and evaluated on arti- ficial data	128
6.17	Performance of the system when trained on artificial data and evaluated on the CEDEL2 corpus	129
6.18	Number of collocation errors in artificial + CEDEL2 corpus	131
6.19	Number of multiple collocation errors in artificial + CEDEL2 corpus	131
6.20	Performance of the system when trained on artificial data en- riched with CEDEL2 data and evaluated on the remaining CEDEL2 corpus	131
6.21	Performance of the system when trained on artificial data and evaluated on 40% of the CEDEL2 corpus	132
A.1	Frequently used Association Measures	155
A.2	“Improved” Association Measures	156

Introduction

The present PhD thesis focuses on the topic of collocations and *Computer Assisted Language Learning (CALL)*, and particularly, on the development of techniques for automatic detection and classification of collocations and collocation errors and automatic generation of collocational resources that can help non-native speakers in their learning and using of collocations in Spanish.

1.1 Collocations and Foreign Language Learning

It is commonly accepted nowadays that language learning cannot be fully accomplished through the learning of words and grammar only. Vocabulary and grammar are necessary but they are far from being sufficient. Take, as an example, the expression *give [a] walk*. Although constructed with correct English words and following the syntactic rules of the language, any native speaker would find that there is something “wrong” with it because in English *walks* are not *given*, but *taken*. This type of “wrongness” is produced when the lexical restriction “rules” are broken. In other words, lexical elements cannot always be freely combined to form bigger units. In order to properly learn a language, it is also necessary to know the lexical restrictions that apply to the lexical elements of the language in question.

“Collocations” are phenomena in which these restrictions are brought forward. Collocations are combinations of two lexical items between which a direct syntactic dependency holds, where one of the elements (the *base*) is freely chosen by the speaker, but the occurrence of the other (the *collocate*) is restricted by the base. In recent years, there has been an increasing in-

terest in the study of collocations. Their importance in language is well acknowledged today, and the degree to which they are mastered is generally considered as a mark of native-like fluency and language command. As has been stressed by various scholars, “language learning is collocation learning” (Hausmann, 1984) and “language knowledge is collocational knowledge” (Nation, 2001). Research has consistently shown that collocations are a serious problem for language learners, irrespective of their native language (L1) (Bahns and Eldaw, 1993; Nesselhauf, 2003; Laufer and Waldman, 2011). The reason behind this fact is that, unlike grammar, which is regular and easier to master, collocations are idiosyncratic combinations of words, and thus more demanding.

Since it is not enough to know the individual words, but also the lexical restrictions that operate on them, collocational knowledge lags behind vocabulary knowledge, rather than developing alongside it (Bahns and Eldaw, 1993). It is also for this reason that the challenge of collocation learning lies in production, rather than in comprehension (Nesselhauf, 2003). In particular, the learner’s discourse in a foreign language (L2) is affected by three main types of problems: (1) avoidance of collocations, (2) disproportionate use of collocations, and (3) miscollocations, or collocation errors.

Regarding the avoidance of collocations, Farghal and Obiedat (1995) report that learners consciously abstain from using collocations by repeatedly resorting to strategies such as paraphrasing. Besides disrupting the discourse flow, paraphrasing collocations is not always simple, and can easily lead to the production of complex sentences which are prone to be erroneous.

As for the density of collocations in learner writing, research has led to differing conclusions. For instance, Howarth (1996) and Laufer and Waldman (2011) report that learners underproduce collocations with respect to native speakers, while Siyanova and Schmitt (2008) and Orol-González and Alonso Ramos (2013) found that the proportion of collocations in texts written by native and non-native speakers was similar. Other studies reconcile both views showing that learners tend to overuse some collocations, relying on familiarity and general-purpose collocates, and to underuse more idiosyncratic ones. For instance, Granger (1998) found that general-purpose amplifiers such as *totally*, *completely* or *very* were overused, while stereotyped co-occurrences such as *acutely aware* or *painfully clear* were underused. Frequent verbs like *have*, *take* or *get* have also been found to be overused with regard to other idiosyncratic collocations used by natives (Wang and Shaw, 2008; Nesselhauf, 2005). These findings have been con-

firmed by Orol-González and Alonso Ramos (2013). The results of their study, in which the collocational richness of learners and native speakers is compared, show that the collocations used by learners lack the variety and sophistication of those by native speakers.

An even more serious issue is the high number of collocation errors or atypical word combinations produced by L2 learners, like the word pairs **take* [a] *decision* or **pay effort*. Nesselhauf (2005) reports that a third of the verb-noun collocations found in an English L2 corpus by German speakers were erroneous and Alonso Ramos et al. (2010) found that 39% of the collocations in a corpus of Spanish L2 were incorrect. Similar findings have been reported by Laufer and Waldman (2011), who, interestingly, also discovered that the amount of collocation errors increases along with the proficiency level. This could be explained by the fact that advanced students, who are usually more confident, do not depend so much on familiar collocations and are more willing to be creative, generating a higher number of errors in the process.

Regarding their causes, collocation errors can be prompted by both interlingual and intralingual factors. Nesselhauf (2005) found that about half of the errors were influenced by L1, and Laufer and Waldman (2011) report that most of the errors were caused for the same reason. According to Yamashita and Jiang (2010)'s study, learners make significantly more errors than native speakers even on collocations in which the involved lexical items are the same in L2 as in L1. This supports the idea that learners not always rely on their L1 when producing collocations or, in other words, that transfer does not always happen, even if the two languages are closely related. Intralingual factors are as important as interlingual ones. They account for errors such as over-generalization, that is the use of highly polysemous collocates (e.g. the verbs *make* or *get*), or the avoidance of collocations that seem a direct translation from L1.

As for their types, collocation errors can concern more than the incorrect choice of the lexical items. For instance, Nesselhauf (2003) noticed that learners make other types of errors, such as the incorrect use of the prepositions or determiners involved in the collocations, the usage of which is also usually restricted. To capture all types of collocation errors, Alonso Ramos et al. (2010) suggest a fine-grained typology that considers three parallel dimensions. The first dimension, the "location" dimension, captures which element is affected by the error (or whether the error concerns the collocation as a whole). The "descriptive" dimension models the nature of the

error, distinguishing between register, lexical and grammatical collocation errors. Register errors capture context-inappropriate use of *per se* correct collocations. Lexical errors capture a mistake with respect to one of the collocation elements (either wrong word or creation of a non-existing word) or the collocation as a whole (creation of an artificial single word instead of a collocation, creation of an artificial collocation, or use of a collocation with a different sense than intended). Grammatical errors concern the grammar of collocations (missing or superfluous determiner, wrong preposition, wrong subcategorization, etc.). Finally, an “explanatory” dimension models the cause of the errors, that is, whether they are caused by interlingual or intralingual reasons.

Given that collocations represent such a great challenge to L2 learners, it seems necessary to take some action in order to improve their collocational knowledge. The following measures have been suggested:

- **Explicit teaching** (Bahns and Eldaw, 1993; Nesselhauf, 2003; Sadeghi, 2009). If, because of their transparency, collocations remain unnoticed by learners, an obvious first step would be to make the learners aware of their difficulty and the importance of paying attention to the restrictions that apply to lexical elements.
- **Exposure and repetition** (Schmitt, 2008; Webb et al., 2013). Given that the likelihood of coming across a collocation, as compared to its respective lexical elements, is much smaller, even though learners may notice the restriction rules, they could easily forget them if the same collocation is not seen within a certain period of time. For this reason, it is important to regularly provide access to examples of usage.
- **Autonomous learning.** Ying and O’Neill (2009) and Boers and Lindstromberg (2009) defend that, for a better acquisition of collocations, learners should be involved in their own learning process, being forced to reflect on it and seek the strategies that suit them best.

A computational tool able to meet these needs would be, without any doubt, desirable.

1.2 Computational aids to collocation learning

Computational tools that focus on collocations are more recent and scarce than tools addressing other types of errors, such as spelling and grammar.

This is due to the challenges in computational collocation processing, such as collocation extraction or collocation error detection. However, recent research in the field has already led to significant results in collocation identification. This, along with the growing interest in the topic, has resulted in an increased number of tools that were developed during the last years (Chang et al., 2008; Park et al., 2008; Wu et al., 2010; Wanner et al., 2013b). Among the functionalities of these tools are the verification of the correctness of isolated collocations, the search of collocates for a given base, the retrieval of examples of correct usage of a given collocation, the detection of collocation errors in texts, and the suggestion of lists of possible corrections for incorrect collocations. Thus, they focus on lexical restrictions, provide opportunity for exposure and repetition and favour autonomous learning. Nevertheless, despite the evident advance that these tools represent for language learners, there is still space for improvement, since they still present several limitations:

- They do not consider the heterogeneity of collocation errors, focusing exclusively on lexical errors and leaving the grammatical ones aside, on the one hand, and not taking into account the different types of lexical errors, on the other. Besides offering additional information to the learner, the identification of the specific types of errors could lead to more accurate type-targeted error correction techniques.
- They do not provide examples of incorrect usage. It is not enough for the student to have access to native collocation-tagged corpora in order to learn the correct usage of collocations. It is also crucial to confront them with the variety of collocation errors they or other learners make, typify these errors, and indicate the corrections.
- They offer, as correction suggestions (or as the result of the collocate search), lists of words that tend to co-occur with the keyword, usually ordered by frequency and POS-pattern, but whose semantics is ignored. The usefulness of such lists is limited in the sense that the learners are likely not to know the meaning of all of these co-occurrences (i.e., collocations). It would be desirable to group the resulting collocates according to their meanings.

1.3 Our proposal

The present thesis attempts to meet the needs of learners of Spanish and propose a solution to the limitations identified above with respect to the existing collocation identification/correction tools in the following ways:

- Development of computational techniques for the automatic retrieval and semantic classification of collocations, according to a typology based on *Lexical Functions* (LFs) (Mel'čuk, 1996).
- Development of computational techniques for the automatic identification and classification of collocation errors according to a fine-grained typology (Alonso Ramos et al., 2010), which considers lexical and grammatical errors and their different subtypes.
- Design and implementation, in view of the lack of annotated learner corpora, of an algorithm for the automatic creation of an artificial collocation learner corpus, as done in the context of automatic grammar error detection and correction research (Foster and Andersen, 2009; Rozovskaya and Roth, 2010b; Felice and Yuan, 2014).

1.4 Outline of the thesis


The remainder of the thesis is organized as follows.

Chapter 2 gives an overview of the theoretical aspects of collocations, and the semantic and collocation error typologies that are used in the experiments.

In Chapter 3, we introduce the current state of the art in collocation detection and semantic classification, grammatical and collocation error detection and correction, and artificial error generation.

Chapters 4 – 6 represent the experimental section of the thesis. First, Chapter 4 presents the experiments on retrieval and semantic classification of collocations. Then, Chapter 5 describes our work regarding collocation error classification. In Chapter 6, we describe the design, implementation and use of an artificial collocation error corpus for collocation error detection.

Finally, Chapter 7 summarizes the research carried out in the thesis, states its limitations and suggests open lines of future work.



Background on collocations

As stated in the Introduction, the present thesis focuses on collocations in the context of second language learning. In particular, it focuses on the generation of collocational resources and on the automatic detection and classification of collocations and collocation errors. In this chapter, we present the theoretical background related to these topics. Specifically, in Section 2.1, we give an overview of the different concepts of the term “collocation” as understood by the statistical and lexicological schools, paying special attention to the definition and characterization of collocations as considered in the *Explanatory and Combinatorial Lexicology* (ECL) (Mel’čuk, 1995), which we adopt in our work. Section 2.2 presents current collocation typologies, and Section 2.3 describes the only collocation error typology that we are aware of.

2.1 On the nature of collocations

The term “collocation” as introduced by Firth (1957) and cast into a definition by Halliday (1961) encompasses the statistical distribution of lexical items in context: lexical items that form high probability associations are considered collocations. However, in contemporary lexicography and lexicology, an interpretation that stresses the idiosyncratic nature of collocations prevails. According to Hausmann (1984), Cowie (1994), Mel’čuk (1995) and others, a collocation is a binary idiosyncratic co-occurrence of lexical items between which a direct syntactic dependency holds and where the occurrence of one of the items (the *base*) is subject of the free choice of the speaker, while the occurrence of the other item (the *collocate*) is

restricted by the base. Some examples of collocations are: *take [a] walk*, where *walk* is the base and *take* the collocate, or *high speed*, where *speed* is the base and *high* the collocate.

In the remainder of this section we present an overview of the different definitions of the concept of collocation coming from both schools: statistical and lexicological. Given that many definitions have been proposed by a number of scholars, and it is out of the scope of this thesis to review all of them, we only focus on the most relevant and best-known ones.

2.1.1 The statistical approach: The Firthian tradition

According to Firth, part of the meaning of a word is given by its relation to its context, or its tendency to co-occur with other words. The term “collocation” is used to account for this relation or association between words:

Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of night is its collocability with dark, and of dark, of course, collocation with night. (Firth, 1957)

Associations between words can be more or less stereotyped and, in Firth’s view, it is by taking into account the co-occurrence of a word how the acceptability of a word combination can be determined. This is to say that collocations are simply understood as frequent co-occurrences. As an example, he offers collocations of the word *time*, with words *saved*, *spent*, *wasted*, *frittered away*, *presses*, *flies* and even the negation particle *no*.

In an attempt to introduce formal criteria for the study of collocations, and given that grammar can not properly describe the lexical relations in language, Halliday proposes a new linguistic level “the lexis” (Halliday, 1966), in which collocations are still described in terms of frequencies and probabilities, and defined as “*the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item x, the items a, b, c ...*”

Sinclair (1966) introduces the terms of *node* (the word under study), *span* (the number of lexical items on both sides of the node that are considered relevant), and *collocates* (the items in the environment of the node, set by the span). Similar to Halliday, Sinclair also defends the study of collocations in statistical terms. In his own words, a collocation is *the occurrence of two items in a context within a specified environment. Significant collocation is*

regular collocation between two items, such that they co-occur more often than their respective frequencies and the length of the text in which they appear would predict (Sinclair et al., 1970). He proposes a method for determining whether a combination is a collocation or not, which takes several factors into account: the frequency of the combination, the frequencies of each of its components and the distance between them, and which applies significance tests to find *significant* collocations (Sinclair et al., 2004).

2.1.2 The lexicological approach

Collocations in the statistical sense are textual collocations, where only co-occurrence criteria are used to distinguish them from other types of word co-occurrences. In lexicology and lexical semantics, however, syntactic and semantic criteria are preferred, and collocations are generally seen as combinations of syntactically bound and lexically restricted elements that cannot be freely combined.

Cruse (1986) uses the term “collocation” to designate only a subset of the sequences of lexical items that habitually co-occur, i.e., those that are fully transparent in the sense that each lexical constituent is also a semantic constituent, like *fine weather*, *torrential rain*, *light drizzle* or *high winds*. Idioms are excluded from his definition, but he mentions a special type of collocations, *bound* collocations, such as *foot the bill* or *curry favour*, whose constituents habitually co-occur and display some of the characteristic properties of idioms.

Cruse observes that collocations have a kind of semantic cohesion – the constituent elements are, to varying degrees, selective. He distinguishes three types of collocational restrictions to explain the inappropriateness of a combination: systematic, semi-systematic and idiosyncratic. Systematic restrictions can be explained by means of semantics. For instance, while it is possible to say *toast the bread*, **grill the bread* seems less appropriate. It is likely that the reason behind this restriction is that *grill* is used for cooking raw food, *toast* being preferred for cooked food. Semi-systematic restrictions can be only partially explained by means of semantics. For instance, while, in exchange for money, a *customer* typically acquires something material, and a *client* some professional or technical service, people that use the services of a bank can be still called its *customers*. Finally, idiosyncratic restrictions are totally arbitrary, like *immaculate kitchen* and *impeccable taste*, compared to, for example, **immaculate taste*.

Hausmann understands collocations as restricted combinations in which a lexical item cannot be freely combined with any other lexical item in order to express a meaning. He defines them as “restricted and oriented combinations”. Components of collocations are not mutually selective, rather, the selection is oriented from one element to the other. He introduces the term “base” to refer to the stable, autonomous element in a collocation, or the element that *selects*, and the term “collocator” or “collocate” for the element that *is selected* (Hausmann, 1985).

Benson (1989) generalizes the notion of collocation as “arbitrary and recurrent word combinations”. In the *BBI Combinatory Dictionary of English: A Guide to Word Combination* (Benson et al., 2010), collocations are defined as “recurrent, fixed, identifiable, nonidiomatic phrases and constructions”, and the frequency criterion is used to consider a certain combination as collocation or not, and include it in the dictionary, if it is. A distinction is made between lexical and grammatical collocations, each being divided into several classes. Lexical collocations are those in which only nouns, adjectives, verbs and adverbs intervene, while grammatical collocations are made up of a dominant word, such as a noun, adjective or adverb, combined with a preposition or a grammatical structure like *to* + infinitive, *that* + clause, etc.

Within the framework of the Explanatory and Combinatorial Lexicology (ECL) (Mel’čuk, 1995), the lexical component of the *Meaning–Text Theory* (Mel’čuk, 1973, 1997; Polguère, 1998; Kahane, 2003), Mel’čuk (1995) introduced a theory of phrasemes, which has been developed since then (Mel’čuk, 1998, 2003, 2013). Essentially, his idea of collocation is similar to that of Hausmann, which is to say that collocations are syntactically bound binary expressions in which one of the elements selects the other. In his theory, however, a systematic definition of the concept of collocation is proposed, along with its characterization with regards to both free expressions and other types of set expressions, or phrasemes, which are also presented and classified according to their two defining features: restriction and compositionality.

2.1.3 Conclusion

The definition of collocations in terms of frequency presents a practical advantage over the lexicological approach: even though the choice of some of the features, such as the span of the collocation, or the exact meaning of “frequently co-occurring” might be somewhat arbitrary, still, frequen-

cies and probabilities can be objectively quantified. It is for this reason, together with the fact that frequencies and probabilities can be easily computed automatically, that the statistical view of collocations has been widely adopted in many NLP applications. However, this view also presents several problems both from the theoretical and applied perspectives.

One of the main limitations of the statistical approach is the fact that it does not capture the reality of the appearance of a collocation in context: the components of a collocation do not necessarily appear together, but can appear sufficiently wide apart in the text, so that they fall out of the scope of a short span.

Besides, frequent co-occurrence of lexical items does not necessarily mean the existence of a collocation. For instance, the combination *red car*, whose elements, frequent as they can be, are free of any type of lexical selection restrictions, is not a collocation but a free (although frequent) word combination. Even though the criterion of frequency is relevant to language teaching, i.e., the most frequent collocations should be taught first, what truly matters from a pedagogical perspective is the problem of restricted lexical co-occurrence.

For this reason, in our work, we have adopted the lexicological concept of collocation, in particular, as it is understood in the ECL (Mel'čuk, 1995). We present below this concept in some detail.

Collocations in Explanatory and Combinatorial Lexicology

Collocations are a subclass of what is known as *phrasemes*, or *set phrases*, and are characterized in terms of their differences with regard to free expressions and other types of phrasemes. Informally, phrasemes can be defined as non-free expressions, or syntagms that cannot be constructed according to the general rules of language. For a formal definition of what means to be a free or non-free phrase, the concepts of *restriction* and *regularity* (*compositionality*) are introduced (Mel'čuk, 1995; Mel'čuk, 1998), since it is through those two properties that free phrases and phrasemes are characterized. For a phrase to be free, it needs to be both unrestrictedly constructed, and compositional. If any of these properties is broken, the phrase is said to be *set*, *fixed*, *non-free*.

The property of restriction belongs to the paradigmatic axis and refers to the selection of the lexical items. An expression **AB** is unrestricted if and only if (1) it is selected because of its meaning, independently of the ex-

tralinguistic situation, and (2) **A** is selected by its meaning independently of **B**, and vice versa. For instance, in *buy [a] bicycle*, both *buy* and *bicycle* are freely selected for their meaning, but the selection of *take* in *take [a] step* is dependent on *step*, and thus restricted.

The property of compositionality belongs to the syntagmatic axis and refers to *regularity*, or meaning and expression combination rules. Therefore, an expression **AB** is compositional if and only if it is constructed from **A** and **B** according to the general combination rules of the language, such as [*a*] *hot meal*, whose meaning is the sum of the meanings of its components. The idiom [*a*] *hot potato*, however, does not mean ‘a potato that has been heated’, but rather it refers to an (often conflictive) issue which many people are talking about.

A theory of phrasemes

When crossing the properties of restriction and compositionality, would they be independent, four different types of expressions would be obtained. Since they are not independent (unrestricted expressions are always compositional), three types of expressions are obtained: free expressions, and two major types of phrasemes, i.e., compositional phrasemes and non-compositional phrasemes (see Figure 2.1).

Restriction types \ Compositionality	non-compositional	compositional
	lexical	idioms (1)
semantic-lexical	impossible	clichés (3)

Figure 2.1: Types of syntagms (translated from Mel’čuk (2011))

A further distinction can be made when considering restriction types. According to the nature of the restriction, phrasemes can be divided into lexical and semantic-lexical phrasemes.

In a lexical phraseme, the sense is freely constructed by the speaker, independently of the extralinguistic situation, while the selection of the lexical items is restricted. In a semantic-lexical phraseme, the sense is not freely constructed by the speaker, but selected depending on the situation. Usually, the choice of the lexical items is also restricted.

The product of these two types of phrasemes with compositional and non-compositional phrasemes gives as result four main types of phrasemes. Since semantic-lexical non-compositional phrasemes cannot appear in a language¹, we are left with three main types of phrasemes: idioms, collocations and clichés (see Figure 2.2).

Types of syntagms Properties	free expressions (1)	impossible combination	non-free expressions = phrasemes	
			pragmatemes, clichés & collocations (2)	idioms (3)
unrestriction	+	+	-	-
compositionality	+	-	+	-

Figure 2.2: Types of phrasemes (translated from Mel'čuk (2013))

Idioms, such as *miss the boat* 'to miss a chance' or *piece of cake* 'a task or activity that is simple or easy', clichés, such as *time will tell*, or *better late than never* and pragmatemes, such as *no parking*, fall out of the scope of this thesis and will therefore not be presented here. The concept of collocation is explained below.

Collocations

Collocations are compositional lexical phrasemes, such as *genuine affection*, *pay attention*, or *devise [a] strategy*. The phraseological restriction applies only to one of its components (the collocate), the other (base) is freely

¹Mel'čuk (2013) provides the following reason as to the impossibility of semantic-lexical non-compositional phrasemes: if a phraseme is non-compositional, it has, by definition, a sense associated to it as a whole; that sense is, therefore, not constructed by the speaker for the particular situation, and for this reason it is not possible to talk about the restricted character of its construction.

chosen. For instance, in the example from above, *devise [a] strategy*, in order to express the meaning ‘cause to come into existence, create’, the speaker might choose verbs such as *devise*, *develop* or *design*, but not *build* or *construct* or *make*. If, however, the base was *plan*, *make* could also be a valid option.

In ECL, collocations are described uniquely in terms of restriction and compositionality. Other criteria, such as idiosyncrasy, semantic transparency and fixation, which have been traditionally proposed as defining criteria for the description of collocations, are seen as properties related to the concept, rather than defining elements.

Collocations, as any other type of phrasemes, are idiosyncratic phenomena. This means that restriction takes place at the lexical, rather than the semantic level or, in other words, that restriction “rules” are arbitrary, a matter of use. However, the degree of idiosyncrasy varies from collocation to collocation (cf., e.g., *high price* and *paramount importance*).

Regarding transparency, firstly, it is necessary to distinguish it from compositionality. While semantic compositionality is an objective property that can be measured by summing up the individual meanings of the different components of an expression, transparency is a psychological characteristic that depends on the knowledge or ability of the speaker to capture the meaning of the expression. A collocation such as *great pride* is both compositional and transparent. On the contrary, *heavy smoker*, although compositional, since the collocate *heavy*, in the context of the collocation, expresses the meaning ‘intense’, rather than its typical meaning in isolation ‘of great weight’, can be considerably more opaque to the speaker. As can be seen from the examples above, not all collocations possess the same degree of transparency.

Finally, the degree of fixation is also logically independent of phraseology. For instance, even though *[to] pay attention [to N]* is more fixed than *[to] turn one’s attention [to N]*², both are correct English collocations because the base *attention* determines the selection of the collocates *pay* and *turn* respectively.

²To calculate the degree of “fixation” of both expressions, we calculated *Pointwise Mutual Information* (a measure in the statistical paradigm that measures the “collocationality” of a word co-occurrence) on the British National Corpus. We obtained 0.72 in the first case and 0 in the second, which means that *pay attention* is much more fixed than *[to] turn one’s attention [to N]*. Pointwise Mutual Information and other *association measures* are introduced in Chapter 3

2.2 Collocation typologies

Following Hausmann (1989), Heid (1994, 1996) classifies collocations according to the category of their components. N–V collocations are further subclassified based on the grammatical function of the noun. The resulting classes are the following:

- noun + adjective; cf., e.g., *confirmed bachelor*;
- noun + verb (Subj); cf., e.g., *his anger falls*;
- noun + verb (Obj); cf., e.g., *to withdraw money*;
- noun + noun; cf., e.g., *a gust of anger*;
- verb + adverb; cf., e.g., *it is raining heavily*;
- adjective + adverb; cf., e.g., *seriously injured*.

Collocation dictionaries such as the *Oxford Collocations Dictionary* or the *Macmillan Collocations Dictionary* classify collocations in terms of semantic categories such that language learners can find more easily the collocate that communicates the meaning they intend to express. However, as far as we know, the semantic typologies used in both cases are not explicit to the user.

The authors of the *BBI*, on the contrary, propose an explicit typology of lexical collocations that takes into account both the POS-tags of their components and their semantics. Seven classes are defined:

- Verb that means ‘creation’ or ‘activation’ and noun or pronoun; cf., e.g., *start [a] fight*;
- Verb that means ‘eradication’ and/or ‘nullification’ and noun; cf., e.g., *annul [a] law*;
- Adjective and noun; cf., e.g., *strong opinion*;
- Noun and verb that means ‘typical action of the person or thing designed by the noun’; cf., e.g., *wind blows*;
- Noun and noun, one of them designating the unity, and the other, the set; cf., e.g., *[a] bouquet [of] flowers*;

- Adverb and adjective; cf., e.g., *sorely wounded*;
- verb and adverb; cf., e.g., *eat hungrily*.

A more comprehensive typology is proposed within ECL. This typology is based on the *Lexical Functions* (LFs) system, which provides a formal, systematic and rigorous classification of collocations according to their morpho-syntactic and semantic features. In what follows, we present a description of LFs as understood in ECL.

2.2.1 Lexical Functions: a tool for the description of collocations

According to Mel'čuk (1995), a *Lexical Function* (LF) **f** is “a function that associates with a given lexical unit *L*, which is the argument, or keyword, of **f**, a set L_i of (more or less) synonymous lexical items - the value³ of **f**”. In other words, an LF is a meaning that can be lexically expressed in different ways depending on the particular keyword the LF applies to. For instance, the LF **Magn**, which represents the meaning of ‘intensity’ (see Subsection 2.2.1.3 for the introduction of the names of the LFs) can be expressed by means of the values *big*, *sonorous* or *stentorian*, for the keyword *voice*, or through the values *deafening*, *loud* or *thunderous* for the keyword *applause*. Another example of LFs is **Real**, whose meaning is ‘to do with *keyword* what is expected to be done with it, to realize’. For instance, a *movie* is *watched*, a *book* is *read*, a *language* is *spoken*, and so on. The LF **Oper**, ‘to execute the action of *keyword*, to perform’ accounts for *Support Verb Collocations* of the kind of *[to] do sports*, *[to] make [a] call*, *[to] take care*, etc. Each LF corresponds to a unique sense that describes a relation between the components of the collocation.

2.2.1.1 Simple and complex LFs

Simple LFs, like the ones exemplified above, correspond to “single” meanings, including, e.g., ‘perform’, ‘cause’, ‘realize’, ‘terminate’, ‘intense’ and ‘positive’. In total, around 60 simple LFs are distinguished. These simple LFs can be combined to *Complex* LFs. For instance, the combination of the simple LFs **Anti**, which accounts for the basic sense of negation, and

³The *keyword* corresponds to the *base* of the collocation, and the *value* corresponds to the *collocate*

Magn, gives birth to the complex LF **AntiMagn**, with the complex meaning of ‘not intense’; see Kahane and Polguère (2001) for the mathematical apparatus of the combination of LFs.

2.2.1.2 Standard and non-standard LFs

A distinction is also made between *standard* and *non-standard* LFs. For an LF to be standard, it must fulfil the two following conditions:

- (1) The sense that the LF expresses must be general enough so that it can be applied to a considerably high number of arguments.
- (2) The number of possible values associated to the function must be relatively big.

All other functions are considered non-standard. The difference, then, between standard and non-standard LFs is purely quantitative. **Magn** and **Real** are standard LFs, while the LF that expresses the relation between *year* and *leap* (example taken from Mel’čuk (1996)) is exclusive for this collocation and, therefore, non-standard. By means of non-standard LFs it is possible to describe all the collocations of a language, although building a typology that takes them into account would be unrealistic.

2.2.1.3 Representation of standard LFs

LFs are represented by a name that signals the semantics of the collocation. For the sake of brevity, each type is labelled by a Latin acronym: ‘perform’ \equiv “Oper(are)”, ‘realize’ \equiv “Real(is)”, ‘intense’ \equiv “Magn(us)”, etc.

When applicable, subscripts are added that refer to the deep syntactic actants of the keyword to which the LF in question is applied. For instance, for the keyword *invitation*, whose actantial structure could be defined as ‘**X** gives a present **Y** to **Z**’, the first actant is **X**, who gives the present; the second actant, **Y**, is the present itself, and the third actant, **Z** is the person to whom the present is given. Thus:

Oper₁(present) = *give* [ART \sim]⁴

Oper₃(present) = *receive* [ART \sim]

⁴ \sim stands for the keyword

For the sake of further illustration, let us consider another example, the keyword *punishment*. In this case, ‘**X** (actant 1) administers the punishment **Y** (actant 2) to **Z** (actant 3) for a reason **W** (actant 4)’. Thus:

Oper₁(punishment) = *impose* [ART ~]

Oper₃(punishment) = *suffer* [ART ~]

2.2.2 Conclusion

The semantic typologies adopted in current collocation dictionaries present certain limitations, such as not covering all collocations or being unsystematic. For instance, the BBI Dictionary presents an explicit and systematic typology of lexical collocations, but only four semantic classes are defined (for adjective-noun, adverb-adjective and verb-adverb collocations no semantic classes are considered at all). In other cases, the categorization (or classification) is not always homogeneous. For instance, in the Macmillan Collocations Dictionary, the entries for ADMIRATION and AFFINITY contain the categories ‘have’ and ‘show’, each with their own collocates, while for other headwords, such as, e.g., ABILITY, collocates with the meaning ‘have’ and ‘show’ are grouped under the same category; in the entry for ALARM, *cause* or *express* are not assigned to any category, while for other keywords the categories ‘cause’ and ‘show’ are used (see e.g., PROBLEM for ‘cause’ or ADMIRATION for ‘show’); and so on. On the other hand, in the case of some headwords, the categories are very fine-grained (cf., e.g., AMOUNT, which includes glosses like ‘very large’, ‘too large’, ‘rather large’, ‘at the limit’, etc.), while in the case of others, it is much more coarse-grained (cf., e.g., ANALOGY, for which collocates with different semantics are included under the same gloss, as, e.g., *appropriate*, *apt*, *close*, *exact*, *good*, *helpful*, *interesting*, *obvious*, *perfect*, *simple*, *useful* that all belong to the category ‘good’). This lack of uniformity may confuse learners, who will expect that collocates grouped together share similar semantic features.

As it is the case in collocation dictionaries, in computational lexicography, categories of different granularity have also been used for automatic classification of collocations from given lists; cf., e.g., Wanner et al. (2016), who use 16 categories for the classification of V–N collocations and 5 categories for the classification of Adj–N collocations; Moreno et al. (2013), who work with 5 broader categories for V–N collocations, or Huang et al. (2009), who also use very coarse-grained semantic categories of the type ‘goodness’, ‘heaviness’, ‘measures’, etc. But all of these categories have the disadvantage of being *ad hoc*.

LF	Semantic gloss
Magn	‘intense’
AntiMagn	‘weak’
Bon	‘good’
Oper ₁	‘perform’
IncepOper ₁	‘begin to perform’
FinOper ₁	‘stop performing’
CausPredPlus	‘increase’
CausPredMinus	‘decrease’
CausFunc ₀	‘create, cause’
LiquFunc ₀	‘put an end’
Manif	‘show’

Table 2.1: LF / semantic gloss correspondence

For all the above reasons, we follow a different approach. As already Wanner et al. (2006b), Gelbukh and Kolesnikova. (2012) and also Moreno et al. (2013) in their second run of experiments and the *Diccionario de Colocaciones del Español* (DiCE), we use the semantic typology of Lexical Functions (LFs) to classify collocations.

In our experiments, we use a subset of the most frequently used simple and complex LFs. For all of these LFs, we define semantic glosses similar to those used in the Macmillan Collocations Dictionary, in order to make the LFs more transparent to users (see Table 2.1). We present and illustrate these LFs with some examples below. In order to find out which were the most frequent LFs, we annotated the collocations of a Spanish corpus, the AnCora-UPF (Mille et al., 2013), which contains 3,513 sentences (100,892 tokens).

Magn (‘intense’):

Magn(<i>certeza</i> ‘certainty’)	=	{ <i>absoluta</i> ‘absolute’, <i>total</i> ‘total’}
Magn(<i>emoción</i> ‘emotion’)	=	{ <i>fuerte</i> ‘strong’, <i>profunda</i> ‘deep’}
Magn(<i>importancia</i> ‘importance’)	=	{ <i>suma</i> ‘utmost’, <i>capital</i> ‘capital’}
Magn(<i>riesgo</i> ‘risk’)	=	{ <i>alto</i> ‘high’, <i>elevado</i> ‘elevated’}

AntiMagn ('weak'):

AntiMagn(<i>temperatura</i> 'temperature')	=	{ <i>baja</i> 'low'}
AntiMagn(<i>herida</i> 'wound')	=	{ <i>leve</i> 'mild', <i>superficial</i> 'superficial'}
AntiMagn(<i>nivel</i> 'level')	=	{ <i>bajo</i> 'low', <i>reducido</i> 'reduced'}
AntiMagn(<i>riesgo</i> 'risk')	=	{ <i>ligero</i> 'slight', <i>pequeño</i> 'small'}

Bon ('good')

Bon(<i>actuación</i> 'performance')	=	{ <i>espléndida</i> 'splendid'}
Bon(<i>comida</i> 'meal')	=	{ <i>deliciosa</i> 'delicious'}
Bon(<i>ayuda</i> 'aid')	=	{ <i>inestimable</i> 'inestimable'}
Bon(<i>conducta</i> 'behaviour')	=	{ <i>intachable</i> 'irreproachable'}

Oper₁ ('perform'):⁵

Oper ₁ (<i>clase</i> 'lecture')	=	{ <i>dar</i> 'give'}
Oper ₁ (<i>búsqueda</i> 'search')	=	{ <i>hacer</i> 'do', <i>realizar</i> 'realize'}
Oper ₁ (<i>decisión</i> 'decision')	=	{ <i>tomar</i> 'take'}
Oper ₁ (<i>idea</i> 'idea')	=	{ <i>teher</i> 'have'}

IncepOper₁ ('begin to perform')

IncepOper ₁ (<i>fuego</i> 'fire')	=	{ <i>abrir</i> 'open'}
IncepOper ₁ (<i>aventura</i> 'adventure')	=	{ <i>embarcar</i> 'embark'}
IncepOper ₁ (<i>competición</i> 'competition')	=	{ <i>entrar</i> 'enter'}
IncepOper ₁ (<i>error</i> 'fault')	=	{ <i>incurrir</i> 'incurr'}

FinOper₁ ('stop performing')

FinOper ₁ (<i>poder</i> 'power')	=	{ <i>perder</i> 'lose'}
FinOper ₁ (<i>carrera</i> 'career')	=	{ <i>abandonar</i> 'abandon'}
FinOper ₁ (<i>trabajo</i> 'job')	=	{ <i>perder</i> 'lose'}
FinOper ₁ (<i>competición</i> 'competition')	=	{ <i>retirarse</i> 'leave'}

CausPredPlus ('increase')

CausPredPlus(<i>tráfico</i> 'traffic')	=	{ <i>aumentar</i> 'increase'}
CausPredPlus(<i>calidad</i> 'quality')	=	{ <i>mejorar</i> 'improve'}
CausPredPlus(<i>precio</i> 'price')	=	{ <i>elegir</i> 'elevate', <i>subir</i> 'raise'}
CausPredPlus(<i>coste</i> 'cost')	=	{ <i>aumentar</i> 'increase'}

⁵The index indicates the syntactic structure of the collocation. Due to the lack of space, we do not enter here in further details; see Mel'čuk (1996) for a detailed description.

CausPredMinus ('decrease')

CausPredMinus(<i>nivel</i> 'level')	=	{ <i>bajar</i> 'lower'}
CausPredMinus(<i>contaminación</i> 'pollution')	=	{ <i>disminuir</i> 'decrease'}
CausPredMinus(<i>temperatura</i> 'temperature')	=	{ <i>reducir</i> 'reduce'}
CausPredMinus(<i>coste</i> 'cost')	=	{ <i>reducir</i> 'reduce'}

CausFunc₀ ('create, cause')

CausFunc ₀ (<i>edificio</i> 'building')	=	{ <i>erigir</i> 'erect', <i>construir</i> 'build'}
CausFunc ₀ (<i>problema</i> 'problem')	=	{ <i>causar</i> 'cause'}
CausFunc ₀ (<i>marca</i> 'mark')	=	{ <i>dejar</i> 'leave'}
CausFunc ₀ (<i>dificultad</i> 'difficulty')	=	{ <i>generar</i> 'generate'}

LiquFunc₀ ('put an end')

LiquFunc ₀ (<i>agresión</i> 'aggression')	=	{ <i>parar</i> 'stop'}
LiquFunc ₀ (<i>asamblea</i> 'assembly')	=	{ <i>disolver</i> 'dissolve'}
LiquFunc ₀ (<i>problema</i> 'problem')	=	{ <i>resolver</i> 'solve'}
LiquFunc ₀ (<i>miedo</i> 'fear')	=	{ <i>vencer</i> 'defeat'}

Manif ('show')

Manif(<i>sorpresa</i> 'surprise')	=	{ <i>expresar</i> 'express'}
Manif(<i>tendencia</i> 'tendencia' 'tendency')	=	{ <i>manifestar</i> 'manifest'}
Manif(<i>ternura</i> 'sentimiento' 'feeling')	=	{ <i>mostrar</i> 'show'}
Manif(<i>emoción</i> 'emotion')	=	{ <i>mostrar</i> 'show'}

2.3 Collocation error typology

As far as we know, the only existing collocation error typology is the one proposed by Alonso Ramos et al. (2010). It was designed after carrying out the analysis of a Spanish learner corpus, the *Corpus Escrito del Español L2* (CEDEL2) (Lozano, 2009), and considers three parallel dimensions. The first dimension, or *Location* dimension, describes where the error is produced, i.e., which element of the collocation is affected, namely the base or the collocate, or whether the error affects the collocation as a whole. The second, *Descriptive*, dimension accounts for the kind of error that has been produced (register, lexical or grammatical). Finally, the third dimension, the *Explanatory* dimension, captures the possible reasons why an error is produced (intralingual or interlingual causes).

In what follows, we describe the typology in some detail and illustrate the different types of errors with examples. All the examples have been taken

from the CEDEL2 corpus⁶.

2.3.1 1st dimension: Location of the error

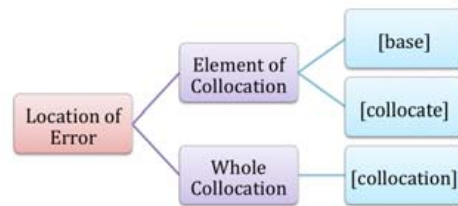


Figure 2.3: Location dimension (from Alonso Ramos et al. (2010))

As shown in Figure 2.3, a collocation error might occur in either the base or the collocate, or it might affect the collocation as a whole. For instance, a wrongly chosen base can be found in **hablar francia*, lit. ‘to speak France’, where the lexical item *francia* is chosen instead of *francés* ‘French’. In **tomar una siesta*, lit. ‘to take a nap’ the erroneous element is the collocate, since *tomar* ‘take’ is chosen instead of the appropriate verb *echar* ‘pour’. Finally, in **hombres sin casas*, lit. ‘men without homes’, the whole collocation is affected, since the intended meaning would be best expressed as *vagabundos* ‘vagabond’ or *sin techo* ‘homeless’.

2.3.2 2nd dimension: Description of the error

With respect to the second dimension, a first distinction is made that separates errors into (1) lexical errors, such as, e.g., **visión aclarada*, lit. ‘clear vision’, where the choice of the collocate is incorrect, (2) grammatical errors, like **tomar una avión*, lit. ‘to take a plane’, where the gender of the base is incorrect, as shown by the use of a determiner with wrong gender, and (3) register errors, such as **tener el deseo personal*, lit. ‘to have a personal wish’, which is a possible combination in Spanish, but not used in the context in which it was found. Lexical and grammatical errors are subdivided into five and eight categories, respectively, as described below and as shown in Figure 2.4.

⁶For simplicity, conjugated verbs are presented in their lemmatized form

2.3.2.1 Lexical errors

According to Alonso Ramos et al. (2010), lexical errors can be divided into the following types:

- **Substitution errors**, which are produced when an existing lexical unit is incorrectly selected as the base or collocate. This is the case of **realizar una meta* ‘to reach a goal’, lit. ‘to realize, to carry out a goal’, where both the base and the collocate are valid lexical units in Spanish, but the correct collocate *conseguir* ‘reach’ has been substituted by *realizar* ‘realize’.
- **Creation errors**, which are produced when a non-existing form is chosen as the base or collocate. An example of this type of error would be **serie televisual*, lit. ‘televisual series’, instead of *serie televisiva* ‘TV series’, where the learner has used the non-existing form *televisual*.
- **Synthesis errors**, in which a non-existing word is used where a collocation should be used instead. Such is the case of **bienvenir*, lit. ‘to welcome’, rather than *dar la bienvenida*, lit. ‘to give the welcome’.
- **Analysis errors**, which concern the opposite case, where a new expression with the form of a collocation is created instead of using a single word. An example of this error is **hacer nieve*, lit. ‘to make snow’, which in Spanish would be better expressed by the verb *nevar* ‘to snow’.
- **Different sense errors**, which are produced when correct collocations are used incorrectly, in the sense that they express a different meaning than the intended one. An example of this case is the use of **voz alta*, lit. ‘loud voice’, instead of *el gran voz*, lit. ‘great voice’. Both are correct collocations, but while the former refers to ‘loud voice’, as opposed to ‘silent voice’, *gran voz* refers to ‘shouting’, to ‘an extraordinarily loud voice’.

Substitution and creation errors capture a mistake that can affect either the base or the collocate. Synthesis, analysis and different sense errors affect the collocation as a whole.

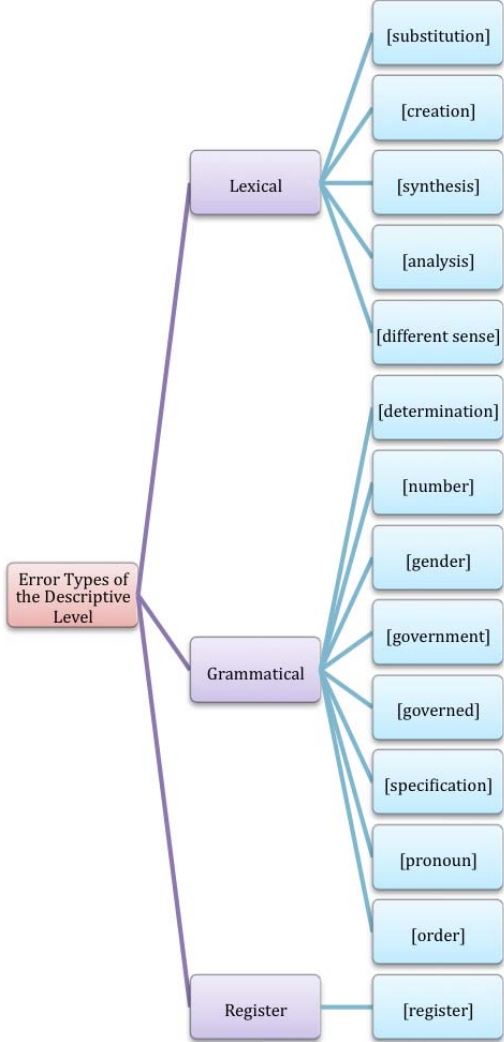


Figure 2.4: Descriptive dimension (from Alonso Ramos et al. (2010))

2.3.2.2 Grammatical errors

Grammatical errors are divided into eight different categories:

- **Determination errors**, which affect the base and are produced when either the determiner is required but is not present, or when it is not permitted but is used. In other words, a determination error is produced when a determiner is incorrectly deleted or inserted. The wrong choice of the determiner is not considered a collocation error, but rather a grammatical one. An example of a collocation affected by this type of error is **terminar escuela*, lit. ‘to finish school’, where a determiner is required in Spanish (*terminar la escuela*), but is missing in the example.
- **Number errors**, such as **estar en vacación*, lit. ‘to be on holiday’, which are produced when either the plural or the singular form of a lexical unit is required for a particular collocation, and the incorrect form is used. In this case the plural form would be correct (*estar en vacaciones*).
- **Gender errors**, which are produced when the incorrect gender of the base is chosen. This is usually manifested as a concordance error due to the incorrect choice of the determiner or adjectival collocates. For instance, in **pasar los vacaciones*, lit. ‘to spend the holidays’, the masculine form of the determiner, *los* instead of *las*, indicates that the base is considered to be masculine.
- **Government errors**, which are made when the preposition required by one of the members of the collocation is missing, mistakenly chosen or used when there should be no preposition. An example of this case is **ver a la película* ‘to watch a movie’, lit. ‘to watch at a movie’. In Spanish, the verb *ver* is followed by the preposition *a* when referring to people, but the use of the preposition is not permitted when referring to things.
- **Governed errors**, such as **estar en buen humor*, lit. ‘to be in a good mood’, which are produced when a wrong preposition is used as governor of the collocates, or a preposition is used where there should be none. In this case, the preposition *de* should have been used instead of *en* (*estar de buen humor*).

- **Specification errors**, which affect exclusively the base and are produced when a modifier is missing. This is the case of the miscollocation **hacer un aterrizaje*, lit. ‘to make a landing’, where the modifier *forzoso* ‘forced’ should be present.
- **Pronoun errors**, which describe the misuse of a reflexive pronoun for a verbal collocate. In **volver loco* ‘to go crazy’, lit. ‘to turn (someone) crazy’ the reflexive pronoun is missing. The correct collocation would be *volverse loco*, lit. ‘to turn (oneself) crazy’.
- **Order errors**, which are produced when the base and the collocate are written in the wrong order, as in **reputación mala*, lit. ‘reputation bad’, instead of *mala reputación* ‘bad reputation’. This type of error affects the collocation as a whole.

2.3.3 3rd dimension: Causes of the error

A first general distinction is made for the third dimension, the explanatory dimension, distinguishing between interlingual and intralingual errors. The former are produced due to the influence of the L1 of the student, while the latter are caused by a lack of knowledge of the L2. Lexical, grammatical and register errors can be explained by means of intralingual and interlingual causes. The combination of the two axes (descriptive and explanatory) gives birth to six main categories of errors: lexical interlingual and intralingual errors, grammatical interlingual and intralingual errors and register interlingual and intralingual errors.

Grammatical interlingual and intralingual errors are not further subclassified. An example of the former is **escuchar a la música*, lit. ‘to listen to the music’, instead of **escuchar música*, where the insertion of the preposition can be explained by interference with the learner’s L1. An example of the latter is **buen educación*, lit. ‘good education’, where the gender of the adjective has been incorrectly chosen so that there is no agreement between the noun and the adjective. The correct form is *buena educación*.

Register errors are not further subdivided either. Only intralingual register errors have been found in the corpus, such as **tener el deseo personal* ‘to have the personal wish’, a correct combination in itself, but incorrect in the learner’s context.

Both interlingual and intralingual lexical errors are divided into more detailed types. Figure 2.5 shows all the different types of errors of the third dimension.

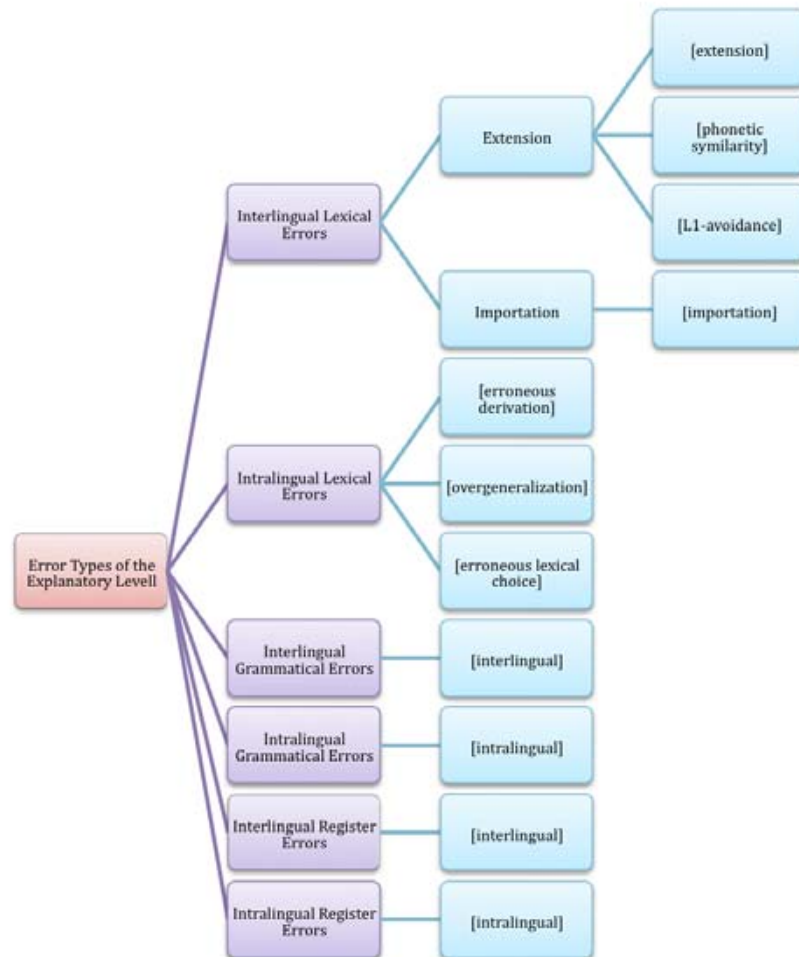


Figure 2.5: Explanatory dimension (from Alonso Ramos et al. (2010))

2.3.3.1 Interlingual lexical errors

Interlingual lexical errors are divided into extension errors and importation errors. Extension errors can be produced by three causes: (1) literal translation from the L1 form into L2, (2) because of the phonetic similarity with the form in the L1, and (3) as an attempt to avoid the use of a form similar to the L1 form. They are explained below:

- **Extension errors proper**, which occur when the meaning of existing words in L2 is extended due to L1 influence. For instance, in **tomar un examen*, lit. ‘to take an exam’, the collocate is literally translated and a wrong collocation is thus created. In this case, the right collocation is *hacer un examen* ‘to do an exam’.
- **Phonetic similarity errors**, which are produced when an incorrect form is chosen due to the phonetic similarity between the incorrectly chosen item and its equivalent in the L1, like in **lograr un gol* ‘to accomplish a goal’, lit. ‘to accomplish a goal (in a football context)’, instead of *lograr un objetivo*.
- **L1-avoidance errors**, which are produced when the student is trying to avoid what he considers to be a too close similarity between the items in the L1 and L2. This is the case of **acudir el teléfono* lit. ‘to turn to the telephone’ instead of *atender el teléfono* ‘to attend the telephone’.

Importation errors are made when a new word is created coming from the learner’s L1, sometimes adapted to Spanish, such as it occurs in **ir de hiking* ‘to go hiking’, lit. ‘to go of hiking’ instead of *ir de acampada*.

2.3.3.2 Intralingual lexical errors

Finally, intralingual lexical errors can be divided into three subtypes:

- **Erroneous derivation errors**, which are produced when the student creates a new word in the L2 inspired by common derivations of other forms of the L2. For example, **tiene limitades*, lit. ‘has limits’, instead of *límites*, where the non-existing form **limitades* has been incorrectly derived from the verb *limitar* ‘to limit’ by adding to the root the correct suffix *ad* (correctly followed by *es* to form the plural) following

the pattern of similar words such as *libertad* ‘freedom’, from *libertar* ‘to free’.

- **Overgeneralization errors**, which are made when a form that is too general is preferred for a collocation over the more appropriate specific lexical item. This is the case of **malos efectos*, lit. ‘bad effects’, instead of *efectos perjudiciales* o *efectos nocivos* ‘harmful, damaging effects’.
- **Erroneous lexical form errors**, which are produced when a wrong form is used for no clear reason. This is the case of **establecer un sufrimiento* ‘to cause suffering’, lit. ‘to establish a suffering’, instead of *provocar un sufrimiento*, lit. ‘to provoke a suffering’.

2.3.4 Conclusion

The typology presented in the previous section is, to the best of our knowledge, the first attempt at offering a detailed classification of collocation errors. It was created by analysing and systematically classifying the collocation errors found in a learner corpus, and it reflects all types of phenomena related to collocations, including the subcategorization frames of their components, and other grammatical aspects of collocations (often ignored in collocation tools) and the possible causes of the errors.

Since it was created from writings of L2 Spanish learners, it cannot be directly applied to languages other than Spanish, because some of the errors, such as *Pronoun* or *Gender* errors, for example, are language-dependent. Nevertheless, the methodology followed for its creation can be extended to other languages, thus setting an example for collocation error research in other languages.

In our work, we experiment with the first and second dimensions (*Locative* and *Descriptive*), and use the third dimension (*Explanatory*) as source of information for the automatic generation of collocation errors.

Collocation error classification

With respect to the classification of collocation errors, we focus on the second dimension, covering both lexical and grammatical errors. Even though some of the grammatical error types (e.g., the *Gender* errors and *Order* errors) can be considered a problem of a grammar checker rather than of a collocation checker, we address them in the context of collocation verification because they make a collocation to be incorrect.

Our studies show that *Substitution*, *Creation* and *Different sense* errors are the most common types of lexical errors. In contrast, learners tend to make rather few *Synthesis* and *Analysis* errors. Therefore, given that *Synthesis* errors are not comparable to any other error class, we decided not to consider them at this stage of our work. *Analysis* errors show in their appearance a high similarity to *Substitution* errors, such that both could be merged without any major distortion of the typology. For this reason, we deal with miscollocation classification with respect to three lexical error classes: *Extended substitution*, *Creation*, and *Different sense*.

Similarly, for grammatical errors we found that *Governed* and *Specification* errors are very seldom. We also opted not to consider them at this stage of the experiments.

Finally, only one case of *Register* error was found in the corpus, so we equally discarded this type of error in our work.

Simultaneous collocation error identification and classification

In our experiments with collocation error identification and classification, we merged the first and second dimensions in order to identify both the component of the collocation affected by the error, and the lexical and grammatical error that is produced. In this case, *Analysis* and *Different sense* errors were also left aside, mainly due to the lack of a clear methodology for their automatic creation.

State of the art

As seen in the Introduction, collocations are a highly important element in language learning, as well as one of the most problematic ones. Despite this, the development of collocation resources and computational tools that focus on collocations is recent, when compared to other computational aids such as spell, grammar or style checkers. This is due to the difficulty related to automatic collocation and collocation error detection. However, recent research in this area has led to significant advances on these topics. This, along with the growing interest in the topic, has resulted in an increased number of tools being developed during the last years. In this chapter, we present an overview of the most significant developments in (1) collocation identification, (2) collocation classification¹, and (3) collocation error correction. Given that existing collocation tools do not deal with the grammatical errors involved in collocations, we also present an overview of the research on grammatical error detection and correction. The chapter is organized as follows. Section 3.1 presents research in collocation extraction. Then, in Section 3.2 we describe the different works on semantic classification of collocations. Finally, Sections 3.3 and 3.4 deal with error detection and correction, both lexical and grammatical. In Section 3.5 an overview of the work on grammatical error detection and correction that uses artificial corpora is given.

¹Classification of collocations is necessary to provide the learners with semantically-motivated collocation resources, and to group collocation corrections according to their meaning.

3.1 Collocation extraction

The task of collocation extraction (and thus identification) from corpora typically involves comparing the distribution of words in combination and in isolation, in a text corpus, as a way to measure their lexical association. In Computational Linguistics, a variety of *Association Measures* (AMs), cf., Appendix A, have been proposed that stem from information theoretical metrics and statistical significance tests. These AMs are based on contingency tables that contain marginal and co-occurrence frequencies. The objective is to assign a score to a given word combination that represents its degree of “collocability”, or association strength. The higher the score, the stronger the association and the probability that it is a collocation.

As noticed by Wanner (2004), even though there is no consensus regarding the concept of “collocation”, the two main interpretations of the term (the distributional and the idiosyncratic one) do not necessarily differ on the judgement of a word combination to be a collocation or a free combination. On the contrary, lexical items that form an idiosyncratic co-occurrence are likely to co-occur in text corpora more often than by chance. Nevertheless, the differing view on the concept has an influence on the collocation extraction models. While within the distributional view purely statistical models are often used, the lexicological approach calls for the addition of linguistic features. For instance, POS or syntactic information can be added by submitting to the statistical models only words in collocation-valid POS patterns or syntactic structures. In any case, the process of collocation extraction is normally performed in two main steps, i.e., (1) extraction, from annotated or raw corpora, of candidates that satisfy the linguistic constraints, if any, and (2) ranking of the candidates by means of statistical models and judgement on the collocation status according to the score, usually by applying a threshold.

Since the preliminary paper by Choueka (1988), where raw frequency counts were used to discriminate collocations from free combinations, and the work by Church and Hanks (1989), who introduced *Mutual Information* (MI) as a statistic metric for measuring lexical association, much research has been carried out in order to develop new models, evaluate and improve existing ones and study the possible combinations of AMs. In what follows, we present an overview of the relevant research in the field.

3.1.1 The onset of collocation extraction

The first works on collocation extraction date from the 80's. Choueka (1988) proposes a method that consists in extracting all n-grams of length 2-6 in a given corpus and then sorting them by frequency. The sequences are considered collocations if their frequency is above a certain threshold. This approach suffers from three main limitations. First, frequent free combinations are often taken as collocations. Second, only uninterrupted sequences of words can be retrieved, and more flexible collocations such as *make - decision*, in which the components can be separated by an arbitrary number of words, are not necessarily found. Finally, rare collocations are not identified, even if their components occur in sequence.

Church and Hanks (1989) present the *Association Ratio* to extract pairs of words that tend to co-occur within a fixed-sized window, thus solving the problem of adjacency. The association ratio is a measure based on the concept of *Mutual Information* (MI) that attempts to capture the binding strength of two words, encoding linear precedence, or the order, in which the words appear. Even though the limitation of word adjacency is overcome with this technique, pairs of related words such as *doctor* and *hospital*, which do not form collocations, are also retrieved.

Smadja (1993) proposes to add parsing information to a statistical method (*Dice Coefficient*) to retrieve collocations from text. Candidate pairs are retrieved in a purely statistical way by taking into account, besides their frequencies, the distance between both elements. Syntactic information is taken into account to filter the candidates in a later stage. Thus, related words such as *doctor* and *hospital*, which are not in a syntactic relation, are discarded.

Like Smadja (1993), Lin (1998) also exploits syntactic information. He first extracts dependency triplets (a head, a dependency type and a modifier) obtained from a shallow-parser. Then, *Mutual Information* (MI) is applied to identify collocations.

Heid (1998) extracts German V-N combinations through query templates. The templates take into account information about sentence boundaries, sequencing and adjacency of word forms, lists of lemmas, and boolean expressions over word forms, lemmas and/or POS-tags. A statistical filter is then applied to select significant collocations.

Pearce (2001) proposes an alternative technique, based on the analysis of the possible substitutions of the candidate pairs by their synonyms extracted

from WordNet. Only word combinations whose elements are linked by a dependency relation are considered as candidates. The technique consists in comparing the co-occurrence of the original word combination with the co-occurrence of one of its components and the synonyms of the other. If the latter is lower than the former, the original combination is considered a potential collocation. The intuition behind this idea is that if a pair of words does not permit substitutions then it is likely to be a collocation.

3.1.2 A step forward in collocation extraction

In the decade of the 2000's, a certain amount of research was carried out in order to evaluate and compare the behaviour of different AMs. In what follows, we first present a summary of this work. A list of the AMs presented in this Section is provided in Appendix A.

Krenn and Evert (2001) evaluate 5 AMs (MI, Dice-coefficient, chi-squared, log-likelihood and t-score) on German support verb constructions and figurative expressions. They extract Prep-N-Verb triplets and calculate the different scores. Their evaluation shows that, overall, t-score achieves best precision values, but that none of the AMs is significantly better suited for the extraction of PP-verb collocations than mere co-occurrence frequency. They also found a negative correlation between support verb constructions and the MI measure. Evert and Krenn (2001) compare the behaviour of MI, the log-likelihood ratio test, chi-squared and t-score on German Adj-N and Prep-N-Verb triplets, showing that the ranking of the different AMs differed depending on the type of collocations, and that none of the measures was able to extract a substantial number of low-frequency collocations (between 58-80% of the collocations, depending on the corpus).

Evert (2005, 2007) presents a comprehensive study of the use of AMs for the task of collocation extraction. He describes the most widely used AMs, including MI, Dice-coefficient, odds-ratio, z-score, t-score, chi-squared or the log-likelihood, for instance, along with methods to understand their mathematical properties and the differences between them. The advantages and mathematical problems of each measure are highlighted, and tools for their empirical evaluation in collocation extraction tasks are provided.

Pecina (2005) evaluates the results of a set of 84 different AMs and describes a new approach that consists of integrating the scores of multiple AMs through Machine Learning techniques. In his experiments, he uses a Logistic Linear Regression algorithm, in which the instances, governor-dependent pairs, are represented by feature vectors consisting of the combination of

the 84 AMs scores, and where the output is also a score for ranking the candidates. In the evaluation of individual AMs, he found that the overall best result was achieved by *Pointwise Mutual Information* (PMI). The combination of multiple AMs leads to a significant performance improvement over the PMI alone. In a follow-up work (Pecina, 2008, 2010), he experiments with other types of machine learning algorithms, i.e., besides Linear Logistic Regression he uses Linear Discriminant Analysis, Support Vector Machines and Neural Networks, on different data sets and tasks. He found that different individual AMs gave different results for the different data and tasks, and confirmed that the combination methods lead to improvements over the individual measures, given that they significantly improved the rankings in all but one dataset.

The comparison of AMs has brought to light several limitations of the use of statistical AMs in the context of collocation extraction. In what follows, we summarize some of these limitations and the efforts that have been done in order to overcome them.

In the first place, it has been observed that PMI tends to assign high scores to low frequency collocations. In order to remove some of this low frequency bias, as well as to provide a measure whose values have a fixed interpretation, Bouma (2009) introduces a normalized variant of PMI. In preliminary experiments, the normalized version of PMI proves to be more effective for the collocation extraction task, showing less bias towards low frequency collocations. A normalized version of MI is also presented, which shows less bias towards high frequency collocations. However, the behaviour of the normalized MI compared to MI is very different, seeming that the normalized MI behaves more like the a PMI measure.

Secondly, Evert (2007) highlights several limitations of the common AMs in general. Among them is the null hypothesis of independence, i.e., assuming that words in language are combined at random. However, this is never the case in natural languages. Rather, words are subject to lexical, syntactic and semantic restrictions. To deal with this issue, Bouma (2010) proposes two measures based on MI and PMI that do not rely on the independence assumption, but on expected probabilities derived from automatically trained Aggregate Markov Models (AMMs). However, in his collocation extraction experiments, carried out on three gold standards, a different behaviour of the new AMs is observed. Depending on the characteristics of the particular corpus, and the number of hidden categories chosen for the AMMs, the new AMs perform better or worse than the baseline. This behaviour prevented

the author to claim the success of the new AMs and made him call for more experimentation and improvement.

Finally, collocations are lexically asymmetrical, however AMs do not consider the lexical asymmetry between the elements of collocations. With regards to this, several ideas have been suggested. Michelbacher et al. (2007) propose two asymmetric AMs: conditional probability and a rank measure derived from the chi-squared test. Their goal was not, however, to identify collocations but to use the measures as a means to model “psychological association”. More recently, Gries (2013) introduced an AM from the associative learning literature, ΔP , which normalizes conditional probabilities and is able to retrieve asymmetric collocations. Carlini et al. (2014) propose an asymmetric normalization of PMI ($NPMI_C$) that uses the probability of the collocate and takes into account dependency information. Their variation is based on the PMI normalization by Bouma (2009).

3.1.3 Conclusions

The general lesson that the comparison of AMs teaches us is that there is no ideal AM. While some enjoy more popularity and are used as standards, e.g., log-likelihood or PMI, the different measures highlight different aspects of collocativity and are better suited for retrieving some types of collocations than others. As pointed out by Evert (2007), for successful collocation extraction it is necessary to have access to a wide range of them, and a good understanding of their properties and behaviour.

3.2 Semantic classification of collocations

While the field of collocation extraction has been the object of extensive research during the last decades, the work on semantic classification of collocations is much more scarce. The few existing systems are based on supervised machine learning algorithms that attempt to classify collocations into semantic classes of varying levels of granularity. In the description that we provide below, we divide these systems into two groups: (1) a first group of systems that make use of external lexico-semantic resources to obtain the semantic features used to classify the collocations, and (2) a second group of more recent systems that infer the semantic features from context.

3.2.1 Lexico-semantic resources as source of semantic information

Among the systems that extract semantic features from external resources is that of Wanner (2004), which also constitutes the pioneer work on semantic collocation classification. Wanner (2004) presents an approach to classification of Spanish VN collocations according to the very fine-grained typology of LFs (Mel'čuk, 1996). An instance-based machine learning method is used, for which training and test sets are disambiguated LF-instances gathered manually. Features are semantic, taken from the hypernym hierarchies of bases and collocates provided by the Spanish part of the EuroWordNet (SpWN). The approach presupposes a componential description of bases and collocates, that is that they can be represented by means of semantic components (in this case, extracted from SpWN). This makes it possible to generalize across collocates with similar meanings. For each LF, the score of a "prototypical" collocation (centroid) is calculated, based on the association strength of each base-component and collocate-component pair. In the classification stage, a score is calculated for each candidate, and compared to the score of the centroid of all LFs. A candidate is assumed to belong to an LF if its similarity with respect to the centroid of that particular LF is higher than a certain threshold. Experiments are performed for collocations belonging to the field of emotions, and for field-independent collocations, focusing on 5 LFs in each experiment ($Oper_1$, $ContOper_1$, $Caus_2Func_1$, $IncepFunc_1$ and $FinFunc_0$ for the domain-dependent experiment; $Oper_1$, $Oper_2$, $Real_1$, $Real_2$ and $CausFunc_0$ for the domain-independent experiment). The main limitation of this work is that the evaluation is performed on manually disambiguated collocations. Although an F-score of 0.7 for domain-independent collocations is achieved, it is unclear how the system would behave with non-disambiguated collocations.

Building upon Wanner (2004), Wanner et al. (2006a) and Wanner et al. (2006b) address this issue by providing a method for automatic disambiguation of test instances. The disambiguation is performed by building the cross-product of all possible senses of each base-collocate pair and using a voting strategy, where each sense bigram "votes" for an LF. The word bigram is assigned the LF with most votes. They report an average F-score of 0.6. Wanner et al. (2006a) and Wanner et al. (2006b) also experiment with other classification techniques, namely Nearest Neighbours (NN), Naïve Bayes (NB) and Tree-Augmented Network (TAN). Their results show that the most suitable technique is the instance-based algorithm. Although its performance is in certain cases somewhat lower than that of

NB and/or TAN, it is more stable.

Wanner et al. (2005) focus on V–N collocations, trying not only to classify them but also to discriminate whether V–N combinations are collocations in the first place. The training data is manually tagged and disambiguated, but the test data is automatically disambiguated. The data comes from a law corpus, but the V–N combinations are not necessarily specialized instances. Their typology consists of 6 syntactically generalized LFs (Real, AntiReal, CausFunc, Oper, IncepOper and Func₀), chosen because of their high frequency. The work builds on Wanner (2004), and Wanner et al. (2006b). They use NN and TAN classifiers trained with features taken from SpWN. They report a precision (p) of 0.68, and a recall (r) of 0.62 for the identification and classification experiment.

Similarly to Wanner et al. (2005), Gelbukh and Kolesnikova. (2012) also target identification of collocations. In their work, they focus on Spanish V–N combinations. They experiment with 68 ML algorithms for classifying V–N combinations as one of 8 LFs (Oper₁, Oper₂, IncepOper₁, ContOper₁, Func₀, CausFunc₀, CausFunc₁ and Real₁) or as the “free-combination” class. Features are lexical: word senses and their hypernyms, manually chosen from SpWN. Although an average F-score of 0.75 is achieved (considering only the algorithm that, for each FL, performs best), the main drawback of this proposal is the need for manual selection of lexical features.

Finally, Huang et al. (2009) classify Adj–N and V–N English collocations according to very coarse-grained semantic categories of the type ‘goodness’, ‘heaviness’, ‘measures’, ‘materials’, etc. They use a thesaurus (Longman Lexicon of Contemporary English) and a sense inventory (WordNet) and employ a random walk to automatically disambiguate and classify the collocates. They report an average $p = 0.76$ and $r = 0.69$.

3.2.2 Contextual features as source of semantic information

The main disadvantages of the proposals described above is that they rely on external resources, which makes them highly dependent on the lexico-semantic resources available for a particular language. In order to overcome this limitation, some works have been put forward that extract semantic information from context, rather than using external resources. These are presented below.

Moreno et al. (2013) focus their experiments on V–N and N–Adj Spanish collocations from the field of emotions. Their data comes from the corpus of examples provided by the *Diccionario de Colocaciones del Español* (DiCE), where collocations are annotated with LFs. They use the Support Vector Machine technique (SVM) to classify collocations according to both the LFs typology, focusing on 10 frequent LFs (Oper₁, Oper₂, Real₁, Func₁, Fact₁, IncepOper₁, IncepFunc₁, IncepPredPlus, CausPredMinus and CausPredPlus) and on 5 broader categories made of LF generalizations (‘intensity’, ‘phase’, ‘manifest’, ‘cause’ and ‘experimenter’). Contextual (i.e., lexical, POS, morphologic and syntactic dependency) features are used to train binary classifiers. They report higher results for the generalized typology than for the individual LFs, showing that it is feasible to automatically classify collocations according to a somewhat more generic typology, of the kind found in collocation dictionaries. In a further experiment, they remove lexical features to see to what extent these are needed or, which is the same, to check whether it is convenient to use domain-dependent classifiers. Even though the accuracy was consistently lower without lexical features, it was not to an extent that would suggest the need of the use of individual classifiers for each semantic field.

In a follow-up work, Wanner et al. (2016) classify Spanish V–N and N–Adj collocations from the field of emotions into broad semantic categories as encountered in general public collocation dictionaries. 5 categories are selected for N–Adj collocations (‘intense L²’, ‘weak L’, ‘positive L’, ‘negative L’ and ‘attributed to someone’) and 16 for N–V collocations (‘begin L’, ‘perform’/‘carry out’/‘experience L’, ‘manifest L’, ‘cause the involvement in L’, ‘cause the existence of L’, ‘cause to be target of L’, ‘undergo L’, ‘suppress manifestation of L’, ‘free oneself of L’, ‘cease or diminish L’, ‘increase L or its manifestation’, ‘L begins to affect someone’, ‘L involves someone’, ‘L continues’, ‘L ceases or diminishes’ and ‘L concerns something’). As Moreno et al. (2013), Wanner et al. (2016) use contextual features (lexical, POS, morphological and syntactic) to train an SVM classifier on data from the DiCE corpus. Besides, they perform a further experiment, adding semantic features to investigate their effect on the classification task, i.e., to see to what extent the contextual features are enough to classify collocations or whether the use of external resources is necessary. Their results show that, even though performance is best when all features are used, semantic features do not lead to a significant rise of accuracy. This demonstrates that contextual features suffice for semantic typification of collocations. They

²‘L’ stands for “base”

report an average F-score of 0.8 for V–N collocations, without semantic features, and even better results for N–Adj.

3.2.3 Conclusions

With few exceptions (Wanner et al., 2005; Gelbukh and Kolesnikova., 2012), the approaches to semantic classification of collocations take for granted collocation extraction and focus solely on classification, leaving the identification of collocations aside.

As shown by Moreno et al. (2013); Wanner et al. (2016), the need of resorting to external lexical resources, as done in the first proposals, is now overcome by the use of contextual features.

Finally, recent work on semantic classification of collocations that takes different typologies as base for their classification has proven the feasibility of classifying collocations with respect to a more generic typology than that of LFs, found, for instance, in collocation dictionaries.

3.3 Collocation error detection and correction

To the best of our knowledge, the first attempt at collocation correction has been carried out by Shei and Pain (2000). The authors propose a method to identify and correct collocation errors in L2 learners' writings that heavily relies on lists of correct and incorrect collocations. In a first stage, word co-occurrences are extracted from the learner text based on POS patterns. These co-occurrences are then checked against a list of automatically retrieved collocations from a reference corpus (RC) and a list of manually revised unacceptable collocations from a learner corpus (LC). The judgement on the status of a combination as a correct or incorrect collocation is made via string or pattern-matching. For those co-occurrences that are not found in any of the lists, alternative collocations are generated and shown to the learner.

Despite the fact that Shei and Pain (2000)'s method proves to be able to offer some help to L2 learners, it presents two important limitations. Firstly, the creation of a database of collocation errors is time-consuming and the resulting database is not easily maintainable, and secondly, the usefulness of a database as such is limited by the manual collection of pre-stored suggestions. For this reason, since Shei and Pain (2000)'s work, multiple strategies that leave behind the use of manually pre-compiled lists

have been proposed. Although the proposed methods vary from work to work, typically, however, the miscollocation correction process follows 2 main steps:

- **Detection of collocation errors.** In a first stage, all word combinations that are likely to form collocations are extracted from the learner's text and judged as correct or incorrect.
- **Suggestion and ranking of potential corrections.** For the collocations that are considered incorrect, alternatives are generated, ranked, and suggested to the learners as potential corrections.

In what follows, we present a summary of how the processes of detection and correction of collocation errors have been addressed in the literature. But before going into any details, let us make three clarifications:

1. Firstly, some of the research on collocation checking focuses solely on collocation error detection, i.e., on judging whether a given word combination is correct or incorrect. Also, some of the proposals focus on detecting collocation errors, leaving the correction aside and thus omitting the second stage.
2. Secondly, it is often the case that, for detecting and judging the status of collocations, methods from collocation extraction on error-free data (such as the use of POS-patterns, syntactic bigrams and AMs) are borrowed. However, the scenery is somewhat more complex when working with learner writings. While in error-free data it suffices to determine whether a word combination is a collocation or a free combination, in non-native data a third phenomenon comes into play that must be differentiated from the other two: collocation errors.
3. Finally, in some occasions, the judgement on the status of collocations and the suggestion of possible corrections are combined in one step. This typically occurs when classifiers are used. Otherwise, a decision is first made regarding the correctness of a collocation, and suggestions are then generated and ranked according to different techniques.

The remainder of this Section is organized as follows. Section 3.3.1 presents a summary of the existing work in collocation error detection. Section 3.3.2

reviews research regarding the generation and ranking of correction candidates. Finally, in Section 3.3.3, the approaches that combine collocation error detection and correction in one step, i.e., those based on Machine Learning classification techniques, are presented.

3.3.1 Detection of collocation errors

The detection of collocation errors is typically done in two steps. A first, optional, step is candidate extraction, in which word combinations that are likely to form collocations are extracted from the learners' writings. A second step is the detection of the errors proper, that is, the judgement of a given combination as correct or incorrect.

The extraction of collocation candidates –word combinations that are likely to form collocations– from learner corpora has been carried out, similarly to the extraction of collocations from error-free corpora, in different ways. The proposed techniques range from the mere use of frequency counts to more complex techniques that incorporate linguistic information.

Among the first are Park et al. (2008), who use n-grams frequencies to discriminate between collocations and non-collocations. A similar method, although slightly more sophisticated, is followed by Tsao and Wible (2009), who propose to use hybrid n-grams, instead of regular n-grams, for more flexibility. A hybrid n-gram, as presented by the authors, is any combination of the word forms, lemmas and POS-tags of an n-sized sequence. For instance, [*PP* made an important decision], [he *make* an *JJ* decision], [*PP* *VV* a important *NN*], etc., are all hybrid n-grams of the sequence *he made an important decision*.

Extracting collocation candidates through n-grams leads, as seen in Section 3.1, to several problems: on the one hand, free combinations that co-occur frequently are taken as collocations. On the other hand, only continuous word co-occurrences can be found, while collocations such as *take - step*, whose components may be separated by an arbitrary number of words, are not necessarily retrieved. The incorporation of morpho-syntactic information into the extraction techniques helps solve these problems, and this solution has been adopted in multiple works. Syntactic information has been used through chunking (see e.g., Yi et al. (2008) and Chang et al. (2008), who also incorporate clause information via a clause model based on Hidden Markov Models) and dependency parsing (Wu et al., 2010; Gao, 2013). However, given the decrease in accuracy of parsers when dealing with non-native language, the most widely adopted approach is POS-pattern match-

ing (Shei and Pain, 2000; Futagi et al., 2008; Ferraro et al., 2011; Wanner et al., 2013b,a; Ferraro et al., 2014).

In order to determine whether a given word combination is a correct or incorrect collocation, most of the methods that have been proposed rely on frequency comparison. A widely adopted strategy consists in calculating, over a RC, the frequencies or association scores of the target collocation, and check them against a threshold or, which is the same, extracting automatically, from a RC, a list of collocations (considering as such the combinations whose scores are above the given threshold) and searching the target combination in the list. Chang et al. (2008) check VN combinations in a collocation list automatically extracted from a RC, reporting a precision of 93.9%. Ferraro et al. (2014) check the frequency of a given combination in a RC, being able to determine its status with an accuracy of 91%. Gao (2013) check combinations in a list of collocations automatically extracted from a RC, reporting an accuracy of around 75%. Tsao and Wible (2009)'s system, based on hybrid n-grams, carried out a matching operation between the word combination (along with its hybrid n-gram versions) and a hybrid n-gram bank extracted from a RC. They do not report any accuracy.

Another strategy based on frequency comparison consists in comparing the frequencies or association scores of the original candidate with the scores of some variants of it. The scores are automatically generated, for instance, by changing a collocate by its synonyms, or by changing word inflections. Park et al. (2008) generate variants of the original collocation by using the Levenshtein distance and calculate a collocation error score based on the frequency of the input and its variants. Futagi et al. (2008) generate alternatives by applying spell checking, varying articles and inflections and changing the collocate by its synonyms. A collocation is judged as correct or incorrect based on its own co-occurrence score and scores of its variants. Futagi et al. (2008) recognize correct collocations with an f-score of 0.91 and incorrect ones with an f-score of 0.34. Yi et al. (2008) use web frequency counts to identify collocation errors. They launch different queries to a web search engine to retrieve examples of correct usage and use their frequencies to decide on the status of the collocations. A precision of 77.8% and a recall of 11.3% are reported.

A final work that relies on frequency comparison to detect collocation errors is that of Wanner et al. (2013b). Focusing on Support Verb Constructions (SVCs), the authors use both pre-compiled lists of collocations and the context of a target combination along with its frequency and the frequency

of known SVCs to decide whether a given SVC is correct or not: a VN co-occurrence whose context of use shows significant similarity with the average context of an SVC, but whose frequency is significantly below the average frequency of known SVCs, is assumed to be an SVC miscollocation.

Although comparison-based techniques can perform relatively well, they still suffer from a major drawback, mainly that they are not able to deal with unseen data. This means that uncommon collocations and collocations that, by mere chance, do not appear in the RC, are judged as incorrect. Kochmar and Briscoe (2014) propose a novel approach to collocation error detection that is able to deal with unseen data. The approach does not rely on direct corpus-based comparison. The error detection task is cast as a binary classification problem, where the features are derived from a semantic analysis of the collocations. The system uses word vector representations that are combined to derive phrase representations, i.e., “collocation” representations (focusing on Adj–N representations). For each combination vector, the original co-occurrence counts of each combination with regard to its context are transformed into MI scores, and 13 semantic measures (based on the vectors’ length, distance and overlap) are calculated. These 13 measures are used as features to train a decision tree, along with the lexical feature that captures the adjective. The authors perform experiments for combinations whose status is independent of the context (achieving an accuracy of 81.13%) and for combinations whose correctness depends on the context (achieving an accuracy of 65.35%), showing in both cases a significant improvement over a baseline based on NPMI (cf. Appendix A) comparison. In a follow up work, Herbelot and Kochmar (2016) integrate context information into the classifier to improve the recognition of erroneous collocations whose status depends on the context. Herbelot and Kochmar (2016) obtain similar results as Kochmar and Briscoe (2014) with an SVM trained on the output of 3 measures of topic coherence plus the adjective itself, but the combination of these features with those by Kochmar and Briscoe (2014) does not lead to significant improvements. Their results show that the system is particularly accurate in classifying form-related errors, but in other cases, such as the incorrect choice of a semantically related word or the use of a general adjective instead of a more specific one, the accuracy of the system is dependent on the adjective. For this reason, different classifiers were proposed for different adjective classes. A preliminary experiment shows that performance improves using adjective-specific information. However, the lack of data prevents the authors from carrying out complete experiments.

Kochmar and Shutova (2016) bring forward a new proposal, also using word vector representations, which consists in training an SVM classifier with a combination of L1 and L2 semantic features to detect errors in VN combinations. L2 features include PMI of the original combination, the lexical features that capture the verb and the noun, and semantic vector space features (such as those used by Kochmar and Briscoe (2014)). L1 features include PMI of the combination in the L1, and the difference between the PMIs in L1 and L2. Kochmar and Shutova (2016) perform several types of experiments on different L1s and report improvement over a baseline classifier whose single feature is L2 co-occurrence frequency of the VN pair. The results show that a combination of L1 and L2 lexico-semantic features improves the performance. The best accuracy they obtain is 71.19% for Spanish L1 Subject-Verb combinations.

3.3.2 Suggestion and ranking of potential corrections

Once a combination has been identified as incorrect, suggestions for correction are generated and ranked to be offered to the learners. In the literature, different ways of generating these suggestions have been explored. Usually, the techniques use statistics or take into consideration different sources of collocation errors, such as semantic similarity, L1 transfer, or form similarity. The idea behind the use of error patterns for the generation of correction suggestions is that knowing the sources of the error can lead to the development of strategies able to generate corrections that are better suited than those generated with the exclusive use of statistical information. In what follows, we describe the different proposals that have been brought forward, and present a summary of the research relevant to each of them.

Co-occurrence based candidate suggestion

In the context of spell checking, a common strategy for error detection is to consider a target word incorrect when it is not present in a reference dictionary. In that case, a list of possible corrections is generated and ranked, often through edit distance metrics. As seen in Section 3.3.1, in the context of collocation checking, the use of similar strategies for collocation error detection and correction has also been proposed (Park et al., 2008; Tsao and Wible, 2009). While in Section 3.3.1 we focused on the detection of the errors, in this section the focus is shifted to error correction.

Park et al. (2008) propose, for an input combination, a set of alternatives, and develop an acceptability metric that relies on a dictionary of n-grams,

to decide on the status of the combination and to rank the generated suggestions, if the combination is considered to be incorrect. The alternatives are generated with the Levenshtein distance by applying inverse error transformation to the original phrase. The errors that are considered are insertion, deletion, transposition and substitution errors. Insertion errors are produced by the addition of an extra word. On the contrary, deletion errors occur when a word is missing. Transposition errors refer to a swap in the order of words. Substitution errors occur when a word is incorrectly replaced by another.

A similar approach is followed by Tsao and Wible (2009), who use hybrid n-grams for error detection and correction. Once an error has been identified, hybrid n-grams³ which nearly match the input (using edit distance to measure proximity) are found. N-grams involving more than one edit are ignored, the rest is kept as correction suggestions. The suggestions are ranked by a “weighted edit distance” that takes into account the types of differences between the input and the candidate. The least distant candidate from the input string is ranked as top suggestion. Neither Park et al. (2008) nor Tsao and Wible (2009) report any performance.

Yi et al. (2008) detect collocation errors based on web frequency counts. First, a set of queries is generated to find appropriate examples of a collocation, and errors are detected depending on the frequencies. If an error is detected, new queries are generated without the erroneous item. The result is a set of alternatives, retrieved from the web, that share similar contexts with the original input. The authors report a precision of 77.8% and a recall of 11.3%.

Focusing only on SVCs, Wanner et al. (2013b) assume that a noun is more likely to co-occur with its support verbs than with other verbs and retrieve the most prominent verbal co-occurrences of a noun as correction suggestions. The product of the z-score and the frequency of the co-occurrence of the base and the collocate candidate is used to measure co-occurrence. The authors report a mean reciprocal rank (MRR) of 0.88 for French SVCs.

Semantic similarity based candidate suggestion

As shown by Kochmar and Briscoe (2014), one of the most frequent causes of erroneous lexical choice is the selection of a lexical unit that is semantically related to the correct one, yet incorrect in the context of the collocation.

³See Section 3.3.1 for the definition of n-grams

tion. For instance, although **forceful* and *strong* are semantically related words, only *strong* is a valid collocate for the base *opinion*. Since semantic similarity is one of the main sources of erroneous collocation usage, a widely used strategy for collocation error correction is based on generating the correction suggestions by substituting the elements of the collocation by semantically related words, such as synonyms or hypernyms, and then ranking the alternatives by means of frequency or AMs scores.

Shei and Pain (2000) generate alternatives with synonyms and other “related words” found in a dictionary of synonyms and definitions that includes definition keywords, collocation paraphrases, words that appear in the context of the collocations, etc. If any alternative is found in a list of correct collocations, it is shown to the learner as a possible correction.

Focusing on the correction of Swedish miscollocates, Östling and Knutsson (2009) generate potential corrections by substituting the collocates with their synonyms and then evaluating them with MI and the log-likelihood measures. The algorithm is evaluated on collocation errors that are artificially generated by replacing the collocate with a synonym or another related word, e.g., a word that shares the same L1 translation. The authors report a p@1 of 57%. No evaluation is performed on “real” learner errors.

Similarly to Östling and Knutsson (2009), Ferraro et al. (2011) and Ferraro et al. (2014) obtain correction candidates by looking for synonyms of the collocate and checking the frequency of the base with each of the synonyms in a RC. Wanner et al. (2013a) extend the algorithm by also obtaining candidates from L1 translations. The combinations above a certain threshold, considered as correct, are ranked using three metrics: affinity metric, lexical context measure and context feature metric. The affinity metric calculates the association strength of the candidate and the base through the log-likelihood measure (see Appendix A), and also takes into account the graphic similarity to the “erroneous” collocate in terms of the Dice Coefficient (see Appendix A), and the synonymy of the candidate with the original collocate. The two other measures are based on distributional semantics. The lexical context measure calculates the affinity of the synonym with the context of the original collocate. The context feature metric considers features other than lexical (POS-tags, grammatical functions, punctuation, etc.) and retrieves the candidate that is preferred by the contextual features. The authors report an accuracy of 54.2% in their experiments with Spanish L2 miscollocations, being able to find the right collocate in the retrieved list of candidates in 73% of the cases.

Finally, Kanashiro Pereira et al. (2013), who focus on the correction of VN miscollocations in Japanese, build the set of correction candidates by using three word similarity measures: thesaurus-based word similarity, distributional similarity, and a confusion set derived from the corrections given in a learner corpus. The first two measures generate the suggestions by finding words that are analogous to the original word. The third measure generates the alternatives based on the corrections in an annotated learner corpus. In order to rank the alternatives, the association strength is measured both in the original collocation and in each generated combination. Candidates are ranked with the weighted Dice Coefficient, the highest alternatives being suggested as corrections. The authors perform experiments for verb and noun suggestions and report a $p@1 = 0.64$, a $p@5 = 0.95$, $r = 0.97$ and an $MRR = 0.75$ for verbs. For nouns, $p@1 = 0.73$, $p@5 = 0.98$, $r = 0.98$ and $MRR = 0.83$.

L1 transfer based candidate suggestion

L1 interference has proven to be another common source of collocation errors (Nesselhauf, 2005; Laufer and Waldman, 2011; Alonso Ramos et al., 2010). Due to this, several authors have also experimented with L1 translation equivalents for generating the correction suggestions. Among them are Wanner et al. (2013a), referred to above, who, apart from synonyms, use L1 translations obtained from a bilingual vocabulary to generate the alternatives.

Chang et al. (2008) also use L1 translation equivalents, which they found in bilingual dictionaries. The generated combinations are checked in a list of collocations automatically extracted from a RC. The combinations that are found in the list are kept as suggestions and ranked by the log-likelihood ratio. The authors report a precision of 84.4% for an automatic evaluation, and 94.1% for a manual evaluation at suggesting the right correction in a list of 10 candidates, and achieve an MRR of 0.66.

Similarly, Dahlmeier and Ng (2011a) use translation information to generate collocation candidates, but unlike Chang et al. (2008), who use bilingual dictionaries, Dahlmeier and Ng use L1-induced paraphrases automatically extracted from parallel corpora, which allows for higher coverage, longer phrases and translation probability estimates. Their approach is implemented in the framework of phrase-based statistical machine translation. The approach obtains a score for each possible translation of a given phrase through a log-linear model that uses different features. This score is used

for ranking and finding the best correction. Experiments with features derived from spelling, homophones, WordNet synonyms and L1-induced paraphrases show that, in isolation, L1-features perform best, and that a combination of all types of features results in better performance than any of the features in isolation. An $MRR = 0.17$ is achieved with automatic evaluation; for manual evaluation, an MRR of 57.26% is reported, with a $p@1$ of 38.20%.

Multiple sources of information for candidate suggestion

Finally, some proposals have been put forward that combine information from different sources in order to generate the correction suggestions. These are presented below.

Liu et al. (2009) retrieve verbal miscollocate correction suggestions from a RC using three metrics: (1) word association strength given by MI, (2) semantic similarity of the collocate with regard to other potential collocates based on their distance in WordNet, and (3) the membership of the incorrect collocate with other potential collocates in an “overlapping cluster” (Cowie and Howarth, 1996). A probabilistic model combines the suggestions offered by each feature and ranks the candidates. A combination of (2) and (3) leads to the best precision achieved for the suggestion of a correction (55.95%). A combination of the three leads to the best precision when a list of five possible corrections is returned (85.71%).

Kochmar and Briscoe (2015) take into account 4 different sources of errors (form similarity, semantically related words, use of general adjectives instead of more specific ones and vice versa, and L1 transfer). They do so by generating alternatives with the Levenshtein distance, WordNet hypernyms, hyponyms and synonyms, and the corrections of erroneous collocations given by an error-tagged learner corpus. The ranking algorithm combines the confusion probability of the corrections with frequency and NPMI. For correction, the authors report an MRR of 0.51, the correction being ranked as first or second in 50% of the times; in a list of 10, the candidate is found in 71% of the cases. When combined with the detection system by Kochmar and Briscoe (2014), an MRR of 0.25 is achieved (although the authors argue that 24.28% of the cases cannot be corrected because the corrections are longer than two words. When these are removed, an MRR of 0.68 is obtained).

3.3.3 Machine learning algorithms for collocation error correction

Finally, the following systems combine collocation error detection and correction through the use of supervised machine learning classification algorithms. All focus on verbal errors.

Liu et al. (2010) combine semantic features from the output of semantic role labelling (SRL) with contextual features within the perceptron learning framework for retrieving verbal suggestions. Taking an input sentence, the system first generates correction candidates by replacing each verb with verbs in a pre-defined confusion set (manually created for the 50 most frequent verbs by consulting lexical resources). Then, for each candidate, SRL-derived features are extracted. These features are used, along with the context (n-gram, syntactic chunk and chunk's headword) of the verb, to train the Generalized Perceptron algorithm and score the candidates. The best candidate is selected as output. The algorithm is trained with native data. A precision of 65.5% and a recall of 44%, with 6.5% of false alarms, is reported.

Wu et al. (2010) train a Maximum Entropy algorithm to suggest verbal collocates. In order to avoid the manual creation of confusion sets, Wu et al. consider any verb as potential correction of any collocate. The algorithm uses contextual features (the token of the base, and 1- and 2-grams of the contexts) to assign a probability to each of the possible corrections, suggesting as final correction the verb with the higher score. The system is trained and tested with a corpus of published academic abstracts, rather than with authentic L2 data. A MRR of 0.518 is achieved in an automatic evaluation.

Sawai et al. (2013) generate the candidate sets from corrections of learner writings taken from *Lang-8*⁴, a Social Networking Service where native users correct the writings of non-natives, and use a classifier to score the candidates. They train the system on native data, adapted to learner writings with the Feature Augmentation technique (Daumé III, 2009), and use contextual features (lexical, POS-tags and word clustering features), evaluating on learner corpora. They show that performance increases when the candidate sets are extracted from learner corpora rather than resources such as WordNet or bilingual dictionaries, and obtain an MRR ranging between

⁴<http://lang-8.com>

0.269 and 0.412 depending on the corpus, with a coverage between 57.6% and 68.9%.

3.3.4 Conclusions

As can be observed from the summary of the work in collocation error detection and correction, most approaches to detect errors are based on AMs candidate comparison (cf., Chang et al. (2008); Ferraro et al. (2014); Futagi et al. (2008), among others). As shown by Ferraro et al. (2011), a drawback of frequency-based criteria is that infrequent collocations, such as literary collocations, are classified as incorrect. Also, they are not able to distinguish between “true” collocations and frequent free co-occurrences. Another serious problem of this type of techniques is that they are not able to deal with unseen data. Approaches based on word vector representations, such as the one proposed by Kochmar and Briscoe (2014), overcome this limitation and outperform the techniques based on AMs.

As for correction, research shows that results are still low for practical applications and there is still much room for improvement.

Firstly, since the tools are not able to correct the error, they must offer a list of correction candidates among which the learner has to choose. Ferraro et al. (2011) is the first proposal that attempts to suggest the exact correction of a collocation error, achieving an accuracy of 54.2%.

Lists of correction candidates are not sufficient, since collocates with other meanings that the original can be also shown. It is necessary to know the semantics of the collocations and propose corrections that fit the meaning intended by the learner.

Existing approaches assume that a unique strategy suffices for the detection and correction of all types of collocation errors. However, collocation errors are very heterogeneous and error-specific techniques that target the specificities of each particular type of error would lead to more accurate error detection and correction.

With the exception of Futagi et al. (2008), who takes into account several grammatical errors related to collocations, all proposals focus on lexical errors. As shown by Alonso Ramos et al. (2010), grammatical collocation errors can be as frequent as lexical ones, and should not be neglected in comprehensive collocation checking tools.

Similarly, apart from few exceptions, such as the work by Kanashiro Pereira et al. (2013), most proposals focus on the correction of the collocate, leaving incorrect bases untouched. Alonso Ramos et al. (2010)'s study shows that errors affecting the base are rather common. Thus, they should not be neglected either.

3.4 Grammatical error detection and correction

The grammatical collocation errors as defined in our typology are different from the grammatical errors addressed in the literature. For instance, not all types of grammatical errors are collocation errors. ‘Subject-verb agreement’ or ‘incorrect verb conjugation’ are considered to be grammatical errors independent of the collocation and are thus not included into the typology. Besides, in other error types, such as ‘Number’ errors, there exist some restrictions regarding their inclusion as collocation errors. In the particular case of ‘Number’ errors, for example, only when a collocation requires that the base needs to be used either in plural or singular, there is an error. This means that a lack of number agreement does not necessarily lead to a collocation error. Despite the above, there is nevertheless a large overlap between the objectives of the works on grammatical error recognition/correction and ours.

Grammatical error correction has attracted much attention in the last years (see Leacock et al. (2014) for a detailed review), especially after the organization of a number of “shared tasks” that provide a common ground for system comparison and evaluation (Dale and Kilgarriff, 2011; Dale et al., 2012; Ng et al., 2013, 2014). Both rule-based and corpus-based techniques have been used, the choice of one or the other usually being subject to the specific types of errors that are targeted by the system. For instance, certain error types, such as subject-verb agreement, can be easily handled by sets of rules and lists, but other types of errors, like determiner or preposition misuse, are more diverse and require the consideration of the context, such that stochastic models are commonly used.

Different systems address different types of errors and, while some focus on a particular kind of error, others attempt to correct errors of any type. Still, most of the work on grammatical error detection and correction concerns the erroneous use of prepositions and determiners. In what follows, we provide an overview of the research on grammatical error detection and correction that is directly related to our work, i.e., that focuses on the types of errors

that are considered in the typology by Alonso Ramos et al. (2010) and that we target in our work.

3.4.1 Determiner and preposition errors

Determiner and preposition errors are among the most common types of errors produced by language learners. For this reason, there is a significant amount of work that deals with determiner and preposition errors. Supervised machine learning has been widely explored, but other methods such as those based on n-grams or machine translation, have also been put forward. Below we present a summary of the work on determiner and preposition errors, dividing the proposals into three main groups, according to the approach that is followed in each case.

Machine learning classifiers

Maximum entropy (ME) models have been repeatedly applied. For instance, Izumi et al. (2003) use ME classifiers to detect 45 classes of errors, including article errors. The features are contextual, and include lexical, morphological and POS information. The classifier detects the presence of an error and assigns it to a specific error type. The training data consists of learner sentences, although the authors also experiment with the addition of the corrected sentences and artificially generated data to the training dataset. For article errors, the addition of these results in an improvement of 30% of precision, while recall remains the same.

Han et al. (2004) and Han et al. (2006) train a ME model on native data, which consists of 6 million noun phrases from the MetaMetrics Lexile corpus, to detect article errors in learner texts. Lexical and POS features are used to compute the probability that the noun phrase has the definite, indefinite or the zero article. The authors evaluate the performance of the model on held-out data from the MetaMetrics Lexile corpus, and also on TOEFL essays, reporting a drop in performance when evaluating on non-native data.

De Felice and Pulman (2008) apply ME to preposition and determiner error correction. They target errors involved in the nine most frequent prepositions in the data, and the definite, indefinite and zero articles. The classifier is trained with examples of correct usage and linguistically motivated features that include lexical, morphological, POS-tags, syntactic features extracted by a parser, and semantic information, extracted by a named entity recognizer and from WordNet. The classifiers are trained on the British

National Corpus (BNC) data and evaluated both on a held-out fragment of BNC and on a portion of the Cambridge Learner Corpus. Results show that accuracy of correct instances is similar in L1 and L2 test sets. However, for articles, the accuracy of incorrect instances is less than 10% in the L2 test set. This drop in accuracy is in accordance with Han et al. (2004) and Han et al. (2006)'s results. In a follow-up work, De Felice and Pulman (2009) found that performance also dropped for the detection of incorrect usage of prepositions.

Tetreault and Chodorow (2008) model the correct usage of prepositions with ME classifiers but, unlike previous approaches, the classifiers are combined with rule-based filters. Tetreault and Chodorow (2008) target insertion and substitution errors for the 34 most common prepositions. The classifier is trained on 7 million instances from the MetaMetrics Lexile corpus, with contextual features that include lexical, POS-tags and syntactic information from phrase-chunking. Tetreault et al. (2010) experiment with the addition of parse features, and obtain a modest improvement in both precision and recall in the detection of incorrect preposition in L2 texts. In order to improve recall, the model is enriched by non-native usage information (Tetreault and Chodorow, 2009). In absence of a large L2 annotated corpus, they use the Google and Yahoo search APIs to compare the distribution of a particular construction in texts found on web pages in an English-speaking country and non-English speaking countries. A marked difference between the distributions is understood as a sign that the construction might be problematic for learners. They build small models specifically tuned for each construction and integrate them into the system of Chodorow et al. (2007) and Tetreault and Chodorow (2008), obtaining an increase of recall in four out of the five cases they experiment with.

Differently to previous approaches, that use native data to build the models, Han et al. (2010) train a ME classifier on learner data. The goal is to detect and correct a subset of the 10 most frequent English prepositions. The classifier is trained on approximately one million instances from the Chungdahm English Learner Corpus, which is composed of essays by Korean learners of English. The corpus lacks exhaustive annotations, which means that some of the instances can be incorrectly tagged as 'correct'. Features include lexical, POS and syntactic information extracted from the context. Evaluation is performed on held-out data from the Chungdahm Corpus. A set of experiments in which the algorithm is trained on different fragments of the Lexile Corpus, ranging from 1M to 5M instances, shows that the model trained on L2 data, even when the error annotations are not

exhaustive and the size smaller, outperforms models trained on well-formed texts, independently of their size.

Gamon et al. (2008) use a decision tree and a language model trained on native data for the detection and correction of determiner and preposition errors. The classifiers are trained on the English Encarta Encyclopedia and Reuters news data. Contextual lexical and POS-tag features are used, along with the relative position of the tokens surrounding the potential insertion point, to train the algorithm. Two classifiers are employed for each type of error, one to determine the presence or absence of the target element, and the other for suggesting which preposition or determiner is most likely to be correct. All suggestions are collected and passed through a language model, trained on the Gigaword corpus, in order to filter out wrong suggestions. When evaluating on L1 data, the use of the language model increases precision dramatically. When evaluating on L2 data (Chinese Learners of English Corpus), encouraging results are obtained, since errors can be corrected with reasonable accuracy (55% for articles; 46% for prepositions). In a follow-up work, Gamon (2010) combines the output of error-specific classifiers and a language model in a meta-classification approach. In this case, he chooses ME classifiers. The meta-classifier takes the output of the classifiers and language model as input. Training instances are gathered by collecting the suggested corrections by the classifiers and the language model on error annotated data, while the class is given by the correct annotation. A decision tree is used as meta-classifier. Annotated learner data stem from the Cambridge Learner Corpus. Results show that, when evaluated separately, the language model notably outperforms the classifier, and that the combination of scores from the classifier and the language model outperforms the language model in isolation.

The idea of using artificial instances to train the classifiers is resumed by Rozovskaya and Roth (2010b), who train an average perceptron model on artificial data for wrong article detection and correction. The way in which the artificial errors are created is described in Section 3.5. Errors are inserted into Wikipedia sentences. Their results show that, with respect to L1 data, accuracy increases when training on artificial data.

N-grams and language models

Despite their simplicity and the severe limitation that they can only estimate probabilities of phrases that appear in the corpus, methods based on counting n-grams have been repeatedly applied for grammatical error

detection and correction. In contrast, language models are able to estimate the probability of phrases that have never been seen in the training corpus. However, they have been rarely used. Among the little research based on language models is the work by Gamon (2010), referred to above, and the work by Lee and Seneff (2006). Lee and Seneff (2006) propose a method for correcting article and preposition errors, among others. In a first step, the input is paraphrased and a word lattice of candidate corrections is generated. Then, a language model is used to select and rank the best candidates. An evaluation on domain-specific artificial data shows that 88.7% of the corrections is accurate.

Below we present a summary of the work based on n-grams. All proposals take the web as source of frequency information.

Yi et al. (2008) use web frequency counts to identify determiner errors. They launch a series of queries, according to different granularity levels, to a web search engine to retrieve examples of correct usage and use the frequencies of these examples to correct the errors.

Bergsma et al. (2009) approach preposition and other errors based on Web counts. They gather counts of a given target word's context in the Web, along with counts of its correction candidates. Contexts are defined as a series of n-grams of various lengths that include the candidates in different positions. Two approaches are proposed for the use of the Web counts. Firstly, the counts are used as features in a supervised classifier, weighting the context's size and position. A Support Vector Machine is trained with 1 million prepositions taken from the New York Times section of the GigaWord Corpus. The second approach consists in summing up the (log) counts for each candidate n-grams. Results show that, without requiring any training data, the second system almost achieves the performance of the classifier.

Elghafari et al. (2010) focus on preposition prediction. They take a 7-gram from the original preposition context (with the preposition in the middle), and modify it by replacing the original preposition by one of the target corrections (the nine most frequent English prepositions). When a preposition is closer to the beginning or end of the sentence, or when 0 frequency is retrieved for all queries, n-grams are reduced down to 3-grams. Elghafari et al. (2010) come to the same conclusion as Bergsma et al. (2009), i.e., that a surface-based approach based on Web counts is competitive compared to classification approaches that rely on complex features. With an accuracy of 77%, the system can compete with machine learning classifiers.

Hermet et al. (2008) develop a system for preposition correction. In a first step, a rule-based processing is used to prune and generalize the context of the target preposition in the original sentence. Then, alternative phrases are generated by replacing the original prepositions with likely confusable ones and their frequencies in a Web search engine are compared to decide which preposition is the correct one. The algorithm is tested on French L2 sentences, and an accuracy of 69.9% is achieved. The performance degrades when a corpus of n-grams, instead of the Web, is used to evaluate the frequencies.

Other methods

Nagata et al. (2005) use a statistical model for article error detection. The model is based on conditional probabilities of articles, and is trained with contextual features. A set of changes has been later made to this original system. Thus, the mass/count noun distinction is explored in Nagata et al. (2006b), and the assumption that the mass/count status of a noun in a discourse does not change is considered in (Nagata et al., 2006c). Nagata et al. (2006a) add prepositional information to the model, which results in a great improvement over the base system.

Hermet and Désilets (2009) build on their previous system (Hermet et al., 2008) and combine it with Statistical Machine Translation (SMT). They use Google Translate to correct preposition errors in the writings of learners of French. A round-trip translation from L2 to L1 and back to L2 is done to obtain the corrections. The input sentence written by the learner is translated into the learner's language, and then back to L2. Machine translation performs slightly worse than the Web based method (Hermet et al., 2008), but the difference is statistically not significant. When combined, accuracy increases up to 83.1%. SMT has also been used by Yuan and Felice (2013) and Felice et al. (2014), who train the system with artificial errors, obtaining an increase in recall.

Dahlmeier and Ng (2011b) present an approach based on Alternating Structure Optimization for article and preposition correction that combines information from native and non-native texts. The authors compare their results against two baselines: training a classifier on only text data or only learner data. Their system outperforms both.

Conclusions

Given that the corpora used for evaluation and the type of tasks differ from approach to approach, it is not possible to make a direct comparison of the results achieved by the proposed methods. Still, some conclusions can be drawn from the review.

Classifiers have been a widely adopted approach to determiner and preposition error detection and correction. Models trained on L1 data perform well when tested on L1 data, but their performance drops when tested on learner language (Han et al., 2004, 2006; De Felice and Pulman, 2008, 2009).

Models trained on L2 data, even when the error annotations are not exhaustive, outperform models trained on well-formed text, independently of their size (Han et al., 2010). The use of artificial error data has also proven to be helpful (Yuan and Felice, 2013; Felice et al., 2014). The combination of native and non-native data in complex systems such as the one proposed by Dahlmeier and Ng (2011b) leads to better results than classifiers trained on only learner data.

Regarding the techniques, Elghafari et al. (2010) and Bergsma et al. (2009) show that surface-based models can compete with L1-training classification approaches that rely on complex features. Language models outperform classifiers trained on L1, but the combination of scores from classifiers and language models in a meta-classification approach outperforms language models in isolation (Gamon, 2010).

3.4.2 Gender, number and word order errors

Because of their regularity, gender, number and word order errors have often been addressed through rule-based systems. Still, statistical approaches have also been proposed. Below, we present a summary of the work related to gender, number and word order errors.

Among the rule-based approaches are Fliedner (2002), Gill and Lehal (2008) and Ibanez and Ohtani (2012). Fliedner (2002) developed a system based on shallow parsing, for which constraint rules are manually created to check noun phrase agreement in German texts. He reports a precision and recall of 67%. Gill and Lehal (2008) developed a rule-based approach for automatic error detection for Punjabi. The system is based on rules and covers errors such as modifier-noun agreement, subject/object-verb agreement and word order. An overall precision of 76.8% and a recall of 87.1% are reported. Ibanez and Ohtani (2012) describe a grammar for automatic detection of

gender and number agreement errors in Spanish texts written by Japanese learners. They report a precision of 64.52% and recall of 71.43% for gender errors, and a precision of 58.62% and recall of 31.48% for number errors.

Sjöbergh and Knutsson (2005) developed a transformation-based rule learner for recognition of grammatical errors, specifically word order errors and erroneously split compounds in Swedish using synthetic data for training the algorithm. Their system achieves a performance comparable to rule-based state-of-the-art grammar correction systems.

Lee and Seneff (2006) propose a method to correct several types of errors, including noun number errors. In a first step, the input is paraphrased and a word lattice of candidate corrections is generated. Then, language models are used to filter and rank the candidates. 88.7% of the corrections of the sentences are at least as good as the original input.

The Microsoft Research ESL Assistant (Gamon et al., 2009) targets a variety of grammatical errors. While preposition and article errors are addressed by statistical techniques, rule-based techniques are developed for other error types such as noun number or word order. The rules are created by inspecting learner data; they are based on a lexical or POS-tag sequence lookup and grouped into different modules. Noun number rules are in the noun-related module, while word order rules are in the adjective-related module. As done by Lee and Seneff (2006), the candidates generated by the rules are filtered by a language model, in this case trained on the Gigaword corpus. An evaluation of the system shows that the performance of the modules depends largely on the test corpus. For the noun-related module, precision varies between 65% and 71%, but for the adjective-related module it ranges from 14% in email data to 64% in a learner corpus.

Within the context of the CoNLL 2013 & 2014 Shared Tasks, a number of teams submitted work on noun number error correction. For instance, Wu et al. (2014) obtain noun number corrections by generating confusion sets with the singular and plural forms of a noun and choosing the best candidate using a language model. A recall of 46.76% is achieved. Among the systems that performed best for noun number errors is the work by Junczys-Dowmunt and Grundkiewicz (2014), who optimize a SMT system by combining web-scale language models and large-scale corrected texts with parameter tuning according to the task metric and correction-specific features. A recall of 58.74% is achieved. Felice et al. (2014) also address error correction within the framework of SMT. Their proposal consists in the addition of artificial errors to train the system. A recall of 54.11% for

noun number errors is reported. Finally, Rozovskaya et al. (2014) combine different classification models that target different error types and address error interaction via Integer Linear Programming formulation (Rozovskaya and Roth, 2013). They obtain a recall of 56.10%.

As in the case of preposition and determiner errors, in most of the approaches to gender, number and order error detection/correction, the corpora used for evaluation and the tasks themselves are different. For this reason, it is also not possible to directly compare the performance of the proposals in this case. Still, two important conclusions can be drawn. The first is that the modelling of learner data (even if it is by means of artificially generated errors) is helpful (Felice et al., 2014) for error correction. The second is that error interaction, which plays an important role in learner language, and which is often disregarded in error detection and correction systems, should be addressed for more effective error detection and correction (Rozovskaya and Roth, 2013).

3.5 Artificial corpora for error detection/correction

Because of the scarceness of annotated L2 corpora, the generation of artificial errors has been considered as a potential substitute of “real” learner data by a significant number of researchers.

Errors can be generated in a deterministic or probabilistic way. Deterministic approaches are those that do not make use of the error distributions observed on learner corpora. Rather, errors are systematically introduced in all relevant instances. Probabilistic approaches try to reproduce learner data and consider learner error distributions to generate and insert the errors.

Artificial corpora have been used for training and evaluating algorithms, and have been developed for different tasks and error types. With the exception of the proposals by Östling and Knutsson (2009) and Herbelot and Kochmar (2016), who introduce lexical errors, all work in artificial error generation focus on the creation of grammatical errors.

In what follows, we present the work on artificial error generation. Section 3.5.1 summarizes the proposals for grammatical errors. Section 3.5.2 focuses on lexical errors.

3.5.1 Grammatical error generation

Below we describe the research on grammatical error generation. We have grouped the proposals according to the way the errors are generated, i.e., deterministically or probabilistically.

Deterministic approaches

Izumi et al. (2003) generate an artificial corpus of article errors to detect omission and replacement article errors. They replace correctly used articles in native texts by incorrect ones (definite, indefinite, and the *zero* article), randomly selected, in order to create a dataset for training a Maximum Entropy model. Evaluation is carried out on real L2 data. Performance improves for omission errors when compared to the use of the original native text, but the detection rate for replacement errors remains the same. The authors attribute this absence of improvement to the features used by the model, since they might not be able to capture well enough the context of the article as to decide whether a definite or indefinite article should be used.

Sjöbergh and Knutsson (2005) create an artificial corpus for the detection of split compounds and another one for the detection of word order errors in Swedish texts. Compounds are automatically splitted by a modified spell checker. Word order errors are created by using two strategies: (1) switch places of a randomly selected word and a neighbouring word, and (2) switch places of a verb and the negative particle and a neighbouring word. The training data consisted of the modified and original sentences. A transformation-based rule learner is trained to detect erroneous words. The evaluation is carried out on learner data. Compared to state-of-the-art rule-based grammar checkers, their system achieves higher recall, but the precision decreases. State-of-the-art rule-based grammar checkers and Sjöbergh and Knutsson's system are complementary; their combination gives best results. No comparison with regard to the use of their system trained on the modified (synthetic) vs. the original (mostly error free) corpus, however, is reported.

Brockett et al. (2006) generate artificial errors for correcting countability errors associated with 14 mass nouns often used incorrectly by English learners. Errors are generated through hand-constructed regular expressions based on typical examples in learner corpora, and injected in all sentences that contain any of the 14 target nouns. The introduced errors change quantifiers (*much* to *many*; *same* to *a*), convert singular to plural, delete

counters (*piece(s), item(s)... of*) and insert determiners. These are combined to account for multiple errors, often found in L2 writings. A balanced training set that consists of a similar number of correct and incorrect sentences is used to train a STM system. Results show that the artificial data proves to be useful for modelling learner countability errors.

Lee and Seneff (2008) create an artificial corpus of verb form errors to improve the correction of subject-verb agreement, auxiliary agreement and complementation errors. The errors are generated by modifying verb forms (bare infinitive, *to*-infinitive, third person singular, *-ing* participle, *-ed* participle) into one of the other forms. The authors investigate how these errors affect parse trees and use their findings to improve the correction of the above mentioned errors.

Ehsan and Faili (2013) generate artificial errors to correct grammatical errors and context-sensitive spelling mistakes in English and Persian using a STM-based approach. Grammatical errors are inserted into correct sentences using predefined error templates so that each sentence contained just one error. If more than one error can be inserted into the same sentence, separate versions are generated, each containing a different error. For the generation of context-sensitive errors, a confusion set is first produced by using the Levenshtein distance. Then, a sentence is generated for each of the elements of the confusion set. The training data consists of the modified sentences along with their original counterparts. As Sjöbergh and Knutsson (2005), Ehsan and Faili compare their results with those of rule-based grammar checking, finding that the combination of these with their systems results in best performance.

Probabilistic approaches

Wagner et al. (2007) and Wagner et al. (2009) generate 4 types of grammatical errors in correct sentences from the BNC (missing word errors, extra word errors, real-word spelling errors and agreements errors) to evaluate different approaches to grammatical error detection. For each sentence, an attempt is made to produce four ungrammatical sentences, one for each type of error. Their approach to error generation is based on an analysis of the types and frequency distribution of errors in an error corpus. However, it is not fully probabilistic since an attempt is made to generate errors in all sentences. For missing word errors, a word is automatically removed following the POS error distributions in the reference learner text: the analysis of the error corpus shows that among all the deletions, 28% are determiners,

23% verbs, 21% preps, etc. For extra word errors, they either duplicate randomly any word or POS or insert an arbitrary token in the sentence. For real-word spelling errors, they compile a list of commonly mistaken words based on the error corpus and insert an error when one of the elements of the pair is found on the sentence, substituting it with the other. Only an error per sentence is generated. Finally, they introduce Subj-V and Det-N agreement errors by replacing a determiner, noun or verb in singular with their plural counterparts, and vice versa. Both types of agreement errors are considered equally likely. Wagner et al. (2009) also perform an evaluation on real learning data, reporting a loss of accuracy, which confirms that training and test data should be as similar as possible.

Foster and Andersen (2009) present a tool for automatic generation of errors that takes as input a corpus that is assumed to be grammatical, and an error analysis file, and produces as output a tagged corpus of erroneous sentences. The generation of errors is based on the insert, delete, move and substitute operations. For each operation, different subtypes of errors can be generated according to the specifications of the user, who can also specify the error frequencies, in which case the system would try to generate errors until satisfying the given proportion or until all the sentences have been tried. If no statistical information is supplied, the tool tries to insert one error per type in each sentence. A replication of the experiments in Wagner et al. (2009) with an artificial corpus created by the tool resulted in an increase of accuracy. In another experiment, the authors first try to mimic the errors found in the Cambridge Learner Corpus (CLC) and then train a classifier to detect the errors. A decrease of accuracy of 6.2% compared to training on real learner data is reported. This drop shows that the tool fails to accurately mimic the real learner corpus, which can be explained by the spelling errors, error combinations and other types of errors that are present in the CLC but not included on the error analysis file.

Rozovskaya and Roth (2010b) propose four methods for artificial error generation. Classifiers are then used to detect and correct article errors in texts written by non-native speakers. The proposed methods are the following:

General: the target words (articles) are replaced by another randomly chosen article with a given probability, remaining unchanged otherwise.

Distribution before annotation: the distribution of errors in the automatically generated corpus is changed according to the distribution found in non-native texts before correction is done.

Distribution after annotation: the distribution of errors in the artificial

corpus is changed according to the distribution found in non-native texts after correction is performed.

L1-dependent error distribution: the errors are generated according to particular confusions observed in the non-native texts, and their probabilities: $P(\textit{source}|\textit{target})$.

For the experiments, errors are injected into Wikipedia sentences. The performance of an Average Perceptron classifier trained with native data is compared to the performance achieved when training with the four types of artificial corpus. Training on the artificial data produces higher accuracy than training on error-free data. This is true for each of the generation strategies that are proposed, however the best results are achieved by the methods that use knowledge about error patterns and error distribution observed in non-native text.

Rozovskaya and Roth (2010a) arrive at the same conclusions for preposition error correction. An Average Perceptron classifier trained on artificial data that reflect the confusions and probabilities observed in non-native texts outperforms other training paradigms based on training on error-free data.

Given that errors are rather sparse, recall tends to be low. In order to overcome this issue, Rozovskaya et al. (2012) proposes an inflation method that increases the number of errors while at the same time keeping the confusion probabilities. The method improves recall while maintaining high precision in experiments with determiner and preposition error correction. It has been used in later research (Putra and Szabó (2013), Rozovskaya et al. (2013), Rozovskaya et al. (2014)).

Dickinson (2010) describes an approach to morphological error generation for Russian. Errors are created by a *guided* random combination of morphemes into full forms. The randomness is constrained by taking into account POS information, frequency and the types of errors that are likely to occur (not based on real error corpus distribution, but rather on their initial ideas of linguistic plausability). The artificial dataset is evaluated on POS-tagging accuracy, but not on error detection tasks.

Imamura et al. (2012) use artificial data based on Rozovskaya and Roth (2010b)'s method in order to correct Japanese particle errors drawing upon discriminative sequence conversion and domain adaptation. Their experiments show that training on the artificial corpus and applying a domain adaptation technique leads to the best results, higher than when training on the real-error corpus alone or in combination with the artificial corpus.

Yuan and Felice (2013) and Felice et al. (2014) use artificial errors within an SMT framework, obtaining an increase in recall. Error patterns, or *rules*, are automatically extracted from a learner corpus and the corrections of its errors. Then, these rules are randomly applied to the correct sentences, generating synthetic errors. Felice and Yuan (2014) generate a more sophisticated artificial corpus, following Rozovskaya et al. (2012). Unlike previous approaches they use linguistic information, such as morphology, POS, semantic classes or word senses, to refine the probabilities of errors, based on the distribution of errors found in error corpora. The technique was applied to a larger number of error types, including open-class errors. The use of the different linguistic information is evaluated for each type of error. In general, error distributions and POS information produce the best results, but the different types of linguistic information are better suited for different error types. Results also show that their methods improve precision at the expense of recall.

Cahill et al. (2013) compare the three paradigms of training on error-free data, error data and artificial data for the task of preposition correction. They choose the widely adopted system by Tetreault and Chodorow (2008) (a multinomial logistic regression classifier). Native corpora are extracted from the Wikipedia and news texts; error data from the Wikipedia Revisions and lang-8.com website, and their corrections. Artificial corpora are generated with the error distributions of the two corrected real-error corpora. Their evaluation shows that the performance of the classifiers when trained on the Wikipedia Revisions is stable across different test sets, and when trained on artificial data following the distributions from the Wikipedia revisions, they perform equally well.

3.5.2 Lexical error generation

Östling and Knutsson (2009) generate collocation errors by replacing the collocate with a synonym or another related word that shares the same L1 translation. The artificial corpus is used for evaluation of a collocation error correction system.

Similarly, Herbelot and Kochmar (2016) generate an artificial dataset for the evaluation of their algorithm for A–N lexical error detection. The errors are inserted by random substitution of adjectives, assuming that in most cases the substitution would produce a lexical error. The authors experimentally show that training and evaluation on this dataset produces a decrease in ac-

curacy, claiming that quality artificial data is harder to generate for content words than for grammatical errors.

3.5.3 Conclusions

In the view of the lack of sufficiently large error-annotated resources, the use of artificial corpora has repeatedly been proposed as way to model L2 errors, or as test data to carry out experimental evaluation of a given approach.

Regarding grammatical errors, despite the fact that the proposed methods for error generation do not faithfully reproduce L2 language, in the sense that only specific error types are injected into otherwise error-free texts, research shows that synthetic datasets can be successfully used for training algorithms, leading to better performance than training only on native data (Rozovskaya and Roth, 2010b). Still, artificial corpora cannot yet be taken as a replacement of L2 writing, as shown by the poorer performance of the system by Wagner et al. (2009) and Foster and Andersen (2009) when the evaluation is carried out on “real” L2 data, rather than on artificial data.

As for lexical errors, the drop in accuracy achieved by Herbelot and Kochmar (2016) when training and evaluating their algorithm on artificial data shows that random substitution is not a valid approach for the generation of lexical errors (as opposed to grammatical errors (Rozovskaya and Roth, 2010b)).

The scarce work on the development of artificial lexical error corpora along with the poor results achieved to date calls for new strategies that better mimic learners’ errors. For instance, Herbelot and Kochmar (2016) show that random substitution is not a realistic way to reproduce L2 errors. In fact, language learners do not produce errors randomly. A deep analysis of the error causes, error patterns and error types should be considered for the generation of quality artificial lexical error corpora.

Semantics-driven recognition of collocations

As stated in Chapter 1, one of the main goals of this thesis is to propose a method for the automatic extraction and semantic classification of collocations. We pointed out that this task is necessary given the lack of collocational resources and the high cost of their compilation. In order to effectively support L2 learners, techniques are thus needed that are able not only to retrieve collocations, but also provide, for a given base (or headword) and a given semantic gloss of a collocate meaning, the actual collocate lexeme. In this chapter, we describe two techniques that have been developed to that end and present the set of experiments that are performed to evaluate the techniques. Section 4.1 introduces word vector representations, or word embeddings, on which our work is grounded, and justifies our choice of the approaches. Sections 4.2 and 4.3 are devoted to the techniques and experiments, each focusing on a different technique. Finally, Section 4.4 summarizes the conclusions.

4.1 Word embeddings based techniques

As seen in Chapter 3, the task of collocation extraction has been largely explored in the last decades, cf. e.g., (Choueka, 1988; Church and Hanks, 1989; Smadja, 1993; Kilgarriff, 2006; Evert, 2007; Pecina, 2008; Bouma, 2010). The work on semantic classification of collocations (Wanner et al., 2006b; Moreno et al., 2013; Wanner et al., 2016) is more scarce and, generally, takes for granted that collocation extraction is done a priori, using,

as input to the classifiers, lists of collocations that have previously been manually compiled. However, the integration of collocation extraction and collocation classification into one pipeline is not convenient since, giving as input to the classifier the output of a collocation extraction system, i.e., word combinations that may or may not be collocations, would lead without doubt to a decrease in performance. Simultaneous extraction and classification would be, thus, desirable.

On the other hand, research has consistently shown that word embeddings, or word vector representations, are a very effective resource in tasks involving semantics. Taking as inspiration the Neural Probabilistic Model (Bengio et al., 2006), Mikolov et al. (2013b) proposes an approach for computing continuous vector representations of words from large corpora by predicting words given their context, while at the same time predicting context, given an input word. The vectors computed following the approaches described in (Mikolov et al., 2013b,c) have been extensively used for semantically intensive tasks, mainly because of the facility that word embeddings have to capture relationships among words which are not explicitly encoded in the training data. Among these tasks are: Machine Translation (Mikolov et al., 2013a), where a transition matrix is learned from word pairs of two different languages and then applied to unseen cases in order to provide word-level translation; Knowledge Base (KB) Embedding (transformation of structured information in KBs such as *Freebase* or *DBpedia* into continuous vectors in a shared space) (Bordes et al., 2011); Knowledge Base Completion (introduction of novel relationships into existing KBs) (Lin et al., 2015); Word Similarity, Syntactic Relations, Synonym Selection and Sentiment Analysis (Faruqui et al., 2015); Word Similarity and Relatedness (Iacobacci et al., 2015); and taxonomy learning (Fu et al., 2014; Espinosa-Anke et al., 2016). From these applications, we can deduce that word embeddings provide an efficient semantic representation of words and concepts, and, therefore, might also be suitable for the acquisition of collocational resources.

In order to examine the above-mentioned hypothesis that word embeddings can be used in collocation-related tasks, we propose two methods for collocation acquisition which strongly rely on relational properties of word embeddings for discovering semantic relations. Specifically, we propose an unsupervised example-based approach and a semi-supervised approach based on the linear relations between semantically similar words in different vector spaces. The two methods target collocation extraction and classification simultaneously.

4.2 Example-based approach

In this section, we present an unsupervised approach to collocation extraction and classification. Section 4.2.1 introduces the algorithm, Section 4.2.2 describes the experimental setup, and Sections 4.2.3 and 4.2.4 focus on the results and discussion.

4.2.1 Exploiting the analogy property

The first approach that we explore is based on the ability of word embeddings to capture syntactic and semantic regularities in language, as discovered by Mikolov et al. (2013c). In their work, the authors found that specific relationships, like the male/female relationship, are characterized by a determined vector offset, i.e., in the embedding space, pairs of words sharing a particular relationship are linked by a similar offset. Both syntactic and semantic tasks were formulated as analogy questions, e.g., $x \sim woman \equiv king \sim man$ (implying $x=queen$). We apply this simple fact to discover and classify collocates, considering, as target relationships, different types of LFs or semantic categories. In our case, therefore, the approach exploits the inherent property of embeddings for drawing analogies between bases and collocates, such as, e.g., $x \sim applause \equiv heavy \sim rain$ (implying $x=thunderous$), cf., Figure 4.1. We calculate, for each semantic category, a prototypical vector offset and use this offset to retrieve other instances belonging to the given category.

Using a state-of-the-art continuous word representation, the algorithm takes as input seed a single representative example (a base and a collocate) of a specific semantic category in order to retrieve from a corpus a set of collocates that belong to the same category, for new bases. We propose a second stage, in which a filtering based on the association metric $NPMIC$ (Carlini et al., 2014), cf., Appendix A, is applied to remove suggested collocates that do not co-occur with the base. In what follows, we outline first the setup of our experiments and present then their outcome and a discussion of the results. Even though we focus in our experiments on Spanish, the technique is scalable and applicable to other languages, since it only requires raw data, for training the word embeddings, and an example of a collocation belonging to each semantic category.

Our algorithm produces, for each LF $\iota \in LF$, and a given base b_ι , a set BC of $(b_\iota, \varsigma_\iota)$ pairs, where ς_ι is a collocate which has been retrieved from a corpus in two stages. In the first stage, the similarity between the relation that ς_ι

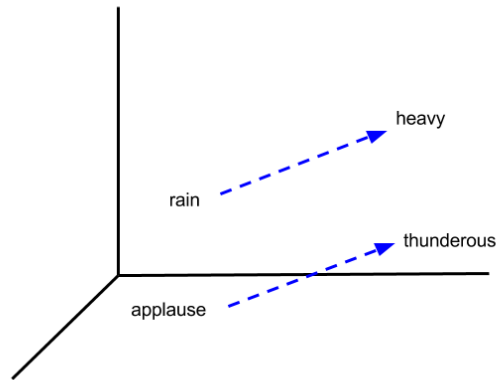


Figure 4.1: Examples of vector offsets for ‘intense’ collocations

holds with b_l and the relation held by the components of a representative collocation ϕ_l (with the base b_l^ϕ and the collocate ζ_l^ϕ) is computed. In the second stage, a filtering is applied, based on the collocation-specific statistical independence metric $NPMI_C$ (Carlini et al., 2014).

Algorithm 1 outlines the two stages. The first stage (lines 4 – 9) consists, first, in retrieving a candidate set by means of the function $relSim$, which computes the similarity of the relation between b_l^ϕ and ζ_l^ϕ to the relation between b_l and a hidden word x . $relSim$ can be thus interpreted as satisfying the well-known analogy “ a is to b what c is to x ”, exploiting the vector space representation¹ of a , b , and c to discover x (Zhila et al., 2013). Specifically, we compute $v_b - v_a + v_c$ in order to obtain the set of vectors closest to v_x by cosine distance. To obtain the best collocate candidate set, we retrieve the ten most similar vectors to x , where x is the unknown collocate we aim to find. This is done over a model trained with $word2vec$ ² on a 2014 dump of the Spanish Wikipedia, preprocessed and lemmatized with Freeling (Atserias et al., 2006).

¹We denote the vector of a word as v , e.g., v_a .

²<http://word2vec.googlecode.com/>

Algorithm 1: Collocate Discovery Algorithm

```

Input:
1   $LF$  // Set of Lexical Functions;
2   $B$  // Set of manually selected bases;
3   $\varepsilon$  // Word embeddings model;
Output:
4   $\Lambda$  // Final resource;
5   $\Lambda = \emptyset$ ;
   initialization;
6  for  $\iota \in LF$  do
7      for  $b_\iota \in B$  do
8           $BC = \emptyset$  // Base and collocates set;
9           $C = relSim(b_\iota, \phi_\iota, \varepsilon)$ ;
10         for  $\varsigma \in C$  do
11              $conf = NPMIC(b_\iota, \varsigma)$ ;
12             if  $conf > \theta$  then
13                  $BC = BC \cup \{(b_\iota, \varsigma)\}$ ;
14             end
15              $\Lambda \cup \{BC\}$ ;
16         end
17     end
18 end
19 return  $\Lambda$ 

```

The second stage (lines 10–14) implements a filtering procedure by applying $NPMIC$, an association measure that is based on Pointwise Mutual Information, but takes into account the asymmetry of the lexical dependencies between a base and its collocate (see Chapter 3). We set the association threshold θ to 0, such that all (b_ι, ϕ_ι) collocation candidates below θ are discarded.

4.2.2 Experimental setup

For these experiments, we focus on eight of the most productive LFs or semantic glosses in Spanish (see Table 4.1 for the list, along with an example; see Table 2.1, in Chapter 2, for the correspondence between LFs and seman-

Semantic gloss	Representative example
‘intense’	<i>gran idea</i> ‘great idea’
‘weak’	<i>leve cambio</i> ‘slight change’
‘create’, ‘cause’	<i>crear [un] entorno</i> ‘create [an] environment’
‘put an end’	<i>romper [una] amistad</i> ‘break [a] friendship’
‘increase’	<i>aumentar [el] precio</i> ‘increase [the] price’
‘decrease’	<i>disminuir [el] precio</i> ‘decrease [the] price’
‘good’	<i>día bueno</i> ‘good day’
‘show’	<i>expresar afecto</i> ‘express affection’

Table 4.1: Seed examples for each semantic gloss

tic glosses). Five of them are verbal collocate glosses in V–N collocations (‘create, cause’, ‘put an end’, ‘increase’, ‘decrease’ and ‘show’), and three are property glosses in Adj–N collocations (‘intense’, ‘weak’ and ‘good’).

As described in the previous section, the algorithm requires a seed example as input to the acquisition of collocates of a given gloss. Therefore, for each gloss, we take a representative collocation, i.e., a collocation whose collocate has a general abstract meaning similar to that of the target gloss, such as *crear [un] entorno* ‘create [an] environment’ for ‘create, cause’, or *disminuir [el] precio* ‘reduce [the] price’ for ‘decrease’. The seed examples chosen for each gloss can be seen in Table 4.1.

Additionally, for each gloss, 20 bases are selected to test the algorithm. These bases are manually chosen nouns for which at least one collocate with the meaning of the targeted gloss is available. For each of these bases, candidate collocates are extracted automatically following the algorithm described above. The retrieved candidates for each test base are tagged as either correct or incorrect, according to two criteria: (1) whether the candidate correlates with the base forming a correct collocation and, if criterion (1) is fulfilled, (2) whether the collocate correctly belongs to the particular semantic gloss.

4.2.3 Outcome of the experiments

To the best of our knowledge, no other work on simultaneous retrieval and classification of collocations has been carried out, therefore we cannot compare our results to any reference work.

Semantic gloss	#candidates	#collocations	#correct gloss
'intense'	74	70	59
'weak'	17	12	0
'create', 'cause'	64	49	44
'put an end'	56	42	15
'increase'	70	61	42
'decrease'	44	40	6
'good'	67	47	24
'show'	26	15	10

Table 4.2: Number of collocations found for each semantic gloss

To assess the performance of our approach, we calculate its precision,³ taking into account: (1) the number of candidates that correctly correlate with the base, and (2) the number of collocates whose semantics matches the given semantic gloss. Tables 4.2 and 4.3 display the outcome of the experiments. Table 4.2 shows the number of collocate candidates obtained for each gloss after the application of the $NPMI_C$ filter (second column); the number of correct collocations formed by the given bases and the retrieved collocate candidates (third column), and the number of correctly typed retrieved collocations with respect to each gloss (fourth column). Table 4.3 shows the achieved precision during the identification of correct collocations and during the typification of the collocations calculated over all candidates of a gloss from Table 4.2 (first value in the third column) and over the correctly identified collocations (second value in the third column).

Some of the collocates retrieved for each semantic gloss can be seen in Table 4.4.

4.2.4 Discussion

As can be observed in Table 4.3, the performance of the algorithm is not uniform across the target set of semantic glosses. On the contrary, its accuracy depends greatly on the particular gloss. In what follows, we present some observations and conclusions derived from the analysis of the retrieved instances.

³In our experiments, we omit recall because the lack of comprehensive semantically-tagged collocation resources for Spanish makes it impossible to have reference gloss-specific collocate lists suitable for its calculation.

Semantic gloss	Precision (p)	
	(identif. collocations)	(glosses)
‘intense’	0.95	0.80 0.84
‘weak’	0.71	0.00 0.00
‘create’, ‘cause’	0.77	0.69 0.90
‘put an end’	0.75	0.27 0.36
‘increase’	0.87	0.60 0.69
‘decrease’	0.91	0.14 0.15
‘good’	0.70	0.36 0.51
‘show’	0.58	0.38 0.67

Table 4.3: Performance of the acquisition and classification of collocations with respect to semantic glosses

With a $p = 0.95$, the system’s performance for ‘intense’ is close to human judgement as far as the identification of collocations is concerned. The precision of the correct recognition of a collocation as belonging to ‘intense’ is somewhat lower ($p = 0.80$). Most of the erroneous typifications as ‘intense’ are due to two reasons: (1) semantic similarity of the collocate to the meaning ‘intense’ (as, e.g., *creciente* ‘growing’), and (2) the failure of word embeddings to distinguish ‘intense’-collocates from their antonyms (as, e.g., *mínimo* ‘minimal’).

In the case of ‘create, cause’, several free combinations are judged as collocations; cf., e.g., *unificar* [*un*] *sistema* ‘to unify [a] system’ or *idear* [*un*] *sistema* ‘to design [a] system’. Still, almost 70% of the obtained candidates are correctly typified as ‘create, cause’; cf., e.g., *desatar* [*una*] *epidemia* ‘to spark [a] pandemic’, *desencadenar* [*una*] *crisis*, ‘to trigger [a] crisis’, *redactar* [*un*] *informe*, ‘to draft [a] report’ or *promulgar* [*un*] *edicto*, ‘to issue [an] edict’, etc.

The number of collocates that do not convey the targeted meaning is considerably higher for ‘increase’ than for other glosses such as ‘intense’ or ‘create, cause’. Unsurprisingly, most of the collocates retrieved for ‘increase’ convey exactly the opposite meaning (‘decrease’), which can be easily explained by the semantic relation that antonyms show when represented by word embeddings. Among the collocations with correct meaning, we obtain *mejorar* [*la*] *estabilidad* ‘to improve stability’, *incrementar* [*la*] *cobertura* ‘to increase coverage’, *fortalecer* [*el*] *liderazgo* ‘to strengthen leadership’, and *estimular*,

Semantic gloss	Retrieved Examples
'intense'	<i>lluvia torrencial</i> 'torrential rain' <i>viento huracanado</i> 'hurricane-force winds' <i>ruido ensordecedor</i> 'deafening noise' <i>valor incalculable</i> 'inestimable value'
'create', 'cause'	<i>desatar [una] epidemia</i> 'to spark [a] pandemic' <i>desencadenar [una] crisis</i> 'to trigger [a] crisis' <i>redactar [un] informe</i> 'to draft [a] report' <i>promulgar [un] edicto</i> 'to issue an edict'
'put an end'	<i>demoler [un] edificio</i> 'to demolish [a] building' <i>apagar [un] fuego</i> 'to extinguish [a] fire' <i>resolver [un] problema</i> 'to solve [a] problem' <i>anular [un] acuerdo</i> 'to nullify [an] agreement'
'increase'	<i>mejorar [la] estabilidad</i> 'to improve stability' <i>incrementar [la] cobertura</i> 'to increase coverage' <i>fortalecer [el] liderazgo</i> 'to strengthen leadership' <i>estimular [la] economía</i> 'to stimulate [the] economy'
'decrease'	<i>minimizar [un] valor</i> 'to minimize [a] value' <i>reducir [una] tasa</i> 'to reduce [a] rate' <i>reducir [un] salario</i> 'to reduce [a] salary' <i>minimizar [un] coste</i> 'to minimize costs'
'good'	<i>posicin excelente</i> 'excellent position' <i>carrera impecable</i> 'impeccable career' <i>resultado satisfactorio</i> 'satisfactory result' <i>forma perfecta</i> 'perfect shape'
'show'	<i>manifestar [una] preocupaci3n</i> 'to manifest [a] concern' <i>reflejar alegr3a</i> 'to reflect joy' <i>evidenciar [una] mejor3a</i> 'to show improvement'

Table 4.4: Examples of correctly retrieved collocates for each semantic gloss

reactivar [la] *economía* ‘to stimulate, revive [the] economy’.

As with ‘create, cause’, in the case of the types ‘good’ and ‘show’, a handful of free combinations are judged as collocations. Some examples of these combinations are *actitud sincera* ‘sincere attitude’, *intención indudable* ‘unquestionable intention’, or *aspecto inusual* ‘unusual aspect’, for the gloss ‘good’; *reafirmar* [el] *apoyo* ‘to reassert support’, *definir* [una] *característica* ‘to define [a] feature’ or *justificar* [un] *temor* ‘to justify [a] fear’, for the gloss ‘show’. Precision for ‘good’ and ‘show’ is somewhat low. Two main reasons could be the cause of this decrease of performance: that the chosen seed examples are not sufficiently common or general, and therefore not representative enough for the glosses in question, or that these glosses present a wider meaning, and are thus more difficult to attain. Further research is needed to assess these issues.

Finally, as far as ‘weak’, ‘put an end’ and ‘decrease’ are concerned, whose meanings are opposite to ‘intense’, ‘create, cause’ and ‘increase’, the number of candidates retrieved by the system that are correct collocates remains high. However, the precision of the classification with respect to the target semantic glosses drops significantly when compared to their *positive* counterparts. This occurs because, as stated above, word embeddings fail to distinguish between antonyms, considering words with opposite meanings as actual synonyms. Most of the collocates found for ‘weak’, ‘put an end’ and ‘decrease’ are correct instances of ‘intense’, ‘create, cause’ and ‘increase’. For instance, we obtain *luz cegadora* ‘blinding light’ and *daño severo* ‘severe damage’ for ‘weak’; *levantar* [un] *edificio* ‘to erect a building’ and *encender* [un] *fuego* ‘to light [a] fire’ were found for ‘put an end’; and for ‘decrease’ cases such as *incrementar* [un] *salario* ‘to increase wages’ and *aumentar* [el] *valor* ‘to increase [a] value’ are obtained.

4.3 Weakly supervised approach

In this Section, we describe a weakly supervised approach to collocation extraction and classification. Section 4.3.1 introduces the algorithm, Section 4.3.2 focuses on the experimental setup, and Sections 4.3.3 and 4.3.4 present the results and discussion.

4.3.1 Methodology for the acquisition of collocation resources

The second method that we explore for automatic retrieval and typification of collocations exploits the fact that semantically related words in two different vector representations are related by linear transformation (Mikolov et al., 2013a). This property has been exploited for word-based translation (Mikolov et al., 2013a), learning semantic hierarchies (hyponym-hypernym relations) in Chinese (Fu et al., 2014), and modelling linguistic similarities between standard (Wikipedia) and non-standard language (Twitter) (Tan et al., 2015). We learn a *transition matrix*, that is a linear transformation matrix, over a small number of collocation examples (base-collocate pairs), which belong to a base space and a collocate space, where collocates share the same semantic gloss, or category. Then, the matrix is applied to unseen bases in order to discover new collocates for the given category.

As already mentioned above, our approach relies on Mikolov et al. (2013a)’s linear transformation model, which associates word vector representations between two analogous spaces. In Mikolov et al.’s original work, one space captures words in language L_1 and the other space words in language L_2 , such that the found relations are between translation equivalents. In our case, we define a base space \mathcal{B} and a collocate space \mathcal{C} in order to relate bases with their collocates that have the same meaning, in the same language (see Figure 4.2 for an illustration, for the gloss ‘intense’). To obtain the word vector representations in \mathcal{B} and \mathcal{C} , we use Mikolov et al. (2013c)’s *word2vec*.⁴

The linear transformation model is constructed as follows. Let \mathbf{T} be a set of collocations whose collocates share the semantic gloss τ , and let b_{t_i} and c_{t_i} be the collocate respectively base of the collocation $t_i \in \mathbf{T}$. The base matrix $B_\tau = [b_{t_1}, b_{t_2} \dots b_{t_n}]$ and the collocate matrix $C_\tau = [c_{t_1}, c_{t_2} \dots c_{t_n}]$ are given by their corresponding vector representations. Together, they constitute a set of training examples Φ_τ , composed by vector pairs $\{b_{t_i}, c_{t_i}\}_{i=1}^n$.

Φ_τ is used to learn a linear transformation matrix $\Psi_\tau \in \mathbb{R}^{\mathcal{B} \times \mathcal{C}}$. Following the notation in (Tan et al., 2015), this transformation can be depicted as:

$$B_\tau \Psi_\tau = C_\tau$$

We follow Mikolov et al.’s original approach and compute Ψ_τ as follows:

⁴<https://code.google.com/archive/p/word2vec/>

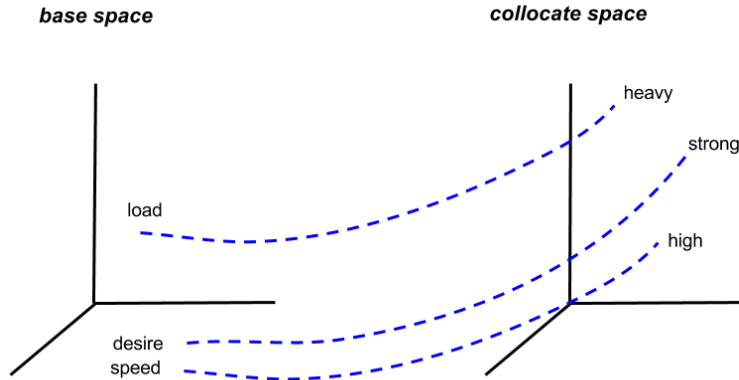


Figure 4.2: Example of relations between bases and ‘intense’ collocates

$$\min_{\Psi_\tau} \sum_{i=1}^{|\Phi_\tau|} \|\Psi_\tau b_{t_i} - c_{t_i}\|^2$$

Hence, for any given novel base b_{j_τ} , we obtain a novel list of ranked collocates by applying $\Psi_\tau b_{j_\tau}$. The resulting candidates are then filtered in two steps. First, only candidates that form with the base a valid collocation POS-pattern are kept. Then, the association measure $NPMI_C$ (Carlini et al., 2014) is applied to the remaining candidates to guarantee the co-occurrence between the base and the candidate.

We carry out experiments with ten different collocate glosses (including ‘do’ / ‘perform’, ‘increase’, ‘decrease’, etc.) with Spanish and English material. For most glosses, an approach that combines a stage of the application of a gloss-specific transition matrix with a filtering stage based on POS-patterns and statistical evidence outperforms approaches that exploit only one of these stages as well as our previous technique presented in Section 4.2.

4.3.2 Experimental setup

As mentioned above, we perform our experiments on the ten semantic collocate glosses listed in the first column of Table 4.5: eight verbal collocate glosses in V–N collocations (‘perform’, ‘begin to perform’, ‘stop performing’, ‘increase’, ‘decrease’, ‘create, cause’, ‘put an end’ and ‘show’) and two property glosses in Adj–N collocations (‘intense’ and ‘weak’), first without filtering the obtained candidate list and then applying the POS and $NPMI_C$ filters⁵.

Training data

As is common in previous work on semantic collocation classification (Moreno et al., 2013; Wanner et al., 2016), our training set consists of a list of manually annotated correct collocations. For this purpose, we randomly select English nouns from the Macmillan Collocations Dictionary and manually classify their corresponding collocates with respect to the glosses.⁶ Note that there may be more than one collocate for each base. Since collocations with different collocate meanings are not evenly distributed in language (e.g., speakers use more often collocations conveying the idea of ‘intense’ and ‘perform’ than ‘stop performing’), the number of instances per gloss in our training data also varies significantly (see Table 4.5).

For Spanish, the training examples for each of the glosses in our experiments are taken from a corpus in which collocations are classified with respect to LFs. The corpus is the first version of the AnCora-UPF corpus (Mille et al., 2013), which contains 3,513 sentences. When gathering the experimental data, duplicated instances of collocations are removed. When more than one collocate for a given base is found, all collocates are kept. See Table 4.5 for the number of training instances for each gloss.

Test data

A total of 10 bases is evaluated for each gloss. The ground truth test set is created in a similar fashion as the training set: for English, nouns from the Macmillan Dictionary are randomly chosen, and their collocates manually

⁵For the calculation of $NPMI_C$ during the post-processing stage, a seven million sentence newspaper corpus is used for Spanish, and the British National Corpus (BNC) for English.

⁶We consider only collocations that involve single word tokens for both the base and the collocate. In other words, we do not take into account, e.g., phrasal verb collocates such as *stand up*, *give up* or *calm down*. We also leave aside the problem of subcategorization in collocations; cf., e.g., *into* in *take [into] consideration*.

Semantic gloss	Example	# En	# Es
‘intense’	<i>absolute certainty</i>	586	174
‘weak’	<i>remote chance</i>	70	23
‘perform’	<i>give chase</i>	393	319
‘begin to perform’	<i>start [a] career</i>	79	67
‘stop performing’	<i>abandon [a] hope</i>	12	3
‘increase’	<i>improve concentration</i>	73	22
‘decrease’	<i>limit [a] choice</i>	73	16
‘create’, ‘cause’	<i>pose [a] challenge</i>	195	181
‘put an end’	<i>break [the] calm</i>	79	31
‘show’	<i>exhibit [a] characteristic</i>	49	5

Table 4.5: Semantic glosses and size of training sets

classified in terms of the different glosses, until a set of 10 unseen base–collocate pairs is obtained for each gloss. For Spanish, 10 bases for each gloss from those annotated in the corpus are kept for evaluation.

Vector spaces

For Spanish, the only available corpus that is sufficiently large for training the vectors is the Spanish Wikipedia. Therefore, both bases and collocates are modelled by training their word vectors over Wikipedia, more precisely, over a 2014 dump of the Spanish Wikipedia.

For English, bases and collocates are modelled with different corpora. Due to the asymmetric nature of collocations, not all corpora may be equally suitable for the derivation of word embedding representations for both bases and collocates. Thus, we hypothesize that for modelling (nominal) bases, which keep in collocations their literal meaning, a standard register corpus with a small percentage of figurative meanings might be more adequate, while for modelling collocates, a corpus which is potentially rich in collocations would likely be more appropriate. In order to verify this hypothesis, we carry out two different experiments for English. In the first experiment, we use for both bases and collocates vectors pre-trained on the Google News corpus (*GoogleVecs*), which is available at *word2vec*’s website. In the second experiment, the bases are modelled by training their word vectors over a 2014 dump of the English Wikipedia, while for modelling collocates, again, *GoogleVecs* is used. In other words, we assume that Wikipedia is a standard register corpus and thus better for modelling bases, while *GoogleVecs*

is more suitable for modelling collocates. The figures in Section 4.3.3 below will give us a hint whether this assumption is correct.

4.3.3 Outcome of the experiments

The outcome of each experiment is assessed by verifying the correctness of each retrieved candidate from the top-10 candidates obtained for each test base. Given that a base can have different collocates to express a meaning, and neither dictionaries nor corpus contain all possibilities,⁷ the evaluation is not performed automatically against the collocates found in the corpus or dictionary; instead, each candidate is manually judged as correct or incorrect.

For the outcome of each experiment, we compute both precision as the ratio of collocates with the targeted gloss retrieved for each base, and *Mean Reciprocal Rank* (MRR), which rewards the position of the first correct result in a ranked list of outcomes:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

where Q is a sample of experiment runs and rank_i refers to the rank position of the *first* relevant outcome for the i th run. MRR is commonly used in Information Retrieval and Question Answering, but has also shown to be well suited for collocation discovery; see, e.g., (Wu et al., 2010).

We compare the performance of our setup with the accuracy achieved with the technique described in Section 4.2, which serves us as baseline, in two variants: without and with POS+ $NPMI_C$ filters. The first baseline (**S1**) is based on the regularities in word embeddings, with the $\text{vec}(\text{king}) - \text{vec}(\text{man}) + \text{vec}(\text{woman}) = \text{vec}(\text{queen})$ example as paramount case. For collocation retrieval, we follow the same schema; cf., e.g., the gloss ‘perform’ $\text{vec}(\text{take}) - \text{vec}(\text{walk}) + \text{vec}(\text{suggestion}) = \text{vec}(\text{make})$ (where *make* is the collocate to be discovered). The second baseline (**S2**) is an extension of **S1** in that its output is filtered with respect to the valid POS-patterns of targeted collocations and $NPMI_C$.⁸

⁷For this reason, we omit recall

⁸At the first glance, a state-of-the-art approach to correction of collocation errors by suggesting alternative co-occurrences, such as, e.g., (Dahlmeier and Ng, 2011a; Park et al., 2008; Futagi et al., 2008), might appear as a suitable baseline. We discarded this option given that none of them uses explicit fine-grained semantics.

Semantic gloss	Precision (p)			
	S1	S2	S3	S4
‘intense’	0.25	0.12	0.17	0.44
‘weak’	0.00	0.00	0.10	0.45
‘perform’	0.09	0.00	0.20	0.16
‘begin to perform’	0.15	0.00	0.03	0.15
‘stop performing’	0.04	0.04	0.06	0.44
‘increase’	0.20	0.08	0.25	0.50
‘decrease’	0.04	0.09	0.08	0.19
‘create’, ‘cause’	0.07	0.13	0.21	0.20
‘put an end’	0.06	0.05	0.16	0.23
‘show’	0.02	0.00	0.31	0.33

Table 4.6: Precision achieved by the system configurations (S3 – S4) and baselines (S1 –S2) tested on Spanish data

For Spanish, the results of our experiments are shown in Tables 4.6 and 4.7 (**S3** stands for our current transformation matrix-based setup without filtering; and **S4** for the matrix-based setup with POS+ $NPMI_C$ filtering).

For English, we evaluate four different configurations of our technique against the two baselines. The four configurations that we test are: **S3**, which is based on the transition matrix for which *GoogleVecs* is used as reference vector space representation for both bases and collocates; **S4**, which applies POS-pattern and $NPMI_C$ filters to the output of **S3**; **S5**, which is equivalent to **S3**, but relies on a vector space representation derived from Wikipedia for learning bases projections and on a vector space representation from *GoogleVecs* for collocate projections; and, finally, **S6**, where the **S5** output is, again, filtered by POS collocation patterns and $NPMI_C$. The results of the experiments are displayed in Tables 4.8 and 4.9.

4.3.4 Discussion

In general, the full pipeline promotes good collocate candidates to the first positions of the ranked result lists and is also best, in terms of accuracy, compared to the system configurations that do not exploit POS+ $NPMI_C$ filtering. For English, the configurations S3 – S6 of the system largely outperform the baselines, with the exception of the gloss ‘increase’, for which S2 equals S6 as far as p is concerned. However, in this case too MRR

Semantic gloss	MRR			
	S1	S2	S3	S4
‘intense’	0.52	0.10	0.31	0.42
‘weak’	0.00	0.00	0.75	0.60
‘perform’	0.19	0.00	0.44	0.25
‘begin to perform’	0.29	0.00	0.04	0.08
‘stop performing’	0.10	0.07	0.35	0.53
‘increase’	0.51	0.17	0.63	0.67
‘decrease’	0.07	0.10	0.35	0.43
‘create’, ‘cause’	0.38	0.13	0.57	0.38
‘put an end’	0.26	0.02	0.32	0.43
‘show’	0.20	0.00	0.85	0.55

Table 4.7: Mean Reciprocal Rank achieved by the system configurations (S3 – S4) and baselines (S1 – S2) tested on Spanish data

Semantic gloss	Precision (p)					
	S1	S2	S3	S4	S5	S6
‘intense’	0.08	0.43	0.04	0.50	0.24	0.72
‘weak’	0.09	0.11	0.23	0.45	0.27	0.39
‘perform’	0.05	0.17	0.01	0.06	0.13	0.40
‘begin to perform’	0.03	0.08	0.24	0.30	0.22	0.38
‘stop performing’	0.00	0.00	0.11	0.15	0.12	0.20
‘increase’	0.16	0.53	0.31	0.43	0.35	0.53
‘decrease’	0.07	0.05	0.28	0.25	0.27	0.28
‘create’, ‘cause’	0.10	0.16	0.01	0.15	0.14	0.53
‘put an end’	0.05	0.09	0.15	0.20	0.08	0.25
‘show’	0.10	0.55	0.24	0.49	0.49	0.70

Table 4.8: Precision achieved by the system configurations (S3 – S6) and baselines (S1 – S2) tested on English data

is considerably higher for S6, which achieves the highest MMR scores for 6 and the highest precision scores for 7 out of 10 glosses (see the S6 columns in Tables 4.8 and 4.9). For Spanish, configurations S3 – S4 also outperform the baseline, except for the gloss ‘begin to perform’, in which S1 achieves a p comparable to that of S4, and a higher MRR than both S3 and S4 (Tables 4.6 and 4.7).

As we can observe from the number of instances in Table 4.5, certain glosses seem to exhibit less linguistic variation, requiring a less populated trans-

Semantic gloss	MRR					
	S1	S2	S3	S4	S5	S6
‘intense’	0.18	0.35	0.35	0.15	0.66	0.82
‘weak’	0.31	0.15	0.69	0.64	0.61	0.47
‘perform’	0.22	0.32	0.01	0.35	0.70	0.79
‘begin to perform’	0.17	0.05	0.61	0.64	0.70	0.71
‘stop performing’	0.01	0.00	0.90	0.66	0.71	0.65
‘increase’	0.47	0.72	0.78	0.86	0.86	0.90
‘decrease’	0.18	0.04	0.57	0.38	0.37	0.30
‘create’, ‘cause’	0.41	0.23	0.11	0.11	0.48	0.58
‘put an end’	0.28	0.10	0.38	0.36	0.33	0.38
‘show’	0.44	0.54	0.87	0.82	0.73	0.81

Table 4.9: Mean Reciprocal Rank achieved by the system configurations (S3 – S6) and baselines (S1 – S2) tested on English data

Semantic gloss	S6
‘intense’	0.82
‘weak’	0.45
‘perform’	0.40
‘begin to perform’	0.42
‘stop performing’	0.22
‘increase’	0.55
‘decrease’	0.37
‘create’, ‘cause’	0.59
‘put an end’	0.43
‘show’	0.85

Table 4.10: Precision of the coarse-grained evaluation of the English S6 configuration

formation function from bases to collocates. Consider the case of ‘show’, for English, which generates with only 49 training pairs the second best transition matrix, with $p=0.70$. Similarly, for Spanish, for example, the transformation function of ‘stop performing’, trained with only 3 instances, achieves the second best results both for p and MRR.

Comparing S1, S3, S5 to S2, S4, and S6, we may conclude that the inclusion of a filtering module contributes substantially to the overall precision in nearly all cases (‘decrease’ being the only exception for English). The comparison of the precision obtained for configurations S3 and S5 also reveals

Semantic gloss	Base	Retrieved candidates
'intense'	<i>velocidad</i> 'speed'	<i>alto</i> 'high', <i>máximo</i> 'maximum' <i>constante</i> 'constant', <i>gran</i> 'great' <i>considerable</i> 'considerable' <i>vertiginoso</i> 'vertiginous'
'weak'	<i>plazo</i> 'period'	<i>breve</i> 'brief', <i>corto</i> 'short' <i>largo</i> 'long', <i>prorrogable</i> 'extendable'
'perform'	<i>viaje</i> 'trip'	<i>hacer</i> 'make', <i>embarcar</i> 'load' <i>efectuar</i> 'carry out', <i>realizar</i> 'make' <i>iniciar</i> 'initiate', <i>preparar</i> 'prepare' <i>topar</i> 'bump into'
'begin to perform'	<i>éxito</i> 'success'	<i>alcanzar</i> 'attain', <i>medir</i> 'measure' <i>suponer</i> 'suppose', <i>rebasar</i> 'overflow' <i>propiciar</i> 'propiciate', <i>presumir</i> 'boast' <i>presagiar</i> 'foretell'
'stop performing'	<i>escondite</i> 'hiding place'	<i>abandonar</i> 'abandon'
'increase'	<i>producción</i> 'production'	<i>incentivar</i> 'incentive', <i>fomentar</i> 'foster' <i>promover</i> 'promote', <i>alentar</i> 'encourage' <i>potenciar</i> 'improve', <i>fortalecer</i> 'strengthen'
'decrease'	<i>pérdida</i> 'loss'	<i>reducir</i> 'reduce', <i>moderar</i> 'moderate' <i>frenar</i> 'brake', <i>compensar</i> 'compensate' <i>disminuir</i> 'decrease', <i>elegir</i> 'increase'
'create', 'cause'	<i>templo</i> 'temple'	<i>construir</i> 'build', <i>erigir</i> 'erect' <i>levantar</i> 'raise', <i>edificar</i> 'build' <i>derribar</i> 'demolish'
'put an end'	<i>duda</i> 'doubt'	<i>resolver</i> 'solve', <i>solventar</i> 'resolve' <i>plantear</i> 'set out', <i>zanjar</i> 'settle'
'show'	<i>opinión</i> 'opinion'	<i>expresar</i> 'express', <i>manifestar</i> 'manifest' <i>reflejar</i> 'reflect', <i>resumir</i> 'summarize' <i>plasmear</i> 'express', <i>exponer</i> 'expound'

Table 4.11: Examples of retrieved Spanish collocations

Semantic gloss	Base	Retrieved candidates
‘intense’	<i>caution</i>	<i>extreme</i>
‘weak’	<i>change</i>	<i>slight, little, modest, minor, noticeable, minimal, sharp, definite, small, big</i>
‘perform’	<i>calculation</i>	<i>produce, carry</i>
‘begin to perform’	<i>cold</i>	<i>catch, get, run, keep</i>
‘stop performing’	<i>career</i>	<i>abandon, destroy, ruin, terminate, threaten, interrupt</i>
‘increase’	<i>capability</i>	<i>enhance, increase, strengthen, maintain, extend, develop, upgrade, build, provide</i>
‘decrease’	<i>congestion</i>	<i>reduce, relieve, cut, ease, combat</i>
‘create’, ‘cause’	<i>challenge</i>	<i>pose</i>
‘put an end’	<i>ceasefire</i>	<i>break</i>
‘show’	<i>complexity</i>	<i>demonstrate, reveal, illustrate, indicate, reflect, highlight, recognize, explain</i>

Table 4.12: Examples of retrieved English collocations

that for 7 glosses the strategy to model collocates and bases on different corpora paid off. This is different as far as MRR is concerned. Further investigation is needed for the examination of this discrepancy.

It is also informative to contrast the performance on pairs of glosses with opposite meanings, such as e.g., ‘begin to perform’ vs. ‘stop performing’; ‘increase’ vs. ‘decrease’; ‘intense’ vs. ‘weak’; and finally ‘create, cause’ vs. ‘put an end’. Better performance is achieved consistently on the *positive* counterparts (e.g., ‘begin to perform’ over ‘stop performing’). A closer look at the output reveals that in these cases positive glosses are persistently classified as negative, but the opposite also occurs. Consider the following examples as illustration:

- (1) *voz tenue* ‘faint voice’ (belongs to ‘weak’ instead of ‘intense’)
- (2) *fuerte tensión* ‘strong tension’ (belongs to ‘intense’ instead of ‘weak’)

- (3) *aumentar* [*una*] *tasa* ‘to increase [a] rate’ (belongs to ‘increase’ instead of ‘decrease’)
- (4) *derribar* [*un*] *templo* ‘to demolish [a] temple’ (belongs to ‘put an end’ instead of ‘create’, ‘cause’)
- (5) *plantear* [*una*] *duda* ‘to raise [a] question’ (belongs to ‘create’, ‘cause’, instead of ‘put an end’)

This occurs with both the English and Spanish data. Further research is needed to first understand why this is the case and then to come up with an improvement of the technique in particular on the *negative* glosses.

The fact that for some of the glosses precision is rather low may be taken as a hint that the proposed technique is not suitable for the task of semantics-oriented recognition of collocations. However, it should be also stressed that our evaluation is very strict: a retrieved collocate candidate is considered as correct only if it forms a collocation with the base, and if it belongs to the target semantic gloss. In particular the first condition might be too rigorous, given that, in some cases, there is a margin of doubt whether a combination is a free co-occurrence or a collocation; cf., e.g., *huge challenge* or *reflect* [*a*] *concern*, which are rejected as collocations in our evaluation. Since for L2 learners such co-occurrences may be also useful, we carry out a second evaluation in which all the suggested collocate candidates that belong to a target semantic gloss are considered as correct, even if they do not form a collocation.⁹ Cf. Table 4.10 for the outcome of this evaluation for the S6 configuration. Only for ‘perform’ the precision remains the same as before. This is because collocates assigned to this gloss are support verbs (and thus void of own lexical semantic content).

Tables 4.11 and 4.12 show some of the collocates that are obtained.

4.4 Summary and conclusions

In Section 4.2, we presented an unsupervised technique for collocation discovery and classification, based on the ability of word embeddings for drawing analogies. With an average precision of 0.78, the system performs well

⁹Obviously, collocate candidates are considered as incorrect if they form incorrect collocations with the base. Examples of such incorrect collocations are *stop* [*the*] *calm* and *develop* [*a*] *calculation*.

as far as the retrieval of collocates is concerned. However, for the semantic classification of collocates, the performance varies greatly. For instance, a precision of up to 0.80 is obtained for the gloss ‘intense’, which means that the system is able to perform satisfactorily for some of the semantic glosses that are tested. Still, much room for improvement is left for other glosses, such as ‘weak’ or ‘decrease’, and the impact of the chosen seed example on the collocates retrieved by the system is yet to be investigated. Nevertheless, the fact that starting from just one example we are able to obtain highly idiosyncratic collocations such as *desencadenar* [una] crisis ‘to trigger [a] crisis’, *demoler* [un] edificio ‘to demolish [a] building’ or *fortalecer* [el] liderazgo ‘to strengthen leadership’, as well as less idiosyncratic ones such as *reducir* [un] salario ‘to reduce [a] salary’ or *incrementar* [la] cobertura ‘to increase coverage’ shows the potential of the approach.

In Section 4.3, we presented a weakly-supervised algorithm for collocation retrieval and classification. The technique is based on the linear relations that are held between semantically similar words in different vector spaces. Experiments on both Spanish and English show that this approach behaves better than the baseline (the technique described in Section 4.2), outperforming it in nearly all cases, even when the size of the training set is very limited, i.e., below 10 instances. This suggests that a small investment in annotation can lead to better resources.

The experiments presented in Section 4.3 also point to two important conclusions: (1) not all corpora are equally suitable for obtaining word embedding representations for bases and collocates. A standard register corpus with definitorial semantic language is more appropriate for modelling bases, and a corpus richer in collocations is more appropriate for modelling collocates. (2) a $NPMI_C$ filtering module, which helps to disregard candidates that do not co-occur with the base, contributes substantially to the overall precision in almost all of the cases.

We focus on Spanish and English, and only on a small amount of collocations. However, since the resources required by both approaches, i.e., the example-based and the weakly-supervised approaches, are easily obtained, the proposed approaches are highly scalable and portable to other languages. Given the lack of semantically tagged collocation resources for most languages, our work has the potential to become influential, especially in the context of second language learning.

Collocation error classification

The second of our goals for the present thesis work is, as stated in Chapter 1, the development of computational techniques for the automatic identification and classification of collocation errors. We argue that specific techniques are necessary for the identification of each type of error because, on the one hand, more accurate type-targeted error correction techniques can be developed and, on the other, information about the error types the learners struggle more with can be given to the learners. In our work, we approach the identification and classification of collocation errors as two different tasks, namely (1) collocation error classification, which consists in classifying incorrect collocations that have been manually identified as such, and (2) simultaneous identification and classification of collocation errors in learners' writings. For the first task, we propose a hybrid approach that takes as input erroneous collocations and classifies them according to their error type(s). The second task is left for the next chapter. The remainder of this chapter is organized as follows: Section 5.1 describes the method for collocation error classification, along with the experiments and the discussion, and Section 5.2 presents some conclusions.

5.1 Collocation error classification

In this Section, we present a hybrid approach to collocation error classification based on Machine Learning classification techniques and rule-based techniques. Section 5.1.1 justifies the use of a hybrid approach, and presents the techniques. Section 5.1.2 describes the experimental setup, and Sections 5.1.3 and 5.1.4 present the results and the discussion.

5.1.1 A hybrid approach to collocation error classification

In corpus-based linguistic phenomenon classification, it is common to choose a supervised machine learning method that is used to assign any identified phenomenon to one of the available classes. In the light of the diversity of the linguistic nature of the collocation errors and the widely diverging frequency of the different error types, this procedure seems not optimal for miscollocation classification. A round of preliminary experiments confirmed this assessment. It is more promising to target the identification of each collocation error type separately, using for each of them the identification method that suits its characteristics best. For this reason, we propose a hybrid approach, which uses Machine Learning classification algorithms along with rules to classify the different types of errors.

In our work, we use the collocation error typology by Alonso Ramos et al. (2010), described in detail in Chapter 2. Concretely, the types of errors that we address are presented in Section 2.3.4, at the end of Chapter 2. As a summary, we present below for the convenience of the reader the lexical and grammatical errors types that we target in our work, along with a brief description and an example for each type.

Lexical errors

- **Extended substitution:** Includes **Substitution errors** (erroneous choice of the base or collocate, as in **realizar una meta*, lit. ‘to realize, to carry out a goal’; corr.: *conseguir una meta*, lit. ‘to reach a goal’) and **Analysis errors** (creation of an erroneous word combination with the form of a collocation, as in **hacer nieve*, lit. ‘to make snow’; corr.: *nevar*, lit. ‘to snow’).
- **Creation:** Erroneous choice of a non-existing base or collocate, as in **hacer un llamo*, lit. ‘make a *llamo* [non-existing word meaning *call*]’; corr.: *hacer una llamada*, lit. ‘make a call’.
- **Different sense:** Erroneous choice of a correct collocation in a particular context, as in **voz alta*, lit. ‘loud voice’; corr.: *gran voz*, lit. ‘great voice’. Both are correct collocations, but while the former refers to ‘loud voice’, as opposed to ‘silent voice’, *gran voz* refers to ‘shouting’, to ‘an extraordinarily loud voice’.

Grammatical errors

- **Government:** Erroneous omission, insertion or use of the government of any of the members of the collocation, as in **ver a la película*, lit. ‘watch

at a movie’; corr.: *ver la película*, lit. ‘watch a movie’.

- **Determiner:** Erroneous omission or insertion of a determiner of the nominal base, as in **hablar el inglés*, lit. ‘speak the English’; corr.: *hablar inglés*, lit. ‘speak English’.
- **Pronoun:** Erroneous omission or insertion of a reflexive pronoun for a verbal collocate, as in **volver loco*, lit. ‘turn (someone) crazy’; corr.: *volverse loco*, lit. ‘turn (oneself) crazy’.
- **Gender:** Erroneous gender, as in **aumentar las precios*, lit. ‘raise the[fem] prices’; corr.: *aumentar los precios*, lit. ‘raise the[masc] prices’.
- **Number:** Erroneous number of the base or the base determiner, as in **dar bienvenidas*, lit. ‘give welcomes’; corr.: *dar la bienvenida*, lit. ‘give the welcome’.
- **Order:** Erroneous word order, as in **educación buena*, lit. ‘education good’; corr.: *buena educación*, lit. ‘good education’.

5.1.1.1 Lexical errors

In what follows, we describe the methods that we use to identify lexical miscollocations. All of these methods perform a binary classification of all identified incorrect collocations as ‘of type X’ / ‘not of type X’. The methods for the identification of ‘Extended substitution’ and ‘Creation’ errors receive as input the incorrect collocations (i.e., grammatical, lexical or register-oriented miscollocations) recognized in the writing of a language learner by a collocation error recognition program¹, together with their sentential contexts. The method for the recognition of ‘Different sense’ errors receives as input ‘Different sense’ errors along with the correct collocations identified in the writing of the learner.

Extended Substitution Error Classification. For the classification of incorrect collocations as ‘Extended substitution’ error / ‘Not an extended substitution’ error, we use supervised machine learning. This is because ‘Extended substitution’ is, on the one side, the most common type of error (such that sufficient training material is available), and, on the other side,

¹Since in our experiments we focus on miscollocation classification, we use as “writings of language learners” a learner corpus in which both correct and incorrect collocations have been annotated manually and revised by different annotators. Only those instances for which complete agreement is found is used for the experiments.

very heterogeneous (such that it is difficult to be captured by a rule-based procedure). After testing various ML-approaches, we have chosen the Support Vector Machine (SMO) implementation from the Weka toolkit (Hall et al., 2009).²

Two different types of features have been used: lexical features and co-occurrence (or *PMI*-based) features. The lexical features consist of the lemma of the collocate and the bigram made up of the lemmas of the base and collocate. The *PMI*-based features consist of: $NPMI_C$ of the base and the collocate, $NPMI_C$ of the hypernym of the base and the collocate, $NPMI$ (see Appendix A) of the base and its context, and $NPMI$ of the collocate and its context, considering as context the two immediate words to the left and to the right of each element. Hypernyms are taken from the Spanish WordNet; $NPMIs$ and $NPMI_Cs$ are calculated on a 7 million sentences reference corpus of Spanish. Table 5.1 summarizes the features used in the experiments.

Features	
Lexical	lemmaCollocate lemmaBase + lemmaCollocate
Statistical	$NPMI_C$ (base,collocate) $NPMI_C$ (base_hyponym,collocate_hyponym) $NPMI$ (base,baseContext) $NPMI$ (collocate,collocateContext)

Table 5.1: Features for the classification of ‘Extended Substitution’ errors.

Creation Error Classification. For the detection of ‘Creation’ errors among all miscollocations, we have designed a rule-based algorithm that uses linguistic (lexical and morphological) information; see Algorithm 2.

If both elements of a collocation under examination are found in the RC with a sufficient frequency (≥ 50 for our experiments), they are considered valid tokens of Spanish, and therefore ‘Not creation’ errors. If one of the elements has a low frequency in the RC (< 50), the algorithm continues to examine the miscollocation. First, it checks whether a learner has used an English word in a Spanish sentence, considering it as a ‘transfer Creation’ error. If this is not the case, it checks whether the gender suffix is wrong,

²Weka is University of Waikato’s public machine learning platform that offers a great variety of different classification algorithms for data mining.

Algorithm 2: ‘Creation’ Error Classification

```

Input:
1   $b + c$  // Set of base+collocate pairs;
2   $b_L / c_L$  // Set of lemmatized bases/collocates;
3   $b_r / c_r$  // Set of stems of bases/collocates;
   initialization;
4  for  $b + c$  do
5  |   if  $b_L, c_L \in RC$ 
   |   and  $freq('b_L') > 50$ 
   |   and  $freq('c_L') > 50$  then
6  |   |   echo “Not a creation error”;
7  |   else if  $b_L \vee c_L \in \text{English dictionary}$  then
8  |   |   echo “Creation error (Transfer)”;
9  |   else if  $check\_gender(b_L) = false$  then
10 |   |   echo “Creation error (Incorrect gender)”;
11 |   else if  $check\_affix(b_r) \parallel check\_affix(c_r)$  then
12 |   |   echo “Creation error (Incorrect derivation)”;
13 |   else if  $check\_ortography(b_L) \parallel check\_ortography(c_L)$  then
14 |   |   echo “Not a creation error (Ortographic)”;
15 |   else if  $freq('b_L') > 0$  or  $freq('c_L') > 0$  then
16 |   |   echo “Not a creation error”;
17 |   else
18 |   |   echo “Creation error (Unidentified)”
   end

```

considering it as a ‘gender Creation’ error, as in, e.g., **hacer regala* instead of *hacer regalo*, lit. ‘make present’. This is done by alternating the gender suffix and checking the resulting token in the RC.

If no gender-influenced error can be detected, the algorithm checks whether the error is due to an incorrect morphological derivation of either the base or the collocate — which would imply a ‘derivation Creation’ error, as in, e.g., **ataque terrorístico* instead of *ataque terrorista* ‘terrorist attack’. For this purpose, the stems of the collocation elements are obtained and expanded by the common nominal / verbal derivation affixes of Spanish to see whether any derivation leads to the form used by the learner.

Should this not be the case, the final check is to see whether any of the elements is misspelled and therefore we face a ‘Not creation error’. This is done by calculating the edit distance from the given forms to valid tokens in the RC.

In the case of an unsuccessful orthography check, we assume a ‘Creation’ error if the frequency of one of the elements of the miscollocation is ‘0’, and a ‘Not creation’ error for element frequencies between ‘0’ and ‘50’.

Different Sense Error Classification. Given that ‘Different sense’ errors capture the use of correct collocations in an inappropriate context, the main strategy for their detection is to compare the context of a learner collocation with its prototypical context. The prototypical context is represented by a centroid vector calculated using the lexical contexts of the correct uses of the collocation found in the RC.

The vector representing the original context is compared to the centroid vector in terms of cosine similarity; cf. Eq. (5.1).

$$\text{sim}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (5.1)$$

A specific similarity threshold must be determined in order to discriminate correct and incorrect uses. Although further research is needed to design a more generic threshold determination procedure, in the experiments we carried out so far, 0.02543 was empirically determined as the best fitting threshold.

5.1.1.2 Grammatical errors

For the classification of grammatical errors, we have developed a set of functions. Each function focuses on the identification of one specific type of grammatical error in given miscollocations. Each function has thus been designed taking into account both the specific characteristics of the type of error it deals with and the possibility of a collocation being affected by several errors at the same time, either grammatical, lexical, register or any combination of them. All six functions receive as input miscollocations found in writings by learners of Spanish, and most of them use a RC of Spanish. In what follows, we briefly describe each one of them.

Determination errors. This function queries the RC to look up common occurrences of both the base and the collocate of the miscollocation, includ-

ing those with the presence of a determiner and those in which no determiner is found. If the number of occurrences with the determiner is significantly higher than the number of occurrences without the determiner, the collocation is considered to require a determiner. In this case, if the context of the miscollocation does not contain a determiner, a ‘Determination’ error is flagged. Along the same lines, if it is determined that the collocation does not take a determiner, but the learner uses one, again, a ‘Determination’ error is flagged.

Number errors. ‘Number’ errors can affect both the base and the collocate and are not necessarily manifested in terms of a lack of agreement, as, e.g., in **tener una vacación* ‘to have a holiday’, **dimos bienvenidas* ‘to welcome’, **gané pesos* ‘to put on weight’, etc. In order to check whether a collocation contains a ‘Number’ error, the corresponding function retrieves from the RC combinations of the lemmas of the base, the collocate and the prepositions that depend on the dependent element. In other words, given a preposition, all possible combinations of the forms of the base and the collocate with that particular preposition are retrieved. Then, alternative number forms of the base and collocate are generated (i.e., if an element in the miscollocation is in plural, its singular form is generated, and vice versa) and occurrences of their combinations are retrieved from the RC.

If the original form is not one of the possible combinations retrieved from the RC, but any of the alternatives is, the miscollocation is assumed to contain a ‘Number’ error.

Gender errors. Only miscollocations that have a noun as their base can contain this kind of error. However, the form of the base is rarely erroneous (cf., e.g., **pasar los vacaciones*). Rather, there is often a lack of agreement between the base and its determiner, or between the base and the collocate (in N-Adj collocations), resulting from the wrong choice of the gender of the determiner respectively collocate. For this reason, the corresponding function checks the gender of the determiner and adjectival modifiers of the base of the given miscollocation. Both the frequency of the miscollocation *n*-gram (i.e., string consisting of the collocate and the base with its determiner) and linguistic information are considered. For each miscollocation, the function retrieves from the RC the frequency of the original *n*-gram. Then, it generates new alternatives by changing the gender of the determiner (in V-N, N-N or prepositional collocations) or the adjective (in N-Adj collocations) and looks for the frequency of the new combinations. If this happens to be higher than the frequency of the miscollocation, a

‘Gender’ error is assumed. Otherwise, the agreement between the base and the determiner respectively collocate is checked. If lack of agreement is identified, a ‘Gender’ error is assigned.

Government errors. To identify this kind of error, we take into account the context in which the miscollocation appears. For this purpose, first, syntactic patterns that contain the miscollocation’s base and collocate and any preposition governed by either of the two are retrieved from the RC. Then, it is looked up whether the original syntactic miscollocation pattern that involves a governed preposition appears in the retrieved list. If this is not the case, the miscollocation is assumed to contain a ‘Government’ error.

Pronoun errors. In order to identify ‘Pronoun’ errors, a similar approach to the one used for recognizing ‘Determination’ errors is followed. In this case, frequencies of the combinations with and without reflexive pronouns are retrieved and compared to the miscollocation.

Order errors. To identify an ‘Order’ error, the frequency of the given miscollocation in the RC is calculated. Then, the frequencies of all the possible permutations of the elements of the collocation are compared to the frequency of the miscollocation. If any of them is significantly higher, the collocation is considered to contain an ‘Order’ error.

5.1.2 Experimental setup

For our experiments, we use a fragment of the Spanish Learner Corpus CEDEL2 (Lozano, 2009), which is composed of writings of learners of Spanish whose first language is American English. The writings have an average length of 500 words and cover different genres. Opinion essays, descriptive texts, accounts of some past experience, and letters are the most common of them. The levels of the students range from ‘low-intermediate’ to ‘advanced’. In the fragment of CEDEL2 (in total, 517 writings) that we use (our working corpus), both the correct and incorrect collocation occurrences are tagged.³ As stated above, collocations were annotated and revised, and only those for which a general agreement regarding their status was found, are used for the experiments.

³The tagging procedure has been carried out manually by several linguists. The first phase of it was already carried out by Alonso Ramos et al. (2010). We carried on the tagging work by Alonso Ramos et al. (2010) to have for our experiments a corpus of a sufficient size.

Table 5.2 shows the frequency of the correct collocations and of the five types of lexical miscollocations in our working corpus. The numbers confirm our decision to discard synthesis miscollocations (there are only 9 of them – compared to, e.g., 565 substitution miscollocations) and to merge analysis miscollocations (19 in our corpus) with substitution miscollocations.⁴

Class	# Instances
Correct collocations	3245
Substitution errors	565
Creation errors	69
Different sense errors	48
Analysis errors	19
Synthesis errors	9

Table 5.2: Number of instances of the lexical error types and correct collocations in CEDEL2.

Analogously, Table 5.3 shows the frequency of the different types of grammatical errors. With only 2 instances of governed errors and one instance of specification error, our decision to disregard these two types of errors seems appropriate.

Class	# Instances
Government errors	225
Determination errors	146
Gender errors	77
Number errors	44
Pronoun errors	28
Order errors	28
Governed errors	2
Specification errors	1

Table 5.3: Number of instances of the grammatical error types in CEDEL2.

To be able to take the syntactic structure of collocations into account, we processed CEDEL2 with Bohnet (2010)'s syntactic dependency parser⁵.

⁴Synthesis miscollocations are too different from the other types of errors to be merged with any other type.

⁵Processing tools' performance on non-native texts is lower than on texts written by natives. We evaluated the performance of the parser on our learner corpus and obtained the following results: LAS:88.50%, UAS:87.67%, LA:84.54%.

As RC, we use a seven million sentence corpus, from Peninsular Spanish newspaper material. The RC was also processed with Bohnet (2010)’s syntactic dependency parser.

5.1.3 Outcome of the experiments

Table 5.4 shows the performance of the individual lexical error classification methods. In the ‘+’ column of each error type, the accuracy is displayed with which our algorithms correctly detect that a miscollocation belongs to the error type in question; in the ‘-’ column, the accuracy is displayed with which our algorithms correctly detect that a miscollocation does not belong to the corresponding error type.

	‘Ext. subst’		‘Creation’		‘Diff. sense’	
	+	-	+	-	+	-
Baseline	0.395	0.902	0.391	0.986	0.5	0.453
Our model	0.832	0.719	0.681	0.942	0.583	0.587

Table 5.4: Accuracy of the lexical error detection systems.

To assess the performance of our classification, we use three baselines, one for each type of error. To the best of our knowledge, no other state-of-the-art figures are available with which we can compare its quality further. For the ‘Extended substitution’ miscollocation classification, we use as baseline a simplified version of the model, trained only with one of our lexical features, namely bigrams made up of the lemmas of the base and the collocate of the collocation. For ‘Creation’ miscollocation classification, the baseline is an algorithm that judges a miscollocation to be of the type ‘Creation’ if either one of the elements (the lemma of the base or of the collocate) or both elements of the miscollocation are not found in the RC. Finally, for the ‘Different sense’ miscollocation classification, we take as baseline an algorithm that, given a bag of the lexical items that constitute the contexts of the correct uses of a collocation in the RC, judges a collocation to be a miscollocation of the ‘Different sense’ type if less than half of the lexical items of the context of this collocation in the writing of the learner is not found in the reference bag.

Table 5.5 shows the classification accuracy of the individual grammatical error identification functions for both the positive cases (collocations containing the type of error that is to be identified) and the negative cases

(incorrect collocations affected by any kind of error, except the one that is dealt with).

Type of error	(+)	(-)
Determination	0.719	0.793
Number	0.659	0.851
Gender	0.818	0.989
Government	0.68	0.708
Pronoun	0.357	0.99
Order	0.75	0.848

Table 5.5: Accuracy of the grammatical error detection functions

5.1.4 Discussion

Before we discuss the outcome of the experiments, let us briefly make some generic remarks on the phenomenon of a collocation in the experiments.

5.1.4.1 The phenomenon of a collocation

The decision whether a collocation is correct or incorrect is not always straightforward, even for native expert annotators. Firstly, a certain number of collocations is affected by spelling and inflection errors. Consider, e.g., *tomamos cervezas* ‘we drank beer’, instead of *cervezas*; *sacque una mala nota* ‘I got a bad mark’, where *sacqué* is the right form, or *el dolor disminúe* ‘the pain decreases’, instead of *disminuye*. In such cases, we assume that these are orthographical or morphological mistakes, rather than collocational ones. Therefore, we consider them to be correct. On the other hand, collocations may also differ in their degree of acceptability. Consider, e.g., *asistir a la escuela*, *tomar una fotografía* o *mirar la televisión*. Collocations that were doubtful to one or several annotators were looked up in a RC. If their frequency was higher than a certain threshold, they were annotated as correct. Otherwise, they were considered incorrect. From the above examples, *asistir a la escuela* is the only collocation considered as correct after the consultation of the RC.

5.1.4.2 The outcome of the experiments

After these considerations, we now discuss the outcome of the experiments, for each of the error types that we target.

Extended substitution errors. Especially in the case of ‘Extended substitution’ and ‘Creation’ miscollocations, the classification accuracy in our experiments has been rather high. Still, the analysis of the incorrectly classified miscollocations is certainly beneficiary for further improvements. For instance, some of the ‘Extended substitution’ miscollocations have not been recognized; cf. (5.1–5.3). Also, some miscollocations that are not of the type ‘Extended substitution’ are classified to be of this type. Consider, e.g., (5.4) and (5.5), both ‘Creation’ errors incorrectly classified as ‘Extended substitution’. Lexical features play an important role in these instances, since all these cases are possibly influenced by the fact that their collocates are rather frequent in correct non-substitution collocations.

(5.1) **rato elevado*, instead of *tasa elevada* ‘high rate’

(5.2) **tener moralidad*, instead of *tener moral* ‘to have moral’

(5.3) **comer café*, instead of *tomar café* ‘to eat coffee’

(5.4) **la data revela*, instead of *los datos revelan* ‘the data reveal’

(5.5) **arena orada*, instead of *arena dorada* ‘golden sand’

Creation errors. Cases of ‘Creation’ miscollocations that are not recognized as such include (5.6–5.8). (5.6) and (5.7). The failure to recognize them is likely to be due to the noise in the RC created by the lemmatizer, since the frequencies of *secto* and *ético* are 145 and 179, respectively.

In other cases, such as in (5.8), our algorithm is not able to find an origin of the possible error, judging it thus to be ‘Not creation’.

(5.6) **formar [un] secto*, instead of *formar [una] secta* ‘to create [a] sect’

(5.7) **tener [un] ético*, instead of *tener ética* ‘to have [an] ethic’

(5.8) **hablar creol*, instead of *hablar criollo* ‘to speak Creol’

Some miscollocations that are not of the ‘Creation’ type have been classified as being ‘Creation’; cf., e.g., (5.9–5.12). All four are, in fact, ‘Substitution’ errors, which have been classified as ‘Creation’. One possible reason for this failure might be the journalistic genre of our RC. Thus, *picnic* ‘picnic’ and *conteo* ‘count’ are perfectly acceptable words in Spanish. However, they are not frequent in journalistic writings of our RC. *Picnic* appears 47 times and *conteo* 29 times. Their low frequency leads them to be considered

‘Creation’ errors: *picnic* is an English word, such that the system classified its collocation as ‘Creation by transfer’ error. *Conteo* is assumed to be a ‘Creation’ error because its affix is interpreted incorrectly. *Indisputable* is also a correct word in Spanish, but it does not appear in our RC at all, such that its collocation is classified as ‘Creation by transfer’ error.

Other misclassifications with respect to ‘Creation’ are due to errors during lemmatization. This is, for instance, the case of (5.12), where *caminata* ‘hike’ has been lemmatized as the infinitive form *caminatar*, a non-existing word in Spanish.

(5.9) **tener [un] picnic*, instead of *hacer [un] picnic* ‘to have [a] picnic’

(5.10) **derecho indisputable*, instead of *derecho indiscutible* ‘indisputable right’

(5.11) **hacer [un] conteo*, instead of *hacer [un] conteo* ‘to take stock, to count’

(5.12) **hacer caminata*, instead of *ir de caminata* ‘to go [for a] hike’

Different sense errors. The performance figures show that the correct identification of ‘Different sense’ miscollocations is still a challenge. With an accuracy somewhat below 60% for both the recognition of ‘Different sense’ miscollocations and recognition of ‘Correctly used’ collocations, there is room for improvement. Our cosine-measure quite often classifies ‘Different sense’ errors as correctly used collocations, such as (5.13 and 5.14), or leads to the classification of correct collocations as ‘Different sense’ miscollocations (cf., e.g., 5.15–5.18). This shows the limitations of an exclusive use of lexical contexts for the judgement whether a collocation is appropriately used: on the one hand, lexical contexts can, in fact, be rather variant (such that the learner may use a collocation correctly in a novel context), and, on the other hand, lexical contexts do not capture the *situational* contexts, which determine even to a major extent the appropriateness of the use of a given expression. Unfortunately, the capture of situational contexts remains a big challenge.

(5.13) *gastar [el] tiempo*, instead of *pasar [el] tiempo* ‘to spend time’

(5.14) *tener opciones*, instead of *ofrecer posibilidades* ‘to offer possibilities’

(5.15) *ir [en] coche* ‘to go [by] car’

(5.16) *tener [una] relación* ‘to have [a] relationship’

(5.17) *tener impacto* ‘to have impact’

(5.18) *tener capacidad* ‘to have capacity’

Determination errors. As illustrated in the examples (5.19–5.21), some ‘Determination’ errors are not identified as such because these collocations can be found both with and without determiner, depending on the context. For instance, a determiner can be required by a specifier, as in (5.19). Also, we find a singular form of the collocation with a determiner, as in (5.20) and (5.21), where *tener un hijo* ‘to have a child’ and *hacer una actividad* ‘to do a task’ are correct.

(5.19) **tiene una reputación*, instead of *tiene reputación* ‘he has a reputation’

(5.20) **tiene los hijos*, instead of *tiene hijos* ‘he has children’

(5.21) **hace las actividades*, instead of *hace actividades* ‘he does tasks’

With regard to the negative case, i.e., the classification of miscollocations that contain other kinds of errors than ‘Determination’ error, the same reasons can be identified as the source of error. In the following examples, the forms including a determiner, i.e., the singular forms, are more frequent than the forms that do not have it, such that they are classified as ‘Determination’ errors.

(5.22) **tengo planes*, instead of *tengo planes (de)* ‘I have plans’

(5.23) **dijo secretos*, instead of *contar, revelar secretos* ‘he told secrets’

(5.24) **hacer decisiones*, instead of *tomar decisiones* ‘to take decisions’

Number errors. Most failures to identify a ‘Number’ error are due to the fact that, because of multiple errors appearing in the collocation, no usable patterns are retrieved from the RC. A number of failures occur when a collocation is *per se* valid in Spanish, but incorrect in the particular context in which it is used by the learner; cf. (5.25)–(5.27).

(5.25) **fuimos a un museo*, compared to *fuimos a museos* ‘we went to museums’

(5.26) **hacer barbacoa*, compared to *hacer barbacoas* ‘to barbecue’

(5.27) **tienen razón*, compared to *tienen razones* ‘they have reasons’

The same occurs in miscollocations that contain other types of errors, but are classified as ‘Number’ error; cf.: (5.28) and (5.29). An additional source of failure in the negative case is the appearance of lexical errors in the miscollocation, as, e.g., in (5.30), where a wrong selection of an element of the collocation leads to a correct collocation with a different meaning.

(5.28) **tener los derechos*, compared to *tener el derecho* ‘to have the rights’

(5.29) **tiene opciones*, compared to *tiene opción* ‘he has options’

(5.30) **hacer divisiones*, compared to *causar divisiones* ‘to cause separation’,
lit. ‘to make mathematical divisions’

Gender errors. The analysis of the incorrectly classified instances of both ‘Gender’ and ‘other’ miscollocations shows that the misclassification is mainly due to errors resulting from the automatic processing of the writings of the students. For instance, in the case of (5.31–5.33), the first step of our function returns no information, since all three collocations are affected by several errors and therefore, no valid patterns are retrieved from the RC. To account for this case, agreement is checked. In (5.31), both the determiner and the base have been assigned masculine gender, so no agreement error is found and the collocation is classified as ‘Other’. Similarly, *canoa* ‘canoe’ is incorrectly tagged and no agreement error is found either. Finally, in (5.33) a parsing error is responsible for the incorrect assignation of the class, since the determiner appears as depending on the verb.

(5.31) **rechazar los metas*, instead of *alcanzar, lograr las metas* ‘to reach goals’

(5.32) **hacer el canoa*, instead of *ir en canoa* ‘canoeing’

(5.33) **la idioma habla*, instead of *habla un idioma* ‘he speaks a language’

As already (5.31–5.33), the following ‘other error type’ miscollocations are affected by several kinds of errors at the same time, which means that agreement has to be checked. Thus, (5.34) and (5.35) are incorrectly POS-tagged as N-Adj collocations, such that an agreement between the noun and the adjective is looked for. Since none is found, the collocations are judged to have ‘Gender’ errors.

(5.34) **sentado por sillas*, instead of *sentado en sillas* ‘sat on chairs’

(5.35) **completo mis clases*, instead of *termino mis clases* ‘I complete my
classes’

(5.36) **tuve la chance*, instead of *tuve la suerte* ‘I was lucky’

Government errors. An analysis of the results for this kind of error reveals that, as already with ‘Determination’ errors, there is often a correct version of the collocation, in this case with a different government, and it is the context which requires the selection of one or the other alternative. Thus, in (5.37), *tiene el poder* ‘he has the power’ (whithout preposition) should not be used when followed by a verb, but is a possible expression on its own. The same occurs in (5.38) and (5.39).

(5.37) **tiene el poder + V*, instead of *tiene el poder (de) + V* ‘he has the power (to)’ + V

(5.38) **aprovechaba la oportunidad + V*, instead of *aprovechaba la oportunidad (de) + V* ‘I took the most of an opportunity’ + V

(5.39) **tener idea + V*, instead of *tener idea (de) + V* ‘to have idea (of)’ + V

Other types of collocation errors classified as ‘Government’ errors are usually caused by lexical errors involved in the collocation, as in the following examples. In (5.40), a correct collocation can be found with the given base and collocate (*resolución de este problema* ‘solution to a problem’). The same can be observed in (5.41) (*cambiar de religión* ‘to convert to a religion’) and (5.42) (*manejar un coche* ‘to drive a car’). In all of these cases, there is a correct collocation composed by the original base and collocate and a different preposition, which leads the function to classify them as ‘Government’ errors.

(5.40) **resolución a este problema*, instead of *solución a este problema* ‘solution to a problem’

(5.41) **cambiar a la religión*, instead of *convertirse a la religión* ‘to convert to a religion’

(5.42) **manejamos en coche*, instead of *ir en coche, conducir un coche* ‘we drove a car’

Pronoun errors. The lower accuracy rate for the identification of ‘Pronoun’ errors is due to several reasons. Firstly, due to lexical errors in the same miscollocation, almost a third of the queries to the RC does not retrieve any frequencies. Secondly, lexical errors produce combinations in Spanish that are not necessarily collocations. Thus, *sacar una operación a flote/adelante* ‘to get an operation going’ (cf. 5.43) is correct, but it is not a binary collocation. Thirdly, multiple grammatical errors also give place to possible occurrences, as in (5.44). Finally, there are collocations that accept both the pronominal form and the bare verb form (cf. 5.45), where it is the context that marks one or the other use.

(5.43) **sacar una operación*, instead of *hacerse una operación* ‘to have surgery’

- (5.44) **aprovecharme de la oportunidad*, instead of *aprovechar la oportunidad*
‘to take the most of an opportunity’
- (5.45) **volver loco*, instead of *volverse loco* ‘to go mad’

On the contrary, very few collocations of the class ‘other error type’ have been incorrectly classified as ‘Pronoun’ errors. These are cases in which both the pronominal form and the bare verb form are possible, as in (5.46–5.47), or where a lexical error gives rise to an acceptable combination (5.48).

- (5.46) **ir de vacaciones*, compared to *irse de vacaciones* ‘to go on holidays’
- (5.47) **cambios producido*, instead of *producirse cambios* ‘produced changes’
- (5.48) **darnos la idea*, instead of *hacernos una idea* ‘to get an idea’

Order errors. Misclassified ‘Order’ errors are often produced when neither the original combination nor the generated alternatives are found in the RC. As seen before, this is due to multiple errors, such as in (5.50) and (5.51). Another source of error, however, can be seen in (5.49): the use of superlatives, which make the combinations less likely to appear in the RC.

-
- (5.49) **amigas buenísimas*, instead of *buenísimas amigas* ‘close friends’
- (5.50) **nativa parlante*, instead of *hablante nativa* ‘native speaker’
- (5.51) **sumamente creo*, instead of *creo firmemente* ‘I strongly believe’

As far as other types of errors that are classified as ‘Order’ error are concerned, the most frequent reason of this misclassification is that collocations affected by other types of errors (typically ‘Extended substitution’ errors) can form, when reordered, grammatically and semantically valid sequences with high frequency in the RC. Thus, *el día en* ‘the day in’, *buscar trabajo por* ‘to look for a job in’ and *problemas hacen* ‘problems make’ in the following examples are acceptable combinations within a sentence.

- (5.52) **en el día*, instead of *durante el día* ‘in the day’
- (5.53) **buscar por trabajo*, instead of *buscar trabajo* ‘to look for jobs’
- (5.54) **hacen problemas*, instead of *causan problemas* ‘to cause trouble’

5.2 Summary and conclusions

In this chapter, we presented a hybrid approach to collocation error classification. Preliminary experiments with machine learning techniques showed that, due to the heterogeneity of the nature of the errors and the diverging frequencies of each error type, these techniques are not optimal for the classification of all types of collocation errors. The hybrid approach that we propose takes these two aspects into account and uses both machine learning classification algorithms and rules. As done in general grammatical error checking, rules are reserved for the errors that are likely to be easily handled by sets of rules, such as ‘Gender’ or ‘Number’ errors. In our case, the limited number of instances for several of the more diverse error types, such as ‘Determination’ or ‘Government’, required the use of rules as well. Machine learning techniques were reserved for complex error types for which a sufficiently high number of instances was available. Only ‘Extended substitution’ errors could be dealt with by machine learning classifiers.

Our techniques for ‘Extended substitution’ and ‘Creation’ errors are able to detect an error with high accuracy, although in both cases the number of false positives is slightly higher than with the baseline. Our technique for the ‘Different sense’ type of error shows that an exclusive use of lexical contexts is not enough for the judgement of the correct or incorrect use of a collocation. Lexical contexts are rather variant, such that learners can use a collocation correctly in a novel context, and lexical contexts do not capture *situational* contexts, which are key to determine the appropriateness of a given expression.

In all rule-based functions, the error identification has been negatively influenced by two facts: (1) the presence of multiple errors in collocations, which causes that queries to the RC do not retrieve any information, and (2) the automatic preprocessing of the CEDEL2 corpus (note that we are dealing with writings by language learners; the sentences are thus often ungrammatical, such that the error rate of the preprocessing tools (lemmatizer, POS-tagger, morphology-tagger and parser) is considerably higher than in native texts). Research on general grammatical error detection has shown that it is possible to detect grammatical errors such as ‘Determiner’ or ‘Government’ errors with higher accuracy by using statistical approaches. In the context of miscollocations, and due to the small size of our working corpus, this is yet to be confirmed.

An artificial corpus for collocation error detection

As seen in Chapter 1, current collocation checkers focus mainly on collocation validation or identification of miscollocations (usually using mutual information- or distribution-based metrics) in the writings of learners and display of lists of possible corrections, ordered in terms of the strength of their “collocationality” or similarity to the original miscollocation; see, e.g., Chang et al. (2008); Liu et al. (2009); Wu et al. (2010); Ferraro et al. (2014). However, this is by far not sufficient. Ideally, learners should be given the same kind of feedback as given by language instructors when they mark students’ essays: error type-specific symbols or acronyms during marking (e.g., ‘SV’ for “subject verb agreement”, ‘WO’ for “wrong word order”, ‘WW’ for “wrong word”, etc.); see, e.g., Nott (2008). In other words, language instructors classify the students’ mistakes.

In order to be able to offer such advanced collocation checkers, sufficiently large collocation resources, and, in particular, learner corpora annotated with collocation error information, which could be used for training machine learning techniques, are needed. Unfortunately, in second language learning, corpora are usually too small. To remedy this bottleneck, artificial corpora have often been compiled in the context of automatic grammar error detection and correction; cf. Section 3.5 of Chapter 3. In our work, we explore the same idea for collocation error detection and correction, in a preliminary setup. As we will see, the creation of an artificial collocation error corpus is a much more complex task than the creation of a grammatical error corpus. The performance of the error recognizer drops when trained on the artificial corpus and evaluated on the learner corpus, which can be explained by the differences between both corpora. Further work (including experiments with different machine learning techniques) will be needed to come up with high quality artificial collocation error corpora. Still, we think that our exploration is useful in that it shows where the complexity of the generation of an

#	Type	Description	Example
1	SubB	Erroneous selection of the base	* <i>tener <u>confidencia</u></i> , lit. ‘have confidence [secret]’; corr.: <i>tener <u>confianza</u></i> , lit. ‘have confidence [trust]’
2	SubC	Erroneous selection of the collocate	* <i><u>hacer</u> una <u>decisión</u></i> , lit. ‘make a decision’; corr.: <i><u>tomar</u> una <u>decisión</u></i> , lit. ‘take a decision’
3	CrB	Erroneous use of a non-existing base	* <i><u>hacer un llamo</u></i> , lit. ‘make a <i>llamo</i> [non-existing word meaning <i>call</i>]’; corr.: <i><u>hacer una llamada</u></i> , lit. ‘make a call’
4	CrC	Erroneous use of a non-existing collocate	* <i><u>serie televisual</u></i> , lit. ‘[non-existing word meaning <i>TV</i>] series’; corr.: <i><u>serie televisiva</u></i> , lit. ‘TV series’

Table 6.1: Lexical collocation error typology underlying our work

artificial collocation corpus lies and what the directions of the future work should be.

We present an algorithm for the conversion of the Spanish GigaWord corpus into a collocation error corpus of American English learners of Spanish, and an approach to collocation error detection and classification based on *long short-term memory networks* (LSTMs), which uses this corpus. The corpus is created by simulating both the different types of collocation errors produced by language learners, and the distribution of these errors in a learner corpus. In particular, the algorithm attempts to reproduce the most common types of collocation errors produced by English L1 learners of Spanish, generating and inserting V–N, V–Adv, N–Adj collocation errors into Spanish error-free data, according to the typology introduced in Section 6.1. Section 6.2 presents a statistical analysis of the errors observed in the CEDEL2 corpus, necessary for the probabilistic injection of the errors. Section 6.3 describes the algorithm for the generation of the artificial learner corpus, and Section 6.4 provides an overview of the generated corpus. In Section 6.5, the methodology for collocation error detection and the experiments for the evaluation of the approach and the artificial corpus are introduced. Finally, 6.6 summarizes the chapter and presents the conclusions.

6.1 Adapted collocation error typology

In previous work on error detection and correction, the authors commonly divide errors according to the type of operation that is carried out to make a particular error, that is a *substitution* operation, a *deletion* operation or an *insertion* operation. To take into account these operations in the context of learner collocation resources can be useful for the development of more accurate strategies for error correction, and for provision of a better feedback to the learners. Given that the typology by Alonso Ramos et al. (2010) introduced in Chapter 2 does not consider these operations, we opted to include them and subdivide collocation errors, when possible, into these three extra categories. We arrive, thus, at a fine-grained typol-

#	Type	Description	Example
1	DetD	Erroneous omission of a determiner of the nominal base	* <i>ir a _ escuela</i> , lit. ‘go to school’; corr.: <i>ir a la escuela</i> , lit. ‘to to the school’
2	DetI	Erroneous presence of a determiner of the nominal base	* <i>hablar el inglés</i> , lit. ‘speak the English’; corr.: <i>hablar inglés</i> , lit. ‘speak English’
3	GoBD	Erroneous omission of a preposition governed by the base	* <i>tener la oportunidad _ hacer algo</i> , lit. ‘have the opportunity do something’; corr.: <i>tener la oportunidad de hacer algo</i> , lit. ‘have the opportunity of do something’
4	GoBS	Erroneous choice of the preposition governed by the base	* <i>tener obligación a</i> , lit. ‘have the obligation to ...’; corr.: <i>tener obligación de ...</i> , lit. ‘have the obligation of’
5	GoCD	Erroneous omission of the preposition governed by the collocate	* <i>asistir _ una universidad</i> , lit. ‘assist a university’; corr.: <i>asistir a una universidad</i> , lit. ‘assist to a university’
6	GoCI	Erroneous presence of the preposition governed by the collocate	* <i>perder a clientes</i> , lit. ‘lose to clients’; corr.: <i>perder clientes</i> , lit. ‘lose clients’
7	GoCS	Erroneous choice of the preposition governed by the collocate	* <i>ir por tren</i> , lit. ‘go by train’; corr.: <i>ir en tren</i> , lit. ‘go in train’
8	PrD	Erroneous use of a non-reflexive form of the verbal collocate (omission of the reflexive pronoun)	* <i>el hielo _ descongela</i> , lit. ‘the ice melts’; corr.: <i>el hielo se descongela</i> , lit. ‘the ice melts itself’
9	PrI	Erroneous use of the reflexive form of the verbal collocate	* <i>odio que uno se siente</i> , lit. ‘hatred that one feels themselves’; corr.: <i>odio que uno siente</i> , lit. ‘hatred that one feels’
10	NumB	Erroneous number of the base	<i>dar bienvenidas</i> , lit. ‘give welcomes’; corr.: <i>dar la bienvenida</i> , lit. ‘give the welcome’
11	NumD	Erroneous number of base determiner	<i>buena notas</i> , lit. ‘good[sing] marks’; corr.: <i>buenas notas</i> , lit. ‘good[pl] marks’
12	Gen	Erroneous gender	<i>aumentar las precios</i> , lit. ‘raise the[fem] prices’; corr.: <i>aumentar los precios</i> , lit. ‘raise the[masc] prices’
13	Ord	Erroneous word order	<i>educación buena</i> , lit. ‘education good’; corr.: <i>buena educación</i> , lit. ‘good education’

Table 6.2: Grammatical collocation error typology underlying our work

ogy that takes into account, for each type of error: (1) the location of the error, i.e., base, collocate or collocation as a whole (2) the type of error that is produced, i.e., creation, government, order errors, etc., and (3) the type of operation that has been performed by the learner to make the particular error, i.e., substitution, deletion or insertion.¹ As a consequence, we obtain types of errors such as *Government Base Substitution*, where the preposition of the base is incorrectly chosen, *Pronoun Insertion*, where a reflexive pronoun is incorrectly inserted into the collocation, or *Collocate Creation*, where the collocate is an invented word.

An analysis of the CEDEL2 corpus, annotated with collocation errors, revealed that some of the error types in the typology by Alonso Ramos et al. (2010) tend

¹In our work, we also consider the *Explanatory* dimension, used as source of information for the automatic generation of the errors

Error type	Frequency	%
SuC	470	32.41
Gen	116	8.00
GoCD	98	6.76
SuB	96	6.62
DetD	87	6.00
DetI	78	5.38
CrB	72	4.96
GoBS	48	3.31
GoCI	48	3.31
GoCS	45	3.10
Ord	38	2.62
NumB	33	2.27
GoBD	32	2.21
PrI	27	1.86
CrC	25	1.72
PrD	23	1.59
NumD	10	0.69

Table 6.3: Frequency of collocation errors in CEDEL2

to occur very seldomly. For this reason, we opted to disregard them, arriving at 17 classes of lexical and grammatical collocation errors (see Tables 6.1 and 6.2).²

6.2 CEDEL2 Corpus analysis

In order to obtain relevant information about the error distribution in the learner corpus, we start from CEDEL2, carrying out a statistical analysis of the errors present in this corpus. The error distribution is shown in Table 6.3.³ The second column refers to the number of times that each type of error occurs in the corpus, and the third column shows the percentage of the corresponding error type with respect to the total number of collocation errors found in the corpus.

We observed that a collocation can be often affected by several errors at the same time, for instance, containing an error in the base and another in the collocate, such as in **jugar tenis*, ‘to play tennis’, corr. *jugar al tenis*, lit. ‘to play to the tennis’, where there is an omission of the base determiner *el* ‘the’, and an omission

²Note that the error types targeted in this work are a division of the more general types considered in the previous chapter

³Currently, we only consider error types whose raw frequencies are equal or above 5

Error type	Frequency	%
GoCI + SuC	11	0.76
PrD + SuC	10	0.69
GoCD + SuC	9	0.62
DetI + NumB	6	0.41
DetI + GoBS	5	0.34
PrI + SuC	5	0.34
Ord + SuC	5	0.34

Table 6.4: Multiple error types

of the collocate preposition *a* ‘to’.⁴ In the current state of our work, these cases are treated as separate occurrences of the errors, and the decision whether to insert two errors in a collocation is taken randomly by the system.

Furthermore, we found that a base or a collocate can be affected by several errors.⁵ This occurs with lower frequency but is a phenomenon that needs to be reflected in the artificial corpus. Table 6.4 shows all combinations whose raw frequencies are equal or above 5, and presents their frequencies and percentages with respect to the total number of collocation errors in CEDEL2.

In order to generate errors that simulate “real” errors produced by learners, it is not sufficient to copy the error distribution observed in a learner corpus; an analysis of the most confused words is also needed for the cases in which errors are produced through word replacements. In our case, we perform this analysis only for government errors, since, on the one hand, in lexical errors the number of possible options is infinite and thus the usefulness for our work very limited and, on the other hand, the only type of grammatical error where the incorrect choice of a word is considered an error are government errors⁶. The statistics on the confusion sets of prepositions are presented in Tables 6.5, 6.6, 6.7, 6.8 and 6.9.

6.3 Artificial corpus generation

This section presents the algorithm for the generation of the artificial corpus. In particular, Section 6.3.1 describes the general design of the algorithm. Section

⁴In Spanish, when the preposition *a* ‘to’ is followed by the determiner *el* ‘the’, the contracted form *al* is used

⁵We include here cases where an error that affects the collocation as a whole, i.e., **Ord**, and an error affecting either the base or the collocate is produced in the same collocation

⁶Recall that the incorrect choice of determiner and pronoun are not seen as collocation errors

Correct	Incorrect	#	%
None	<i>a</i> 'at'	16	33.33
	<i>con</i> 'with'	14	29.17
	<i>de</i> 'of'	13	27.08
	<i>en</i> 'in'	2	4.17
	<i>por</i> 'by', 'for'	2	4.17
	<i>para</i> 'to', 'for'	1	2.08

Table 6.5: Confusion set GoCI

Correct	Incorrect	#	%
<i>a</i> 'at'	None	83	84.69
<i>en</i> 'in'		8	8.16
<i>de</i> 'of'		3	3.06
<i>con</i> 'with'		2	2.04
<i>por</i> 'by', 'for'		1	1.02
<i>sobre</i> 'over'		1	1.02

Table 6.6: Confusion set GoCD

Correct	Incorrect	#	%
<i>de</i> 'of'	None	27	84.37
<i>en</i> 'in'		3	9.37
<i>para</i> 'to', 'for'		1	3.12
<i>sobre</i> 'over'		1	3.12

Table 6.7: Confusion set GoBD

Correct	Incorrect	#	%
<i>en</i> 'in'	<i>por</i> 'by', 'for'	16	69.56
	<i>a</i> 'at'	4	17.39
	<i>de</i> 'of'	2	8.69
	<i>con</i> 'with'	1	4.35
<i>de</i> 'of'	<i>en</i> 'in'	4	57.14
	<i>a</i> 'at'	3	42.86
<i>para</i> 'to', 'for'	<i>en</i> 'in'	1	100
	<i>en</i> 'in'	7	77.78
<i>a</i> 'at'	<i>de</i> 'of'	1	11.11
	<i>por</i> 'by', 'for'	1	11.11
<i>por</i> 'by', 'for'	<i>en</i> 'in'	2	66.66
	<i>a</i> 'at'	1	33.33
<i>contra</i> 'against'	<i>con</i> 'with'	1	100

Table 6.8: Confusion set GoCS

Correct	Incorrect	#	%
<i>en</i> ‘in’	<i>de</i> ‘of’	2	50
	<i>sobre</i> ‘over’	1	25
	<i>*in</i> ‘in’	1	25
<i>de</i> ‘of’	<i>para</i> ‘to’, ‘for’	9	42.86
	<i>a</i> ‘at’	8	38.09
	<i>en</i> ‘in’	2	9.52
	<i>que</i> ‘that’	1	4.76
	<i>como</i> ‘as’	1	4.76
<i>para</i> ‘to’, ‘for’	<i>por</i> ‘by’, ‘for’	2	50
	<i>a</i> ‘at’	1	25
	<i>de</i> ‘of’	1	25
<i>a</i> ‘at’	<i>de</i> ‘of’	6	85.71
	<i>en</i> ‘in’	1	14.28
<i>por</i> ‘by’, ‘for’	<i>para</i> ‘to’, ‘for’	6	54.54
	<i>a</i> ‘at’	3	27.27
	<i>de</i> ‘of’	2	18.18
<i>sobre</i> ‘over’	<i>de</i> ‘of’	2	100

Table 6.9: Confusion set GoBS

6.3.2 describes a series of *Error Generators*, which are used for the generation of the errors. Finally, Section 6.3.3 presents the corpora and auxiliary resources that are used.

6.3.1 General design

The algorithm passes through three main stages: (1) collocation extraction, (2) collocation classification, and (3) error generation and injection. Firstly, all the N–V, N–Adj and V–Adj dependencies that occur in the corpus where the errors are to be inserted are retrieved and classified, according to their POS pattern, into three groups, that is N–V, N–Adj and V–Adj candidates. A statistical check is performed to reject non-collocations: we choose the asymmetrical normalized *Pointwise Mutual Information* (PMI) by Carlini et al. (2014) and consider as collocations only those dependencies whose PMI is higher than 0. Collocations are stored with their prepositions, determiners and pronouns, along with relevant information that will be used at later stages, such as their position in the sentence, lemmas, POS-tags, morphological information, and their sentential context.

Secondly, collocations are classified according to the types of errors that they can contain. For instance, N–Adj collocations cannot be affected by pronoun errors, V–Adv collocations cannot contain gender errors, collocations that do not contain

a determiner cannot be affected by a determiner insertion error, etc. Table 6.10 shows the error types that can affect a collocation of a given pattern, given a certain condition. At the end of the process, a list of candidates is created for each type of error.

Finally, errors are generated and inserted according to the error distribution presented in the CEDEL2 corpus. In each iteration, an error type is probabilistically chosen by the system; then a candidate is taken, and an *error generator* produces an error, which is inserted into the sentence; otherwise, the candidate is ignored. In order to preserve the error distribution observed in the CEDEL2 corpus, the creation of the corpus ends when the number of candidates for any of the errors is equal to zero. The *Error Generators* that are used are presented below.

6.3.2 Error generators

A total of 6 Generators is used to produce the 17 types of collocation errors that we target. 5 are developed for grammatical errors, and one generates all types of lexical errors. Table 6.11 presents the types of errors that are created by each of the generators.

1. Order Error Generator (OEG)

The OEG takes as input N-Adj and V-Adv collocations and swaps the order of the base and the collocate, generating order errors (**Ord**). In order to avoid the creation of uncontrolled grammatical errors, only collocations whose components appear in contiguous order are considered.

2. Gender Error Generator (GEG)

The GEG's role is to insert gender errors (**Gen**) into V-N and N-Adj collocations. In both types of collocations, gender errors are produced in the determiner of the base.⁷ In N-Adj collocations, the adjectival collocate is considered as a determiner itself, such that gender errors can be produced either in the base determiner or in the collocate. In the cases where a gender error can be inserted in both places, the GEG randomly chooses where to insert the error, i.e., in the determiner or in the adjective.

⁷Recall that, as stated in Chapter 2, according to Alonso Ramos et al. (2010), gender errors are often manifested by a lack of agreement between the nominal base of the collocation and its determiner. In fact, the application of gender inflection rules to nouns in Spanish usually results in the creation of new words (*decisión* 'decision' – **decisiona* '-'; *respuesta* 'answer' – **respuesta* '-'; *miedo* 'fear' – **mieda* '-') or in existing words carrying other meaning than the intended one (*ánimo* 'encouragement') – *ánima* 'soul'), often involving a change in its PoS (*ética* 'ethics' – *ético* 'ethical'[adj]). These cases are treated as 'Creation' or 'Substitution', rather than 'Gender' errors.

Collocation pattern	Error type	Condition	
V-N	NumB SubB SubC CrB CrC	None	
	Gen NumD DetD	\exists determiner	
	DetI	\neg determiner	
	PrD	\exists pronoun	
	PrI	\neg pronoun	
	GoCS GoCD	\exists collocate preposition	
	GoCI	\neg collocate preposition	
	GoBS GoBD	\exists base preposition	
	N-Adj	Gen NumB NumD SubB SubC CrB CrC Ord	None
		GoBS GoBD	\exists base preposition
V-Adv	SubB SubC CrB CrC Ord	None	

Table 6.10: Possible error types for each collocation pattern

Error generator	Error types
Order Error Generator	Ord
Gender Error Generator	Gen
Number Error Generator	NumB NumD
Substitution Error Generator	GoBS GoCS GoBD GoCD DetD PrD
Insertion Error Generator	GoCI DetI PrI
Lexical Error Generator	SuB SuC CrB CrC

Table 6.11: Error types created by each error generator

The GEG is made up of two main functions, one that changes the gender of the determiner, and one that changes the gender of the adjectival collocate. For determiners, the system first checks whether the input determiner is included in a list of irregular determiners, where both masculine and feminine forms are given. If so, the original determiner is replaced by its alternative form. Otherwise, common gender inflection rules are applied according to the determiner's last letters. For adjectives, a suffix map is used, where masculine suffixes are mapped to feminine ones, and vice versa. The system simply checks whether the adjective's last letters are included in the map, and replaces the original ending with the new one.

As a final step, the existence of the created form is guaranteed by checking its frequency in the RC.

3. Number Error Generator (NEG)

The NEG inserts number errors into V-N and N-Adj collocations. As in the case of 'Gender' errors, 'Number' errors can be produced in the determiner or in the adjectival collocate. Differently to 'Gender' errors, however, 'Number' errors can also affect the nominal base of the collocation. The NEG inserts then two types of errors: **NumD** for errors produced in the determiner and adjectival collocate, and **NumB** for errors produced in the base. In cases where the error can be inserted

in more than one place, the NEG randomly chooses where to insert the error.

The NEG works as the GEG, i.e., two main functions are designed, one that deals with determiners and one that deals with adjectives and nouns. A list of irregular determiners together with number inflection rules is used for the former, while a suffix map is used for the latter.

4. Substitution Error Generator (SEG)

The SEG inserts replacement and deletion errors into N–V and N–Adj collocations. In the case of replacement errors, we consider only government replacement errors (**GoBS** and **GoCS**).⁸ The SEG takes as input collocations in which the target component (the base or the collocate) has a government preposition, and replaces it with another preposition, according to the statistics observed in the learner corpus, presented in Tables 6.9 and 6.8.

Changing a preposition often results in an error, but in some occasions it can lead to a correct collocation that involves a change of meaning. For instance, en *tener [una] deuda* ‘to have [a] debt’, the preposition *de* ‘of’ is used to introduce the amount of the debt, for instance, *una deuda de \$10,000* ‘a debt of \$10,000’. The preposition *con* ‘with’ is preferred to introduce the person or entity to whom the debtor is subject. In this case, replacing the preposition *de* with *con* would not result in a collocation error. In order to avoid the introduction of a false error, we defined an PMI-based association metric (see Equation (6.1)) that calculates the association strength between the collocate, the target preposition and the context of the collocation (a window of 2). Only when the contextual PMI of the original collocation is higher than the contextual PMI of the new collocation, the error is inserted.

$$PMI_{CTX}(A, C) = \frac{1}{n} \sum_{i=1}^n PMI(A, C_i) \quad (6.1)$$

where A is the combination of collocate–preposition and C is the context of the collocation.

Deletion errors can be produced in either prepositions, determiners or pronouns, giving place to **GoBD**, **GoCD**, **DetD** and **PrD** errors. The mechanism of the SEG for deletion errors is the same as for replacement errors, the only difference being that while in replacement errors the replacement is a valid element, in deletion

⁸Recall that, according to the typology by Alonso Ramos et al. (2010), the incorrect choice of a determiner or pronoun is not considered a collocation error; only the wrong choice of preposition is. As a consequence, the only type of replacement errors that the SEG generates are **GoBS** and **GoCS**, depending on the element of the collocation where the error is inserted

errors the replacement is void. Contextual PMI is also computed in deletion errors to check that the generated error is a true error.

5. Insertion Error Generator (IEG)

The IEG behaves as the SEG, with the difference that, in this case, none of the elements is changed nor removed, but rather a new element is inserted instead. The IEG generates government, determiner and pronoun insertion errors (**GoCI**,⁹ **DetI** and **PrI**) in N–V and N–Adj collocations. As with the SEG, the IEG also uses contextual PMI scores to avoid the insertion of false errors.

The manner in which the element that is to be inserted is chosen depends on whether the target element is a preposition, a determiner or a pronoun. Prepositions are probabilistically chosen, according to the error statistics observed in the learner corpus, and inserted after the collocate. For determiners, the IEG inserts an indefinite article before the noun. Since neither the definite/indefinite confusion, nor the confusion of any determiner is considered as a collocation error in Alonso Ramos et al. (2010)’s typology, any determiner could be inserted in any case. For simplicity, we opted to always insert indefinite articles, choosing among the different forms depending on the noun number and gender. Finally, pronouns are inserted in two ways, following the rules of the Spanish grammar. For conjugated verbs, the correct pronoun that corresponds to the verb person and number is inserted before the verb. For infinitive forms, the reflexive pronoun *se* is added to the infinitive.

6. Lexical Error Generator (LEG)

The LEG inserts lexical substitution and creation errors in N–V, N–Adj and V–Adv collocations, in both the base and the collocate. The error types covered by the LEG are, therefore, **SuB**, **SuC**, **CrB** and **CrC**. The LEG finds or creates a replacement base or collocate and changes the original base or collocate by the replacement, an existing word in substitution errors, and a non-existing word in creation errors.¹⁰ For substitution errors, the PMI of the new base/collocate and the unchanged element of the collocation is furthermore calculated, and only those combinations whose PMI is higher than 0 are kept as candidates.

Replacement words can be generated in different ways: (1) **transfer**, where the target base or collocate is translated into L1 (English, in our case), (2) **affix change**, where a suffix (including gender inflection) change is applied to the target element, (3) **transfer + affix change**, (4) **synonymy** (only for substitution errors), where the target element is replaced by one of its synonyms, and (5) **literal translation**, (only for collocate substitution errors), where the base is translated into L1, and the verb that most often co-occurs with the base in the L1 is retrieved, translated

⁹In the CEDEL2 corpus the frequency of GoBI errors was rather small, so we opted for disregarding this type of error

¹⁰As in ‘Gender’ errors, the existence of the replacement words is checked in the RC)

into Spanish and used to replace the original verb. The choice of the method for generating the replacement is random. When unable to generate an error by means of the chosen option, the system selects another option until a valid replacement is found or until the options are exhausted.

After the new base/collocate is generated, its form is inflected according to the morphological features of the original base/collocate.

6.3.3 Resources

The following corpora and auxiliary resources are used for the generation of the artificial corpus:

Base corpus: We use the Spanish GigaWord corpus <https://catalog.ldc.upenn.edu/ldc2011t12> as base corpus for the injection of the errors.

Learner corpus: The CEDEL2 corpus is used in order to obtain relevant information regarding the collocation errors that Spanish L2 learners make in their writings. Our working corpus is formed by 517 essays of levels ranging from pre-intermediate to advanced.

As can be expected, the CEDEL2 corpus contains, beside collocation errors, other type of errors that do not affect collocations. In order to have an idea of their types and frequencies, we carried out an analysis of 50 sentences from the corpus, randomly chosen. Table 6.12 presents the types of errors that were observed, along with the number of instances belonging to each type.

Orthographical errors. In total, 43 spelling errors are found in the CEDEL2 sample. Most of them are produced because an accent has been omitted, but the omission, insertion or replacement of other characters is also common. A lesser frequent source of error is the combination of two words into one. The opposite case, the splitting of a word into two, has also been found.

Grammatical errors. As a result of the lack of proficiency of language learners, the CEDEL2 sample is full with grammatical errors of different types. Verbal, government and determiner errors are the most frequent, but gender and number errors also occur often.

Lexical errors. A total of 8 lexical errors have been found in the CEDEL2 sample outside the context of collocations. Surprisingly, only one of them is a creation error; the remaining are correct Spanish words incorrectly selected.

Punctuation errors. Punctuation errors are common in the CEDEL2 sample, often implying the omission of commas, although misplaced and extraneous commas

Error type	Error subtype	# CEDEL2
Orthographical	Accents	31
	Letter-based	10
	Other	2
Grammatical	Government	9
	Determiner	11
	Gender	3
	Number	1
	Verb agreement	4
	Wrong tense	9
Lexical	Substitution	7
	Creation	1
Punctuation	Wrong use	22
Discourse markers	Wrong use	3

Table 6.12: Non-collocation errors in the CEDEL2 sample

have also been observed.

Discourse marker errors. Finally, some errors related to the use of discourse markers have been observed in the CEDEL2 sample, namely wrong use and insertion.

Reference corpora: We use reference corpora (RC) to check word frequencies and co-occurrences. In particular, the algorithm makes use of two RCs, a Spanish RC and an English RC. The Spanish RC consists of 7 million sentences from newspaper material. For English, we use the British National Corpus (BNC), which contains 100 million words from texts of a variety of genres. In order to obtain syntactic dependency information, both corpora were processed with Bohnet (2010)'s dependency parser.

Spanish WordNet: The algorithm also makes use of the *Spanish WordNet*, from the Multilingual Central Repository 3.0 (Agirre et al., 2012), as a source of synonymy information. The *NLTK* library is used to access its contents.

Google Translate: *Google Translate* is used as bi-directional translation engine, both to translate from Spanish to English, and from English to Spanish. Access to it is provided by the *TextBlob* Python library.

Morphological inflection tool: Finally, the algorithm uses the morphological inflection system by Faruqui et al. (2016) This tool allows for the generation of morphologically inflected forms of a word according to given morphological attributes.

In our case, we use it for the generation of lexical errors, to inflect the words that are automatically created by the algorithm as replacement for bases and collocates.

6.4 Enriching the Spanish GigaWord with artificial errors

For our current experiments, we take a fragment of the Spanish GigaWord corpus¹¹ and generate collocation errors according to the algorithm described in the previous section. The result is the collocation error corpus that will be used as training data in the following experiments. But before diving into the experimental part, let us compare the generated resource to the original CEDEL2 corpus. In order to check to what extent our artificial corpus simulates our learner corpus, we carry out an analysis of both of them. In particular, we take a sample of 50 sentences from each corpus and pay attention to three main aspects: (1) collocation errors, (2) non-collocation errors, and (3) sentence complexity. This is, on the one hand, because the analysis of the generated errors and their comparison to the “real” learners’ errors is crucial for a qualitative evaluation of the resource. On the other hand, a comparison of the non-collocation errors and the sentence complexity between the “real” and the synthetic corpus might shed some extra light regarding the similarity of the two corpora. The analysis is presented below.

6.4.1 Collocation errors

A look at the generated errors points to some important conclusions, mainly that, even when some of the generated errors resemble indeed learners’ errors, in some cases the algorithm fails to generate errors correctly. We provide an appendix (Appendix B) with examples of the artificial errors, organized by error type, in their sentential contexts. Here we summarize the conclusions derived from the observation of these failures.

Firstly, not all the combinations in which errors are inserted are real collocations. Some are free combinations, cf., e.g., *representantes _ las islas* ‘islands’ representatives’; orig. *representantes de las islas* and *llenaba el plaza* ‘filled the square’; orig. *llenaba la plaza*.

Secondly, the injection of an “error” does not always produce a collocation error but, rather, results in a correct collocation involving a change of meaning. For instance, in *la depresión nerviosa que le causó la muerte a su mujer* ‘the nervous depression that caused the death to his wife’; orig. *la depresión nerviosa que le causó la muerte de su mujer* ‘the nervous depression that caused the death of her wife’. In other cases, the injection of the “error” results in a change of determination, such as in *consumir una droga* ‘to use a drug’ lit. ‘to consume a drug’; orig. *consumir _ droga* ‘to use drugs’ lit. ‘to consume drug’.

¹¹Processed with Bohnet (2010)’s syntactic dependency parser.

Finally, the injection of an error may result in the generation of unexpected errors. For example, the substitution of *instrumento* ‘instrument’ by its synonym *herramienta* ‘tool’ in *es un herramienta que manejaremos* ‘it is a tool that we will use’; orig. *es un instrumento que manejaremos* ‘it is an instrument that we will use’, produces a determiner error, since there is no agreement between the changed base *herramienta* and the determiner.

These shortcomings obviously lower the expectations with respect to the quality of the generated corpus for the task of collocation error recognition and correction.

6.4.2 Non-collocation errors

This section summarizes our findings regarding the production of errors outside the context of collocations. In particular, we consider orthographical, grammatical, lexical, punctuation and discourse marking errors. Our base corpus (the GigaWord) is assumed to be well written, and thus to be free of any error, apart from those collocation errors that were automatically generated. However, a closer look at it reveals that it contains some spelling and grammatical mistakes. In particular, spelling errors are present, although their proportion and variety is much smaller than in the CEDEL2 corpus: only an unaccented word and 4 typos have been found. The only type of grammatical error observed in the GigaWord sample are agreement errors. Lexical, punctuation and discourse marker errors have not been observed. The small proportion and variety of non-collocation errors in the artificial corpus is likely to have a negative effect on the performance of the error recognizer over the CEDEL2 corpus.

6.4.3 Sentence complexity

In order to measure the sentence complexity, we select several features that can approximate the level of sentence complexity. These features and the values obtained for the two samples are presented in Table 6.13. In order to obtain the POS and syntactic features, the samples have been processed with Bohnet (2010)’s dependency parser.

As can be observed in Table 6.13, the values for some of the features, such as the coordination of passivization ratios are rather similar in both corpora. However, each corpus also shows its own morpho-syntactic characteristics. For instance, the apposition ratio is 8 times higher in the GigaWord corpus than in the L2 corpus. Nouns and adjectives are also significantly more common in the GigaWord corpus, as is the use of punctuation marks. On the contrary, learners tend to use more adverbs and subordinate clauses. As expected, sentence length is substantially shorter in L2 writings.

All these differences are likely to imply that the algorithm trained on artificial data may not perform as well on L2 data as it may on the artificial data.

Feature	CEDEL2	GigaWord
Total words	1,301	2,021
Average sentence length	26.02	40.42
Sentence noun ratio	5.14	10.10
Sentence adjective ratio	1.54	5.18
Sentence verb ratio	3.62	3.50
Sentence adverb ratio	1.56	0.90
Sentence punctuation ratio	2.24	3.38
Sentence coordination ratio	1.10	1.10
Sentence subordination ratio	0.94	0.50
Sentence relativization ratio	0.72	0.66
Sentence passivization ratio	0.18	0.18
Sentence apposition ratio	0.08	0.66

Table 6.13: Syntactic complexity features in the GigaWord and CEDEL2 samples

6.5 Evaluation of the generated corpus for collocation error detection

In this section, we present an evaluation of the generated corpus as a resource for training (and testing) the algorithm for detection of collocation errors. Firstly, Section 6.5.1 describes the approach chosen for the detection and classification task. Then, Section 6.5.2 describes the setup of the experiments, and 6.5.3 presents their outcome. The results are discussed in Section 6.5.4.

6.5.1 The collocation error marking model

Our model for grammatical collocation error marking is motivated by the way in which L2 teachers mark their students' essays: reading the text sequentially and introducing an error mark where an error has been produced. The model is inspired by recent works on neural network architectures for structure prediction, such as Dyer et al. (2015)'s transition-based parsing model and Dyer et al. (2016)'s generative language model and phrase-structure parser. In particular, it follows Ballesteros and Wanner (2016)'s method for punctuation generation, which is based on the transition-based parsing model by Dyer et al. (2015) and on character-based continuous-space word embeddings using bidirectional LSTMs (Ling et al., 2015b; Ballesteros et al., 2015). For a given piece of input raw text, their system is able to introduce punctuation marks where appropriate by reading words sequentially and taking decisions whether to pass or to introduce a specific punctuation symbol. The method works in a similar way in the case of collocation error identification:

words are read sequentially and, after the reading of each word, a decision is made whether to pass or to introduce an error-specific tag (cf., Figure 6.1).

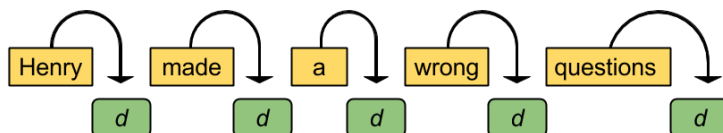


Figure 6.1: Algorithm sequential reading

6.5.1.1 Error Identification and Marking Algorithm

The input of the algorithm is a sentence, and its output is the input sentence that may or may not contain some marks indicating grammatical collocation errors. The input and output are represented in terms of sequential data structures: the input and the output buffers. The input buffer contains the words that are to be processed, while the output buffer contains the words that have already been processed. In order to represent the words, we learn (jointly with the neural network) a vector representation for each word. All out-of-vocabulary words are represented with the same fixed vector representation. In order to train the fixed vector representation, we apply during training stochastic replacement (with $p=0.5$) of words that occur only once in the training set. The obtained word representations are then concatenated with a vector representation from a neural language model provided as auxiliary input to the system (pretrained word embeddings). The resulting vector is passed through a component-wise rectified linear unit (ReLU). To pretrain the fixed vector representations, we use the skip n -gram model introduced by Ling et al. (2015a).

The algorithm starts with an input buffer full of words and an empty output buffer. Looking at an input word, it has the choice to make one of the two following decisions: (a) *pass* to the next word (which means that it decides that no error should be introduced), or (b) *introduce an error-type specific mark* according to the typology in Section 6.1. Whenever an error type mark is introduced, the algorithm treats it as another “word” and adds it to the output sequence. An example for the processing of the sentence *Henry made a wrong choices* is given in Figure 6.2.

At each stage t of the application of the algorithm, the current processing state, which is defined by the contents of the output and input sequences, is encoded in terms of a vector \mathbf{s}_t , which is learned (see Section 6.5.1.2 below for the chosen representation). The vector \mathbf{s}_t is used to compute the probability of the action at time t as follows:

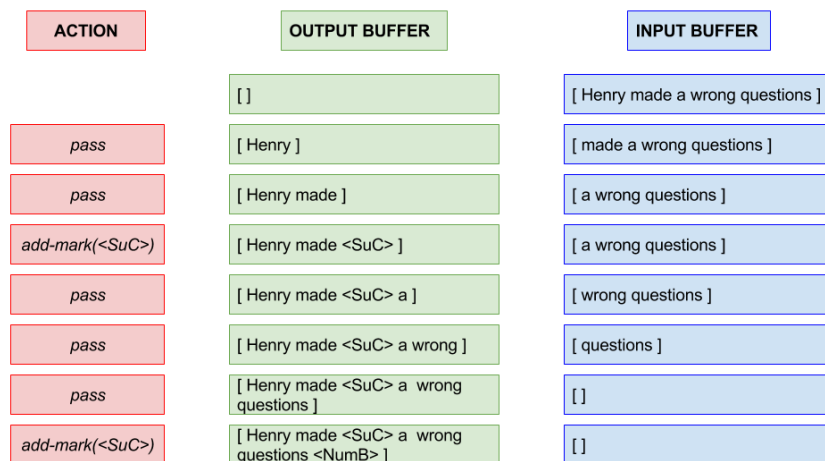


Figure 6.2: Example of the processing steps of the algorithm

$$p(z_t | \mathbf{s}_t) = \frac{\exp(\mathbf{g}_{z_t}^\top \mathbf{s}_t + q_{z_t})}{\sum_{z' \in \mathcal{A}} \exp(\mathbf{g}_{z'}^\top \mathbf{s}_t + q_{z'})} \quad (6.2)$$

where \mathbf{g}_z is a column vector representing the (output) embedding of the action z , and q_z is a bias term for action z . The set \mathcal{A} represents the actions (either PASS to the next word or ADD-MARK(e)) (that adds a mark with the error E).¹²

During training, at the end of each sentence, the parameters are updated with the goal to maximize the likelihood of the reference set of decisions provided (for that sentence) by the annotated corpus. During decoding, the model greedily chooses the best action to take, given the contents of the input and the output.

6.5.1.2 Representing the Processing State

We work with a recurrent neural network model (long short-term memory networks, LSTMs), which encodes the entire input sequence and the entire output sentence, to represent the input and output buffers (i.e., the state \mathbf{s}_t).

¹²Note that ADD-MARK(e) includes all possible errors described in Section 6.1, and thus the number of classes the classifier predicts in each time step is #errors + 1.

LSTM Model LSTMs are a variant of RNNs designed to cope with the vanishing gradient problem inherent in RNNs Hochreiter and Schmidhuber (1997); Graves (2013). RNNs read a vector \mathbf{x}_t at each time step and compute a new (hidden) state \mathbf{h}_t by applying a linear map to the concatenation of the previous time step's state \mathbf{h}_{t-1} and the input, passing then the outcome through a logistic sigmoid non-linearity (cf., Figure 6.3).

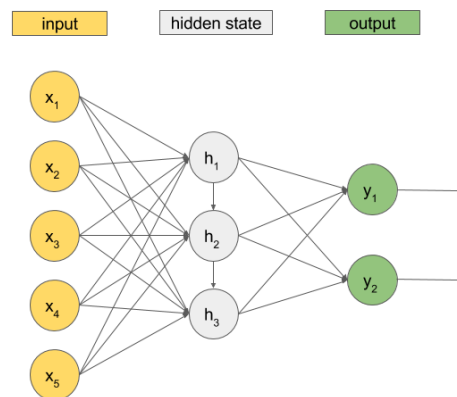


Figure 6.3: Recurrent Neural Networks schema

The input buffer is encoded by an LSTM, into which we add the entire sequence at the beginning, and remove words from it at each time step. The output buffer is a sequence, encoded by an LSTM, into which we add the final output sequence, including the error type marks. As in Dyer et al. (2015), we furthermore use a third sequence with the history of decisions taken, which is encoded by another LSTM. The three resulting vectors (learned with the LSTMs) are passed through a component-wise ReLU and a softmax transformation -a function that transforms the net activations of the output layer into a vector of probabilities- to obtain the probability distribution over the possible decisions to take, given the current state \mathbf{s}_t .

6.5.2 Experimental setup

For our experiments, we take the fragment of the Spanish Gigaword enriched with collocation errors, presented in Section 6.4, and use it for training the LSTMs. From total of 25,894 sentences, 22,415 were used for training and 2,163 for development data. The remaining 1,316 were set apart for testing. Table 6.14 shows the number of collocation errors of each type in the training, development and test corpora. Table 6.15 shows the number of multiple errors.

Error type	Training	Development	Test
SuC	5275	524	300
Gen	1163	126	63
GoCD	1133	104	56
SuB	1002	98	56
DetD	979	83	41
DetI	792	73	40
CrB	785	66	39
GoCI	595	47	37
GoBS	530	49	23
GoCS	493	58	22
Ord	398	41	32
NumB	387	41	23
GoBD	355	25	30
PrI	294	40	20
CrC	274	24	17
PrD	262	17	11
NumD	116	12	5

Table 6.14: Number of collocation errors in training, development and test corpora

Error type	Training	Development	Test
DetI + NumB	58	5	2
PrD + SuC	48	2	1
GoCI + SuC	29	3	0
GoCD + SuC	13	5	1
Ord + SuC	1	0	0
DetI + GoBS	1	0	0

Table 6.15: Number of multiple collocation errors in artificial training, development and test corpora

Error type	Precision (p)	Recall (r)
CrB	1	0.97
CrC	0.92	0.70
SuB	0.92	0.45
SuC	0.94	0.83
DetD	0.85	0.41
DetI	1	0.87
Gen	0.82	0.59
GoBD	1	0.53
GoBS	0.92	0.52
GoCD	0.90	0.66
GoCI	0.97	0.86
GoCS	1	0.45
NumB	1	0.96
NumD	1	0.60
Ord	1	0.62
PrD	1	0.73
PrI	1	0.6
DetI + NumB	1	1
DetI + GoBS	–	–
GoCD + SuC	–	–
GoCI + SuC	–	–
Ord + SuC	–	–
PrD + SuC	–	–

Table 6.16: Performance of the system when trained and evaluated on artificial data

Since our ultimate goal is the detection of errors in learners' writings, we carry out a second evaluation, specifically on the CEDEL2 corpus. The number of instances of each type in the CEDEL2 corpus is presented in Table 6.3, in Section 6.2.

6.5.3 Outcome of the experiments

Table 6.16 shows the performance of the model when evaluating on a fragment of the artificial data. Precision and recall are presented for each of the error types.

Table 6.17 shows the performance of the model when the evaluation is carried out on L2 data, i.e., on original CEDEL2 data.

Error type	Precision (p)	Recall (r)
CrB	0.48	0.63
CrC	0.54	0.38
SuB	0.56	0.27
SuC	0.65	0.58
DetD	0.31	0.20
DetI	0.56	0.52
Gen	0.51	0.42
GoBD	0.70	0.26
GoBS	0.29	0.18
GoCD	0.65	0.50
GoCI	0.56	0.47
GoCS	0.76	0.24
NumB	0.77	0.58
NumD	0.54	0.31
Ord	0.64	0.34
PrD	0.62	0.44
PrI	0.71	0.30
DetI + NumB	0.50	0.60
DetI + GoBS	–	–
GoCD + SuC	–	–
GoCI + SuC	–	–
Ord + SuC	–	–
PrD + SuC	–	–

Table 6.17: Performance of the system when trained on artificial data and evaluated on the CEDEL2 corpus

6.5.4 Discussion

As can be observed in Table 6.16, the performance of the system when applied to artificial data is rather high on single errors. With the exception of two cases, precision is above 0.9. Compared to precision, recall seems to be lower for most of the error types, and there is no evident pattern that the system finds easier to identify. For instance, ‘Creation’ errors in the base present a recall of 0.97, but the recall for ‘Creation’ errors in the collocate descends to 0.70. On the contrary, the recall for ‘Substitution’ errors is almost double for collocate errors than for errors in the base. A look at the number of training instances reveals that, in this case, the number of examples and the achieved recall are directly correlated. However, this is not always the case in grammatical errors; for instance, for ‘Pronoun’ errors, deletions are more easily identified than insertions. Still, deletions are not always better recognized than insertions, as can be observed in the case of ‘Determiner’

errors, where **DetI**, with lower number of training instances, achieves a recall two times higher than its counterpart **DetD**.

With regards to multiple errors, only **DetI + NumB** are identified. Surprisingly, a precision and recall of 1 is achieved.

Despite the encouraging results presented above for the detection of collocation errors in synthetic data, a drop of the performance of the system is observed when evaluating the model on “real” learner data. This is as has been expected (and mentioned before). Both precision and recall decrease notably. Still, a precision of 0.7 or higher is achieved for four error types, and the number of error types that the system is able to detect remains the same. The lack of proficiency of language learners, along with the particular characteristics of each corpus (described in Section 6.4) and the differences in topics and style might be responsible for this drop. Errors outside the collocation may have an influence on the performance of a collocation error detection system trained on artificial data, and domain, style and differences on vocabulary and syntactic structures might represent a certain amount of noise for the system.

In order to minimize these differences, we tried to adapt the model to learner language, opting for enriching the training and development data with examples of learner data. In particular, we split the CEDEL2 corpus into three sections, adding 40% to the training data, and 20% to the development data, and using the remaining 40% as test set, and perform new experiments. Tables 6.18 and 6.19 show the number of training, development and test instances (the number within the parenthesis refers to the number of CEDEL2 instances added to the original artificial datasets), and Table 6.20 shows the results.

As can be observed in Table 6.20, only five classes of errors are identified. Both precision and recall are lower than when trained with artificial instances only, which might suggest at first sight that the addition of L2 data confuses the system. However, it must be noted that the size of the CEDEL2 test set has been reduced to 40% of the corpus, so that a direct comparison of results should not be done. Table 6.21 shows the results achieved by the model trained on artificial data and evaluated on the 40% test set of the CEDEL2 corpus. The performance of the system falls drastically with the reduction of test instances, and an improvement in precision can be observed with the addition of CEDEL2 sentences to the training corpus. For instance, an increase from 0.03 to 0.24 is achieved for **SuC** errors, and **GoCD** and **GoCS** are recognized with a precision of 0.25 and 0.5. Recall, on the contrary, decreases for all the identified error types. Still, the results are somehow encouraging when considering that, because of the small size of the CEDEL2 corpus, the number of learner errors added to the training and development sets is, for most error types, insignificant and, that, when a sufficient number of instances is added (**SuC** errors), precision rises notably. On the other hand, it cannot be claimed from these results that the addition of learner data leads to better performance, but unfortunately, at this moment, we lack the data for further experimentation.

Error type	Training	Development	Test
SuC	5491 (216)	606 (82)	172
Gen	1217 (54)	147 (11)	41
GoCD	1166 (33)	120 (16)	49
SuB	1045 (43)	113 (15)	38
DetD	1006 (27)	91 (4)	52
DetI	829 (37)	84 (11)	30
CrB	811 (26)	74 (8)	38
GoCI	619 (24)	54 (7)	17
GoBS	553 (23)	58 (9)	16
GoCS	518 (25)	61 (3)	17
Ord	409 (11)	47 (6)	21
NumB	405 (18)	44 (3)	12
GoBD	374 (19)	33 (8)	5
PrI	308 (14)	49 (9)	4
CrC	280 (6)	30 (6)	13
PrD	275 (13)	20 (3)	7
NumD	118 (2)	16 (4)	4

Table 6.18: Number of collocation errors in artificial + CEDEL2 corpus

Error type	Training	Development	Test
DetI + NumB	61 (3)	5 (0)	3
PrD + SuC	52 (4)	5 (2)	3
GoCI + SuC	33 (4)	4 (1)	6
GoCD + SuC	15 (2)	5 (0)	7
Ord + SuC	5 (4)	0 (0)	1
DetI + GoBS	3 (2)	0 (0)	3

Table 6.19: Number of multiple collocation errors in artificial + CEDEL2 corpus

Error type	Precision (p)	Recall (r)
CrB	0.10	0.05
SuC	0.24	0.03
Gen	0.06	0.02
GoCD	0.25	0.02
GoCS	0.50	0.06

Table 6.20: Performance of the system when trained on artificial data enriched with CEDEL2 data and evaluated on the remaining CEDEL2 corpus

Error type	Precision (p)	Recall (r)
CrB	0.06	0.18
SuC	0.03	0.04
DetI	0.05	0.1
Gen	0.02	0.07
NumD	0.08	0.25

Table 6.21: Performance of the system when trained on artificial data and evaluated on 40% of the CEDEL2 corpus

6.6 Summary and conclusions

In this Chapter, we have introduced and evaluated an algorithm for the automatic generation and tagging of collocation errors in native writings. The algorithm follows approaches for the generation of grammatical errors, such as those presented by Foster and Andersen (2009); Rozovskaya and Roth (2010b).

In Section 6.1, we present a modification of the typology of collocation errors by Alonso Ramos et al. (2010), which we use in our experiments. The modifications consist in the subdivision of errors into *deletion*, *insertion* or *substitution* errors, where appropriate. Section 6.2 presents the results from the statistical analysis of the CEDEL2 corpus, which we use as model for the design of the error generation and injection algorithm, and Section 6.3 describes the structure of the algorithm and its modules. As far as we know, our algorithm is the first attempt at generating artificial collocation errors, including lexical and grammatical errors.

A qualitative evaluation of the generated corpus is given in Section 6.4. A comparison of the CEDEL2 corpus with our synthetic corpus is carried out, paying attention at three aspects in both corpora: (1) collocation errors, (2) non-collocation errors and (3) sentence complexity. Regarding collocation errors, three types of failures have been observed:

- (a) that not all the combinations into which errors were indeed real collocations, but free combinations, as in *llenaba el plaza* ‘filled the square’; orig. *llenaba la plaza*,
- (b) that the injection of an “error” does not always results in a collocation error but in a correct collocation involving a change of meaning, like in *la depresión nerviosa que le causó la muerte a su mujer* ‘the nervous depression that caused the death of her wife’; orig. *la depresión nerviosa que le causó la muerte de su mujer* ‘the nervous depression that caused the death to her wife’, or in a determination error, as in *consumir una droga* ‘to use a drug’

lit. ‘to consume a drug’; orig. *consumir _ droga* ‘to use drugs’ lit. ‘to consume drug’,

- (c) that the generation of an error results in unexpected errors, for instance, the replacement of *instrumento* ‘instrument’ by its synonym *herramienta* ‘tool’ in *es un herramienta que manejaremos* ‘it is a tool that we will use’; orig. *es un instrumento que manejaremos* ‘it is an instrument that we will use’, generates a determination error, since there is no agreement between the changed base *herramienta* and the determiner.

As expected, non-collocation errors, such as orthographical, grammatical, lexical, punctuation and discourse marking errors occur in the CEDEL2 corpus with higher frequency than in the generated resource. Regarding sentence complexity, several features, such as the average sentence length, and the apposition, noun and adjective ratios suggest that the synthetic corpus presents a higher complexity than the CEDEL2 corpus.

Finally, Section 6.5 provides an evaluation of the artificial corpus as a resource to train an algorithm for the detection and classification of collocation errors in L2 writings. Given the lack of sufficiently large annotated learner corpora, and considering the success of artificial corpora for training grammatical error detection systems (Foster and Andersen, 2009; Rozovskaya and Roth, 2010b; Felice and Yuan, 2014), this strategy seemed optimal. Nevertheless, it has been shown that collocation error corpus creation is much more complex than the creation of a grammatical error corpus. An evaluation of our LSTMs-based technique over artificial and L2 data shows a drop in the performance of the system in the ‘real’ language scenario, which indicates that the CEDEL2 and the artificial corpora display substantial differences. As observed in the qualitative evaluation (Section 6.4), these differences include the presence of non-collocation errors in L2 texts, some of which are similar in form to some of those considered in our typology (determination, government, gender, number, etc.), which may confuse the models, and differences in sentence complexity and style. Non-collocation errors and sentence complexity and style must be addressed for better error generation strategies, for instance, through the use of other types of base corpora that contain simpler grammatical constructions, such as children essays. The existence of ‘erroneous errors’ in the artificial corpus suggests that collocation errors cannot be so easily generated as grammatical errors, such that more sophisticated collocation error creation techniques should be explored. In a final experiment, we add L2 instances to the artificial data for training, but the small size of learner data hinders us from drawing relevant conclusions whether the addition of L2 data results in performance improvement.



Conclusions and Future Work

This chapter presents a summary and the final conclusions of the thesis. Section 7.1 summarizes the contributions of our work, and Section 7.2 provides a list of the resulting publications. Section 7.3 presents the limitations of the thesis. Finally, Section 7.4 lists the tasks that can be tackled next.

7.1 Contributions of the thesis

The first contribution of our work concerns the extraction and semantic classification of collocations. Collocation dictionaries that group the collocates of a base according to their meanings are scarce resources, manually created, and often of limited coverage. In Chapter 4, we propose two techniques, one unsupervised, the other semi-supervised, for collocation discovery and classification. Experiments suggest that a small investment in annotation can lead to better resources, and that with only a few examples of collocations belonging to a target category, it is possible to retrieve pertinent collocates for a given base. Since the resources required by both approaches can be easily obtained, they are highly scalable and portable to any language. Given the lack of semantically tagged collocation resources for most languages, our work has the potential to become influential, especially in the context of second language learning.

The second contribution of the thesis consists in the development of techniques for the identification and classification of collocation errors in the writings of language learners. The advantages that the error-type categorization of collocations could bring are multiple, and include, for instance, the possibility to provide the learners more detailed feedback about their use of collocations. From the point of view of NLP, a classification of miscollocations opens the door to further work on collocations, such as the development of error-specific automatic correction techniques. Despite the advantages, the heterogeneous nature of collocation errors is often ne-

glected in current collocation checkers. To the best of our knowledge, ours is the first work devoted to this task.

In Chapter 5, we present a hybrid approach to collocation error classification according to a fine-grained typology of collocation errors. Rule-based and machine learning techniques are used depending on the specific error type and its characteristics. With diverging performance, our techniques are able to distinguish between most of the error types.

In Chapter 6, we propose a technique for the detection and classification of miscollocations based on LSTMs. Our experiments with an artificial corpus suggest that the technique is valid for the task.

The final contribution concerns the design and implementation of an algorithm for the generation of a Spanish artificial collocation error corpus. Due to the lack of learner corpora annotated with collocation errors, the use of synthetic data is necessary to model statistically collocation errors. The algorithm, described in Section 6, covers lexical and grammatical errors of different types and injects errors probabilistically, according to the error distribution observed in the CEDEL2 corpus. Experiments on collocation error recognition show that the task of collocation error corpus creation is much more complex than, e.g., grammatical error corpus creation, but our work outlines the direction of the future work in this area.

7.2 Publications

The work presented in this thesis has already been disseminated in the following publications:

- Sara Rodríguez-Fernández, Roberto Carlini, and Leo Wanner. Classification of Lexical Collocation Errors in the Writings of Learners of Spanish. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 529–536, Hissar, Bulgaria, 2015a
- Sara Rodríguez-Fernández, Roberto Carlini, and Leo Wanner. Classification of Grammatical Collocation Errors in the Writings of Learners of Spanish. *Procesamiento del Lenguaje Natural*, 55:49–56, 2015b
- Sara Rodríguez-Fernández, Luis Espinosa-Anke, Roberto Carlini, and Leo Wanner. Semantics-Driven Recognition of Collocations Using Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 499–505, Berlin, Germany, 2016a
- Sara Rodríguez-Fernández, Luis Espinosa-Anke, Roberto Carlini, and Leo Wanner. Example-based Acquisition of Fine-grained Collocation Resources. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 2317–2322, Portorož, Slovenia, 2016b

- Sara Rodríguez-Fernández, Luis Espinosa-Anke, Roberto Carlini, and Leo Wanner. Semantics-Driven Collocation Discovery. *Procesamiento del Lenguaje Natural*, 57:57–64, 2016c
- Luis Espinosa Anke, Jose Camacho-Collados, Sara Rodríguez-Fernández, Horacio Saggion, and Leo Wanner. Extending WordNet with Fine-Grained Collocational Information via Supervised Distributional Learning. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 3422–3432, Osaka, Japan, 2016
- Sara Rodríguez-Fernández, Roberto Carlini, and Leo Wanner. Generation of a Spanish Artificial Collocation Error Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan, Accepted for publication

7.3 Limitations of the Thesis

Due to the limited period of time set for the PhD dissertation, this thesis presents a number of limitations.

So far, we focused in our work on semantic classification on a set of 10 semantic glosses of V–N and N–A types. Other categories should be included to generate more complete collocation resources. Also, the performance of both techniques for semantic classification is consistently, and considerably, lower for ‘negative’ glosses than for ‘positive’ ones.

As far as the classification of collocation errors is concerned, the major limitation of our work is the lack of a sufficiently large annotated learner corpus. This has made it necessary to resort to simple rule-based techniques for the identification of complex errors, such as determiner, pronoun or government errors. Research on general grammatical error detection/correction has proven that ‘Determiner’ or ‘Government’ errors can be detected with higher accuracy by using statistical approaches rather than rule-based approaches.

Another important limitation of our work in collocation error classification is the relation between errors. The presence of multiple errors in collocations affects the classification of lexical and grammatical errors, but we have not fully considered this fact yet. Only in our experiments with the LSTMs-based technique multiple errors are targeted.

Regarding the creation of artificial errors, an analysis of the generated corpus reveals that between the created artificial corpus and the CEDEL2 corpus there are differences in at least three aspects: sentence complexity, errors outside the context of the collocations, and errors that affect the collocations. These differences have a negative effect on the performance of our techniques when trained with the artificial corpus and evaluated on the CEDEL2 corpus.

7.4 Future work

In what follows, we suggest some possible improvements to overcome the limitations mentioned above, along with some proposals of future tasks that our work has opened up.

- Expand the set of semantic categories considered up to now in the task of collocation retrieval and semantic classification.
 - Investigate the impact of the chosen seed example on the collocates retrieved by the example-based system for collocation retrieval and classification.
 - Explore how to increase the performance of the collocation retrieval and classification techniques for ‘negative’ glosses.
 - Disambiguate the lexical units for the generation of semantically tagged collocation resources. We have already started to work on the retrieval and classification of disambiguated collocations and their incorporation into WordNet.
 - Enlarge the size of the learner corpus and continue with the annotation process until a sufficiently large corpus is attained.
 - Study the relation between different types of errors in collocations and develop techniques that take this relation into account.
 - Explore different machine learning techniques for collocation error detection, including domain adaptation techniques that minimize the differences between the artificial and real learner corpora.
 - Develop type-targeted strategies for collocation error correction.
-
-
-
-

Bibliography

At the end of each reference the pages where it appears are indicated.

- Aitor González Agirre, Egoitz Laparra, and German Rigau. Multilingual Central Repository Version 3.0: Upgrading a Very Large Lexical Knowledge Base. In *Proceedings of the 6th International Global Wordnet Conference (GWC)*, pages 118–125, Matsue, Japan, 2012. 120
- Margarita Alonso Ramos, Leo Wanner, Orsolya Vincze, Gerard Casamayor del Bosque, Nancy Vázquez Veiga, Estela Mosqueira Suárez, and Sabela Prieto González. Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 3209–3214, La Valetta, Malta, 2010. xii, 3, 6, 21, 22, 23, 24, 27, 48, 51, 52, 53, 90, 96, 108, 109, 114, 117, 118, 132
- Luis Espinosa Anke, Jose Camacho-Collados, Sara Rodríguez-Fernández, Horacio Saggion, and Leo Wanner. Extending WordNet with Fine-Grained Collocational Information via Supervised Distributional Learning. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 3422–3432, Osaka, Japan, 2016.
- Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. Freeling 1.3: Syntactic and Semantic Services in an Open-source NLP Library. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 48–55, Genoa, Italy, 2006. 70
- Jens Bahns and Moira Eldaw. Should we Teach EFL Students Collocations? *System*, 21(1):101–114, 1993. 2, 4
- Miguel Ballesteros and Leo Wanner. A Neural Network Architecture for Multilingual Punctuation Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1048–1053,

- Austin, Texas, USA, 2016. ACL (Association for Computational Linguistics). 123
- Miguel Ballesteros, Chris Dyer, and Noah A. Smith. Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 349–359, Lisbon, Portugal, 2015. 123
- Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Frédéric Morin, and Jean-Luc Gauvain. Neural Probabilistic Language Models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006. 68
- Morton Benson. The Structure of the Collocational Dictionary. *International Journal of Lexicography*, 2(1):1–13, 1989. 10
- Morton Benson, Evelyn Benson, and Robert Ilson. *The BBI Combinatory Dictionary of English: Your Guide to Collocations and Grammar, Third Edition*. Benjamins Academic Publishers, Amsterdam, 2010. 10
- Shane Bergsma, Dekang Lin, and Randy Goebel. Web-scale N-gram Models for Lexical Disambiguation. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1507–1512, Pasadena, California, 2009. 56, 58
- Frank Boers and Seth Lindstromberg. *Optimizing a Lexical Approach to Instructed Second Language Acquisition*. Springer, 2009. 4
- Bernd Bohnet. Very High Accuracy and Fast Dependency Parsing is Not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 89–97, Beijing, China, 2010. Association for Computational Linguistics. 97, 98, 120, 121, 122
- Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning Structured Embeddings of Knowledge Bases. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pages 301–306, San Francisco, California, USA, 2011. 68
- Gerlof Bouma. Normalized (Pointwise) Mutual Information in Collocation Extraction. In C. Chiarcos, R. Eckart de Castilho, and M. Stede, editors, *Proceedings of the Biennial Gesellschaft für Sprachtechnologie & Computerlinguistik Conference (GSCL)*, pages 31–40, Potsdam, Germany, 2009. 35, 36, 156
- Gerlof Bouma. Collocation Extraction beyond the Independence Assumption. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 109–114, Uppsala, Sweden, 2010. 35, 67, 156
- Chris Brockett, William B. Dolan, and Michael Gamon. Correcting ESL Errors Using Phrasal SMT Techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 249–256, Sydney, Australia, 2006. Association for Computational Linguistics. 61
- Aoife Cahill, Nitin Madnani, Joel R. Tetreault, and Diane Napolitano. Robust Systems for Preposition Error Correction Using Wikipedia Revisions. In *Proceed-*

- ings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 507–517, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics. 65
- Roberto Carlini, Joan Codina-Filba, and Leo Wanner. Improving Collocation Correction by Ranking Suggestions Using Linguistic Knowledge. In *Proceedings of the 3rd Workshop on NLP for Computer-assisted Language Learning at the 5th Swedish Language Technology Conference (SLTC)*, pages 1–12, Uppsala, Sweden, 2014. 36, 69, 70, 78, 113, 156
- Yu-Chia Chang, Jason S. Chang, Hao-Jan Chen, and Hsien-Chin Liou. An Automatic Collocation Writing Assistant for Taiwanese EFL learners. A case of Corpus Based NLP Technology. *Computer Assisted Language Learning*, 21(3): 283–299, 2008. 5, 42, 43, 48, 51, 107
- Martin Chodorow, Joel Tetreault, and Na-Rae Han. Detection of Grammatical Errors Involving Prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions at ACL-2007*, pages 25–30. Association for Computational Linguistics, 2007. 54
- Yaakov Choueka. Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. In *Proceedings of the 2nd International Conference on Computer-Assisted Information Retrieval (RIAO)*, pages 34–38, Cambridge, Massachusetts, USA, 1988. 32, 33, 67
- Kenneth W. Church and Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 76–83, Vancouver, Canada, 1989. 32, 33, 67
- Anthony P. Cowie. Phraseology. In R.E. Asher and J.M.Y. Simpson, editors, *The Encyclopedia of Language and Linguistics*, Vol. 6, pages 3168–3171. Pergamon, Oxford, 1994. 7
- Anthony P. Cowie and Peter Howarth. Phraseological Competence and Written Proficiency. *British Studies In Applied Linguistics*, 11:80–93, 1996. 49
- D. Alan Cruse. *Lexical Semantics*. Cambridge University Press, Cambridge, 1986. 9
- Daniel Dahlmeier and Hwee-Tou Ng. Correcting Semantic Collocation Errors with L1-induced Paraphrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 107–117, Edinburgh, Scotland, 2011a. Association for Computational Linguistics. 48, 81
- Daniel Dahlmeier and Hwee-Tou Ng. Grammatical Error Correction with Alternating Structure Optimization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, volume 1, pages 915–923, Portland, Oregon, USA, 2011b. Association for Computational Linguistics. 57, 58
- Robert Dale and Adam Kilgarriff. Helping our own: The HOO 2011 Pilot Shared Task. In *Proceedings of the 13th European Workshop on Natural Language Gen-*

- eration (*ENLG*), pages 242–249, Nancy, France, 2011. Association for Computational Linguistics. 52
- Robert Dale, Ilya Anisimoff, and George Narroway. HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada, 2012. Association for Computational Linguistics. 52
- Hal Daumé III. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 256–263, Prague, Czech Republic, 2009. 50
- Rachele De Felice and Stephen G. Pulman. A Classifier-based Approach to Preposition and Determiner Error Correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, volume 2, pages 169–176, Manchester, UK, 2008. Association for Computational Linguistics. 53, 58
- Rachele De Felice and Stephen G. Pulman. Automatic Detection of Preposition Errors in Learner Writing. *Calico Journal*, 26(3):512–528, 2009. 54, 58
- Markus Dickinson. Generating Learner-like Morphological Errors in Russian. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 259–267, Beijing, China, 2010. Association for Computational Linguistics. 64
- Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. Transition-based Dependency Parsing with Stack Long Short-Term Memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL)*, pages 334–343, Beijing, China, 2015. 123, 126
- Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent Neural Network Grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 199–209, San Diego, California, USA, 2016. 123
- Nava Ehsan and Hesham Faili. Grammatical and Context-sensitive Error Correction Using a Statistical Machine Translation Framework. *Software: Practice and Experience*, 43(2):187–206, 2013. 62
- Anas Elghafari, Detmar Meurers, and Holger Wunsch. Exploring the Data-driven Prediction of Prepositions in English. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 267–275, Beijing, China, 2010. Association for Computational Linguistics. 56, 58
- Luis Espinosa-Anke, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. ExTaSem! Extending, Taxonomizing and Semantifying Domain Terminologies. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 2594–2600, Phoenix, Arizona, USA, 2016. 68
- Stefan Evert. *The statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD Thesis. 2005. 34

- Stefan Evert. Corpora and Collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin, 2007. 34, 35, 36, 67
- Stefan Evert and Brigitte Krenn. Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 188–195, Toulouse, France, 2001. 34
- Mohammed Farghal and Hussein Obiedat. Collocations: A Neglected Variable in EFL. *IRAL-International Review of Applied Linguistics in Language Teaching*, 33(4):315–332, 1995. 2
- Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy, and Noah A. Smith. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1606–1615, Denver, Colorado, USA, 2015. 68
- Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. Morphological Inflection Generation Using Character Sequence to Sequence Learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 634–643, San Diego, California, USA, 2016. 120
- Mariano Felice and Zheng Yuan. Generating Artificial Errors for Grammatical Error Correction. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 116–126, Gothenburg, Sweden, 2014. Association for Computational Linguistics. 6, 65, 133
- Mariano Felice, Zheng Yuan, Øistein E Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. Grammatical Error Correction Using Hybrid Systems and Type Filtering. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 15–24, Baltimore, Maryland, USA, 2014. 57, 58, 59, 60, 65
- Gabriela Ferraro, Rogelio Nazar, and Leo Wanner. Collocations: A Challenge in Computer Assisted Language Learning. In *Proceedings of the 5th International Conference on Meaning-Text Theory*, pages 69–79, Barcelona, Spain, 2011. 43, 47, 51
- Gabriela Ferraro, Rogelio Nazar, Margarita Alonso Ramos, and Leo Wanner. Towards Advanced Collocation Error Correction in Spanish Learner Corpora. *Language Resources and Evaluation*, 48(1):45–64, 2014. 43, 47, 51, 107
- John R. Firth. Modes of Meaning. In J.R. Firth, editor, *Papers in Linguistics, 1934-1951*, pages 190–215. Oxford University Press, Oxford, 1957. 7, 8
- Gerhard Fliedner. A System for Checking NP Agreement in German Texts. In *Proceedings of the ACL Student Research Workshop*, pages 12–17, Philadelphia, USA, 2002. Association for Computational Linguistics. 58
- Jennifer Foster and Øistein E. Andersen. GenERRate: Generating Errors for Use

- in Grammatical Error Detection. In *Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 82–90, Boulder, Colorado, USA, 2009. 6, 63, 66, 132, 133
- Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning Semantic Hierarchies via Word Embeddings. In *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1199–1209, Baltimore, Maryland, USA, 2014. 68, 77
- Yoko Futagi, Paul Deane, Martin Chodorow, and Joel Tetreault. A Computational Approach to Detecting Collocation Errors in the Writing of Non-Native Speakers of English. *Computer Assisted Language Learning*, 21(4):353–367, 2008. 43, 51, 81
- Michael Gamon. Using Mostly Native Data to Correct Errors in Learners’ Writing: a Meta-classifier Approach. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 163–171, Los Angeles, California, USA, 2010. Association for Computational Linguistics. 55, 56, 58
- Michael Gamon, Jianfeng Gao, Chris Brockett, Alexandre Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*, volume 8, pages 449–456, Hyderabad, India, 2008. 55
- Michael Gamon, Claudia Leacock, Chris Brockett, William B. Dolan, Jianfeng Gao, Dmitriy Belenko, and Alexandre Klementiev. Using Statistical Techniques and Web Search to Correct ESL Errors. *Calico Journal*, 26(3):491–511, 2009. 59
- Zhao-Ming Gao. Automatic Identification of English Collocation Errors based on Dependency Relations. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC)*, pages 550–555, Taipei, Taiwan, 2013. 42, 43
- Alexander Gelbukh and Olga Kolesnikova. *Semantic Analysis of Verbal Collocations with Lexical Functions*. Springer, Heidelberg, 2012. 19, 38, 40
- Mandeep Singh Gill and Gurpreet Singh Lehal. A Grammar Checking System for Punjabi. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING): Demonstration Papers*, pages 149–152, Manchester, UK, 2008. Association for Computational Linguistics. 58
- Sylviane Granger. Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In A.P. Cowie, editor, *Phraseology: Theory, Analysis and Applications*, pages 145–160. Oxford University Press, Oxford, 1998. 2
- Alex Graves. Generating sequences with recurrent neural networks. *Computing Research Repository (CoRR)*, *arXiv preprint arXiv:1308.0850*, 2013. 126
- Stefan Th. Gries. 50-something Years of Work on Collocations: What is or should Be Next... *International Journal of Corpus Linguistics*, 18(1):137–166, 2013. 36,

156

- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009. 92
- Michael A.K. Halliday. Categories of the theory of grammar. *Word*, 17:241–292, 1961. 7
- Michael A.K. Halliday. Lexis as a Linguistic Level. In C.E. Bazell et al., editor, *In memory of J.R. Firth*, pages 148–162. Longman, London, 1966. 8
- Na-Rae Han, Martin Chodorow, and Claudia Leacock. Detecting Errors in English Article Usage with a Maximum Entropy Classifier Trained on a Large, Diverse Corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1625–1628, Lisbon, Portugal, 2004. 53, 54, 58
- Na-Rae Han, Martin Chodorow, and Claudia Leacock. Detecting Errors in English Article Usage by Non-Native Speakers. *Natural Language Engineering*, 12(02): 115–129, 2006. 53, 54, 58
- Na-Rae Han, Joel Tetreault, Soo-Hwa Lee, and Jin-Young Ha. Using an Error-Annotated Learner Corpus to Develop an ESL/EFL Error Correction System. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 763–770, Valletta, Malta, 2010. 54, 58
- Franz Josef Hausmann. Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortwendungen. *Praxis des neusprachlichen Unterrichts*, 31(1):395–406, 1984. 2, 7
- Franz Josef Hausmann. Kollokationen im Deutschen Woerterbuch: ein Beitrag zur Theorie des lexicographischen Biespiels. *Lexikographie und Grammatik*, 1985. 10
- Franz Josef Hausmann. Le dictionnaire de collocations. In F.J. Hausmann, O. Reichmann, H.E. Wiegand, and L. Zgusta, editors, *Wörterbücher, Dictionaries, Dictionnaires: An international Handbook of Lexicography*, pages 1010–1019. De Gruyter, Berlin/New-York, 1989. 15
- Ulrich Heid. On Ways Words Work Together-Topics in lexical combinatorics. In *Proceedings of the 6th Euralex International Congress on Lexicography (EURALEX)*, pages 226–257, Amsterdam, the Netherlands, 1994. 15
- Ulrich Heid. Using Lexical Functions for the Extraction of Collocations from Dictionaries and Corpora. In L. Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 115–146. Benjamins Academic Publishers, Amsterdam, the Netherlands, 1996. 15
- Ulrich Heid. Towards a Corpus-based Dictionary of German Noun-Verb Collocations. In *Proceedings of the 8th EURALEX International Congress*, pages 301–312, Liège, Belgium, 1998. 33
- Aurélie Herbelot and Ekaterina Kochmar. ‘Calling on the Classical Phone’: a Distributional Model of Adjective-Noun Errors in Learners’ English. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*,

- pages 976–986, Osaka, Japan, 2016. 44, 60, 65, 66
- Matthieu Hermet and Alain Désilets. Using First and Second Language Models to Correct Preposition Errors in Second Language Authoring. In *Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–72, Boulder, Colorado, USA, 2009. 57
- Matthieu Hermet, Alain Désilets, and Stan Szpakowicz. Using the Web as a Linguistic Resource to Automatically Correct Lexico-Syntactic Errors. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 874–878, Marrakech, Morocco, 2008. 56, 57
- Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780, 1997. 126
- Peter Andrew Howarth. *Phraseology in English Academic Writing: Some Implications for Language Learning and Dictionary Making*, volume 75. Max Niemeyer Verlag, Tübingen, Germany, 1996. 2
- Chung-Chi Huang, Kate H. Kao, Chiung-Hui Tseng, and Jason S. Chang. A Thesaurus-based Semantic Classification of English Collocations. *Computational Linguistics & Chinese Language Processing*, 4(3):257–280, 2009. 18, 38
- Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. SENSEM-BED: Enhancing Word Embeddings for Semantic Similarity and Relatedness. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 95–105, Beijing, China, 2015. Association for Computational Linguistics. 68
- Maria del Pilar Valverde Ibanez and Akira Ohtani. Automatic Detection of Gender and Number Agreement Errors in Spanish Texts Written by Japanese Learners. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 299–307, Bali, Indonesia, 2012. 58
- Kenji Imamura, Kuniko Saito, Kugatsu Sadamitsu, and Hitoshi Nishikawa. Grammar Error Correction Using Seudo-error Sentences and Domain Adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 388–392, Jeju Island, Korea, 2012. 64
- Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, and Hitoshi Isahara. Automatic Error Detection in the Japanese Learners’ English Spoken Data. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 145–148, Sapporo, Japan, 2003. 53, 61
- Marcin Junczys-Dowmunt and Roman Grundkiewicz. The AMU System in the CoNLL-2014 Shared Task: Grammatical Error Correction by Data-intensive and Feature-rich Statistical Machine Translation. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 25–33, Baltimore, Maryland, USA, 2014. 59
- Sylvain Kahane. The Meaning-Text Theory. *Dependency and Valency. An International Handbook of Contemporary Research*, 1:546–570, 2003. 10

- Sylvain Kahane and Alain Polguère. Formal Foundation of Lexical Functions. In *Proceedings of the ACL '01 Workshop COLLOCATION: Computational Extraction, Analysis and Exploitation*, pages 8–15, Toulouse, France, 2001. 17
- Lis W. Kanashiro Pereira, Erlyn Manguilimotan, and Yuji Matsumoto. Automated Collocation Suggestion for Japanese Second Language Learners. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 52–58, Sofia, Bulgaria, 2013. 48, 52
- Adam Kilgarriff. Collocationality (And How to Measure it). In *Proceedings of the 12th Euralex International Congress on Lexicography (EURALEX)*, pages 997–1004, Turin, Italy, 2006. Springer-Verlag. 67
- Ekaterina Kochmar and Ted Briscoe. Detecting Learner Errors in the Choice of Content Words Using Compositional Distributional Semantics. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 1740–1751, Dublin, Ireland, 2014. 44, 45, 46, 49, 51
- Ekaterina Kochmar and Ted Briscoe. Using Learner Data to Improve Error Correction in Adjective–Noun Combinations. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 233–242, Denver, Colorado, USA, 2015. 49
- Ekaterina Kochmar and Ekaterina Shutova. Cross-Lingual Lexico-Semantic Transfer in Language Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 974–983, Berlin, Germany, 2016. 44, 45
- Brigitte Krenn and Stefan Evert. Can We Do Better than Frequency? A Case Study on Extracting PP-Verb Collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse, France, 2001. 34
- Batia Laufer and Tina Waldman. Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Language Learning*, 61(2): 647–672, 2011. 2, 3, 48
- Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. Automated Grammatical Error Detection for Language Learners. *Synthesis Lectures on Human Language Technologies*, 7(1):1–170, 2014. 52
- John Lee and Stephanie Seneff. Automatic Grammar Correction for Second-Language Learners. In *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH - ICSPL)*, pages 1978–1981, Pittsburgh, Pennsylvania, USA, 2006. 56, 59
- John Lee and Stephanie Seneff. Correcting Misuse of Verb Forms. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 174–182, Columbus, Ohio, 2008. 62
- Dekang Lin. Extracting Collocations from Text Corpora. In *Proceedings of the 1st ACL/COLING Workshop on Computational Terminology*, pages 57–63, Montréal, Canada, 1998. 33
- Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning Entity

- and Relation Embeddings for Knowledge Graph Completion. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence and the 27th Innovative Applications of Artificial Intelligence Conference*, pages 2181–2187, Austin, Texas, USA, 2015. 68
- Wang Ling, Chris Dyer, Alan W. Black, and Isabel Trancoso. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1299–1304, Denver, Colorado, USA, 2015a. 124
- Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W. Black, and Isabel Trancoso. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1520–1530, Lisbon, Portugal, 2015b. 123
- Anne Li-E. Liu, David Wible, and Nai-Lung Tsao. Automated Suggestions for Miscollocations. In *Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 47–50, Boulder, Colorado, USA, 2009. 49, 107
- Xiaohua Liu, Bo Han, Kuan Li, Stephan Hyeonjun Stiller, and Ming Zhou. SRL-based Verb Selection for ESL. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1068–1076, Cambridge, Massachusetts, USA, 2010. Association for Computational Linguistics. 50
- Cristóbal Lozano. CEDEL2: Corpus Escrito del Español L2. In C.M. Bretones Callejas, editor, *Applied Linguistics Now: Understanding Language and Mind*, pages 197–212. Universidad de Almería, Almería, 2009. 21, 96
- Igor Mel'čuk. Towards a Linguistic 'Meaning \leftrightarrow Text' Model. In *Trends in Soviet Theoretical Linguistics*, pages 33–57. Springer, 1973. 10
- Igor Mel'čuk. *Vers une linguistique Sens-Texte: leçon inaugurale faite le vendredi 10 janvier 1997*. Collège de France, 1997. 10
- Igor Mel'čuk. Collocations and Lexical Functions. *Phraseology. Theory, Analysis, and Applications*, pages 23–53, 1998. 10, 11
- Igor Mel'čuk. Collocations dans le dictionnaire. *Les écarts culturels dans les dictionnaires bilingues*, pages 19–64, 2003. 10
- Igor Mel'čuk. Tout ce que nous voulions savoir sur les phrasèmes, mais. *Cahiers de lexicologie*, 102:129–149, 2013. xii, 10, 13
- Igor Mel'čuk. Phrasemes in Language and Phraseology in Linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk, and R. Schreuder, editors, *Idioms: Structural and Psychological Perspectives*, pages 167–232. Lawrence Erlbaum Associates, Hillsdale, 1995. 7, 10, 11, 16
- Igor Mel'čuk. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. *Lexical functions in Lexicography and Natural Language Processing*,

- 31:37–102, 1996. 6, 17, 20, 37
- Igor Mel'čuk. Phrasèmes dans le dictionnaire. *Le figement linguistique: la parole entravée*. Paris: Honoré Champion, pages 41–62, 2011. xii, 12
- Lukas Michelbacher, Stefan Evert, and Hinrich Schütze. Asymmetric Association Measures. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 367–372, 2007. 36, 156
- Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting Similarities among Languages for Machine Translation. *Computing Research Repository (CoRR)*, *arXiv preprint arXiv:1309.4168*, 2013a. 68, 77
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119, Montréal, Canada, 2013b. 68
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 746–751, Atlanta, Georgia, USA, 2013c. 68, 69, 77
- Simon Mille, Alicia Burga, and Leo Wanner. AnCora-UPF: A Multi-Level Annotation of Spanish. In *Proceedings of the 2nd International Conference on Dependency Linguistics (DepLing)*, pages 217–226, Prague, Czech Republic, 2013. 19, 79
- Pol Moreno, Gabriela Ferraro, and Leo Wanner. Can we Determine the Semantics of Collocations without using semantics? In *Proceedings of the eLex 2013 Conference*, pages 106–121, Tallinn, Estonia, 2013. 18, 19, 38, 39, 40, 67, 79
- Ryo Nagata, Tatsuya Iguchi, Kenta Wakidera, Fumito Masui, and Atsuo Kawai. Recognizing Article Errors in the Writing of Japanese Learners of English. *Systems and Computers in Japan*, 36(7):54–63, 2005. 57
- Ryo Nagata, Tatsuya Iguchi, Kenta Wakidera, Fumito Masui, Atsuo Kawai, and Naoki Isu. Recognizing Article Errors Using Prepositional Information. *Systems and Computers in Japan*, 37(12):17–26, 2006a. 57
- Ryo Nagata, Koichiro Morihoro, Atsuo Kawai, and Naoki Isu. A Feedback-Augmented Method for Detecting Errors in the Writing of Learners of English. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 241–248, Sydney, Australia, 2006b. Association for Computational Linguistics. 57
- Ryo Nagata, Koichiro Morihoro, Atsuo Kawai, and Naoki Isu. Reinforcing English Countability Prediction with one Countability per Discourse Property. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 595–602, Sydney, Australia, 2006c. Association for Computational Linguistics. 57

- Ian S.P. Nation. *Learning Vocabulary in Another Language*. Ernst Klett Sprachen, 2001. 2
- Nadja Nesselhauf. The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied linguistics*, 24(2):223–242, 2003. 2, 3, 4
- Nadja Nesselhauf. *Collocations in a Learner Corpus*. Benjamins Academic Publishers, Amsterdam, 2005. 2, 3, 48
- Hwee-Tou Ng, Siew-Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 1–12, Sofia, Bulgaria, 2013. 52
- Hwee-Tou Ng, Siew-Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond H. Sussanto, and Christopher Bryant. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 1–14, Baltimore, Maryland, USA, 2014. 52
- David Nott. Marking Students' Written Work: Principles and Practice. In *Good Practice Guide*. Subject Centre for Languages, Linguistics and Area Studies, Southampton, 2008. 107
- Ana Orol-González and M. Alonso Ramos. A Comparative Study of Collocations in a Native Corpus and a Learner Corpus of Spanish. *Procedia—Social and Behavioural Sciences*, 96:563–570, 2013. 2, 3
- Robert Östling and Ola Knutsson. A Corpus-Based Tool for Helping Writers with Swedish Collocations. In *Proceedings of the Nodalida Workshop on Extracting and Using Constructions in NLP*, pages 28–33, Odense, Denmark, 2009. 47, 60, 65
- Taehyun Park, Edward Lank, Pascal Poupart, and Michael Terry. Is the Sky Pure Today? AwkChecker: An Assistive Tool for Detecting and Correcting Collocation Errors. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 121–130, Monterey, California, USA, 2008. ACM. 5, 42, 43, 45, 46, 81
- Darren Pearce. Synonymy in Collocation Extraction. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 41–46, Pittsburgh, Pennsylvania, USA, 2001. 33
- Pavel Pecina. An Extensive Empirical Study of Collocation Extraction Methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18, Ann Arbor, Michigan, USA, 2005. Association for Computational Linguistics. 34
- Pavel Pecina. A Machine Learning Approach to Multiword Expression Extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions*, pages 54–57, Marrakech, Morocco, 2008. 35, 67
- Pavel Pecina. Lexical Association Measures and Collocation Extraction. *Language Resources and Evaluation*, 44(1-2):137–158, 2010. 35

- Alain Polguère. La théorie Sens-Texte. *Dialangue*, 8:9–30, 1998. 10
- Desmond Darma Putra and Lili Szabó. UdS at CoNLL 2013 Shared Task. In *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 88–95, Sofia, Bulgaria, 2013. 64
- Sara Rodríguez-Fernández, Roberto Carlini, and Leo Wanner. Classification of Lexical Collocation Errors in the Writings of Learners of Spanish. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 529–536, Hissar, Bulgaria, 2015a.
- Sara Rodríguez-Fernández, Roberto Carlini, and Leo Wanner. Classification of Grammatical Collocation Errors in the Writings of Learners of Spanish. *Procesamiento del Lenguaje Natural*, 55:49–56, 2015b.
- Sara Rodríguez-Fernández, Luis Espinosa-Anke, Roberto Carlini, and Leo Wanner. Semantics-Driven Recognition of Collocations Using Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 499–505, Berlin, Germany, 2016a.
- Sara Rodríguez-Fernández, Luis Espinosa-Anke, Roberto Carlini, and Leo Wanner. Example-based Acquisition of Fine-grained Collocation Resources. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 2317–2322, Portorož, Slovenia, 2016b.
- Sara Rodríguez-Fernández, Luis Espinosa-Anke, Roberto Carlini, and Leo Wanner. Semantics-Driven Collocation Discovery. *Procesamiento del Lenguaje Natural*, 57:57–64, 2016c.
- Sara Rodríguez-Fernández, Roberto Carlini, and Leo Wanner. Generation of a Spanish Artificial Collocation Error Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan, Accepted for publication.
- Alla Rozovskaya and Dan Roth. Generating Confusion Sets for Context-Sensitive Error Correction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 961–970, Cambridge, Massachusetts, USA, 2010a. Association for Computational Linguistics. 64
- Alla Rozovskaya and Dan Roth. Training Paradigms for Correcting Errors in Grammar and Usage. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 154–162, Los Angeles, California, USA, 2010b. Association for Computational Linguistics. 6, 55, 63, 64, 66, 132, 133
- Alla Rozovskaya and Dan Roth. Joint Learning and Inference for Grammatical Error Correction. *Urbana*, 51:61801, 2013. 60
- Alla Rozovskaya, Mark Sammons, and Dan Roth. The UI System in the HOO 2012 Shared Task on Error Correction. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP*, pages 272–280, Montréal, Canada, 2012. Association for Computational Linguistics. 64, 65

- Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, and Dan Roth. The University of Illinois System in the CoNLL-2013 Shared Task. In *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 13–19, Sofia, Bulgaria, 2013. 64
- Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, Dan Roth, and Nizar Habash. The Illinois-Columbia System in the CoNLL-2014 Shared Task. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 34–42, Baltimore, Maryland, USA, 2014. 60, 64
- Karim Sadeghi. Collocational Differences between L1 and L2: Implications for EFL Learners and Teachers. *TESL Canada Journal*, 26(2):100–124, 2009. 4
- Yu Sawai, Mamoru Komachi, and Yuji Matsumoto. A Learner Corpus-based Approach to Verb Suggestion for ESL. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 708–713, Sofia, Bulgaria, 2013. 50
- Norbert Schmitt. Review Article: Instructed Second Language Vocabulary Learning. *Language teaching research*, 12(3):329–363, 2008. 4
- Chi-Chiang Shei and Helen Pain. An ESL Writer’s Collocational Aid. *Computer Assisted Language Learning*, 13(2):167–182, 2000. 40, 43, 47
- John Sinclair. Beginning the Study of Lexis. In C.E. Bazell et al., editor, *In memory of J.R. Firth*, pages 410–430. Longman, London, 1966. 8
- John Sinclair, Susan Jones, and Robert Daley. *English Lexical Studies: The OSTI Report*. University of Birmingham, 1970. 9
- John Sinclair, Susan Jones, and Robert Daley. *English Collocation Studies: The OSTI Report*. Bloomsbury Publishing, 2004. 9
- Anna Siyanova and Norbert Schmitt. L2 Learner Production and Processing of Collocation: A Multi-Study Perspective. *Canadian Modern Language Review*, 64(3):429–458, 2008. 2
- Jonas Sjöbergh and Ola Knutsson. Faking errors to avoid making errors: Very weakly supervised learning for error detection in writing. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 506–512, 2005. 59, 61, 62
- Frank Smadja. Retrieving Collocations from Text: X-Tract. *Computational Linguistics*, 19(1):143–177, 1993. 33, 67
- Luchen Tan, Haotian Zhang, Charles Clarke, and Mark Smucker. Lexical Comparison Between Wikipedia and Twitter Corpora by Using Word Embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJNLP)*, pages 657–661, Beijing, China, 2015. Association for Computational Linguistics. 77
- Joel Tetreault and Martin Chodorow. The Ups and Downs of Preposition Error Detection in ESL Writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 865–872, Manchester, UK, 2008.

- Association for Computational Linguistics. 54, 65
- Joel Tetreault and Martin Chodorow. Examining the Use of Region Web Counts for ESL Error Detection. In *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, pages 71–78, San Sebastián, Spain, 2009. 54
- Joel Tetreault, Jennifer Foster, and Martin Chodorow. Using Parse Features for Preposition Selection and Error Detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 353–358, Uppsala, Sweden, 2010. Association for Computational Linguistics. 54
- Nai-Lung Tsao and David Wible. A Method for Unsupervised Broad-Coverage Lexical Error Detection and Correction. In *Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 51–54, Boulder, Colorado, USA, 2009. Association for Computational Linguistics. 42, 43, 45, 46
- Joachim Wagner, Jennifer Foster, and Josef van Genabith. A Comparative Evaluation of Deep and Shallow Approaches to the Automatic Detection of Common Grammatical Errors. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 112–121, Prague, Czech Republic, 2007. Association for Computational Linguistics. 62
- Joachim Wagner, Jennifer Foster, and Josef van Genabith. Judging Grammaticality: Experiments in Sentence Classification. *CALICO Journal*, 26(3):474–490, 2009. 62, 63, 66
- Ying Wang and Philip Shaw. Transfer and Universality: Collocation Use in Advanced Chinese and Swedish Learner English. *ICAME journal*, 32:201–232, 2008. 2
- Leo Wanner. Towards Automatic Fine-Grained Semantic Classification of Verb-Noun Collocations. *Natural Language Engineering*, 10(02):95–143, 2004. 32, 37, 38
- Leo Wanner, Bernd Bohnet, Mark Giereth, and Vanesa Vidal. The First Steps towards the Automatic Compilation of Specialized Collocation Dictionaries. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 11(1):143–180, 2005. 38, 40
- Leo Wanner, Bernd Bohnet, and Mark Giereth. What is beyond Collocations? Insights from Machine Learning Experiments. In *Proceedings of the 12th Euralex International Congress on Lexicography (EURALEX)*, pages 1071–1084, Turin, Italy, 2006a. 37
- Leo Wanner, Bernd Bohnet, and Mark Giereth. Making Sense of Collocations. *Computer Speech & Language*, 20(4):609–624, 2006b. 19, 37, 38, 67
- Leo Wanner, Margarita Alonso Ramos, Orsolya Vincze, Rogelio Nazar, Gabriela Ferraro, Estela Mosqueira, and Sabela Prieto. Annotation of Collocations in a Learner Corpus for Building a Learning Environment. In S. Granger et al., editors, *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Presses universitaires de Louvain, Louvain, 2013a. 43, 47, 48

- Leo Wanner, Serge Verlinde, and Margarita Alonso Ramos. Writing Assistants and Automatic Lexical Error Correction: Word Combinatorics. In *Proceedings of the eLex 2013 Conference*, pages 472–487, Tallinn, Estonia, 2013b. 5, 43, 46
- Leo Wanner, Gabriela Ferraro, and Pol Moreno. Towards Distributional Semantics-based Classification of Collocations for Collocation Dictionaries. *International Journal of Lexicography*, 30(2):167–186, 2016. 18, 39, 40, 67, 79
- Stuart Webb, Jonathan Newton, and Anna Chang. Incidental Learning of Collocation. *Language Learning*, 63(1):91–120, 2013. 4
- Jian-Cheng Wu, Yu-Chia Chang, Teruko Mitamura, and Jason S. Chang. Automatic Collocation Suggestion in Academic Writing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 115–119, Uppsala, Sweden, 2010. 5, 42, 50, 81, 107
- Jian-Cheng Wu, Tzu-Hsi Yen, Jim Chang, Guan-Cheng Huang, Jim Chang, Hsiang-Ling Hsu, Yu-Wei Chang, and Jason S Chang. NTHU at the CoNLL-2014 Shared Task. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 91–95, Baltimore, Maryland, USA, 2014. 59
- Junko Yamashita and Nan Jiang. L1 Influence on the Acquisition of 12 Collocations: Japanese ESL Users and EFL Learners Acquiring English Collocations. *Tesol Quarterly*, 44(4):647–668, 2010. 3
- Xing Yi, Jianfeng Gao, and William B. Dolan. A Web-based English Proofing System for English as a Second Language Users. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*, pages 619–624, Hyderabad, India, 2008. 42, 43, 46, 56
- Yang Ying and Marnie O’Neill. Collocation Learning through an ‘AWARE’ Approach: Learner Perspectives and Learning Process. In Andy Barfield and Henrik Gyllstad, editors, *Researching Collocations in Another Language*, pages 181–193. Palgrave Macmillan, UK, 2009. 4
- Zheng Yuan and Mariano Felice. Constrained Grammatical Error Correction Using Statistical Machine Translation. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 52–61, Baltimore, Maryland, USA, 2013. 57, 58, 64
- Alisa Zhila, Wen-Tau Yih, Christopher Meek, Geoffrey Zweig, and Thomas Mikolov. Combining heterogeneous models for measuring relational similarity. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1000–1009, Atlanta, Georgia, USA, 2013. 70

Association Measures

Name	Formula
Conditional probability	$P(w_2 w_1) = \frac{P(w_1, w_2)}{P(w_1)}$
Mutual information	$\sum_{x,y} p(x, y) \ln \frac{p(x,y)}{(p(x)p(y))}$
Pointwise mutual information	$\ln \frac{p(x,y)}{(p(x)p(y))}$
Dice coefficient	$\frac{f(x,y)}{f(x)+f(y)}$
χ^2 test	$\sum_{i,j} \frac{f(i,j) - \hat{f}(i,j)}{\hat{f}(i,j)}$
log-likelihood ratio	$-2 \sum_{i,j} f(i,j) \log \frac{f(i,j)}{\hat{f}(i,j)}$
odds-ratio	$\frac{f(x,y)f(\bar{x},\bar{y})}{f(x,\bar{y})f(\bar{x},y)}$
t-score	$\frac{f(x,y) - \hat{f}(x,y)}{\sqrt{f(x,y)(1 - \frac{\hat{f}(x,y)}{N})}}$
z-score	$\frac{f(x,y) - \hat{f}(x,y)}{\sqrt{\hat{f}(x,y)(1 - \frac{\hat{f}(x,y)}{N})}}$

Table A.1: Frequently used Association Measures

Name	Formula
Rank measure (Michelbacher et al., 2007)	$R(w_2, w_1) = R(w_2 w_1)$
<i>NMI</i> (Bouma, 2009)	$\frac{\sum_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)}}{-\sum_{x,y} p(x,y) \ln(p(x,y))}$
<i>AMM – divergence</i> (MI-based)(Bouma, 2010)	$\sum p(x,y) \log \frac{p(x,y)}{p(x)p_{amm}(y x)}$
<i>NPMI</i> (Bouma, 2009)	$\frac{\ln \frac{p(x,y)}{p(x)p(y)}}{-\ln(p(x,y))}$
<i>AMM – ratio</i> (PMI-based) (Bouma, 2010)	$\log \frac{p(x,y)}{p(x)p_{amm}(y x)}$
<i>NPMI_C</i> (Carlini et al., 2014)	$NPMI_C(\text{collocate}, \text{base}) = \frac{PMI(\text{collocate}, \text{base})}{-\log(p(\text{collocate}))}$
ΔP (Gries, 2013)	$\Delta P(w_1, w_2) = p(w_2 w_1 = \text{present})p(w_2 w_1 = \text{absent})$

Table A.2: “Improved” Association Measures

Artificial error examples

Base substitution

- las **salas** (orig. *ventas*) *netas* consolidadas de la empresa fueron de 26.910 millones de bolívares , lo que representa un incremento en bolívares constantes de 3,8 % respecto al año anterior .
- para mañana se espera el arribo a esta capital de miles de indígenas , quienes realizarán una marcha antes de retornar a sus comunidades para *avisar activamente* [COLLOCATE SUBSTITUTION] en la protesta , informó un vocero del frente por la defensa de la **reina** (orig. *soberanía*) *nacional* y en contra de las privatizaciones .
- la mujer detenida se encuentra en poder de la policía , que *investiga el ocasión* (orig. *caso*) para ubicar a sus presuntos contactos en méxico , señaló la pgr .
- bebeto llegó ayer lunes a río de janeiro sin ocultar su estado depresivo por la pérdida del título del certamen español , por los reclamos de los hinchas del deportivo la coruña , disgustados porque no ejecutó el penal desperdiciado por el serbio djukic , que podría *haber dado un diploma* (orig. *título*) inédito para su club .
- entrevistado por el *político programa* [ORDER] “ panorama ” , de panamericana tv de lima , fujimori dijo que pese a esas acciones de violencia la situación está controlada y que existe un buen seguimiento de sendero tanto en las ciudades como en el **corazón** (orig. *medio*) *rural* .

Collocate substitution

- las circunstancias misteriosas en las cuales logró ser exento suscitaron una **resistente** (orig. *fuerte*) *polémica* durante la campaña electoral de 1992 .

- el barco , bastante antiguo , había chocado contra unos arrecifes en marzo pasado pero sin **hacer** (orig. *causar*) víctimas .
- read acusó directamente a francia , estados unidos , venezuela y Canadá , de **avivar** (orig. *impulsar*) una política tendiente a justificar un eventual desembarco de tropas extranjeras en territorio dominicano para posteriormente invadir haití .
- el gobierno danés mantiene que es éste último , y no la OTAN , quien debe **tener** (orig. *tomar*) la decisión de ordenar un apoyo aéreo ” .
- en Kiev se confirmó el envío de 1.000 a 1.300 miembros de las fuerzas del orden de Ucrania , interpretando esta medida como una decisión de **mandar** (orig. *imponer*) un régimen de administración directa del presidente Kravchuk en la República Autónoma de Tendencias Separatistas .

Base creation

- “ la condición de asilado político que tiene García le otorga un fuero que no es posible vulnerar por una **judgmenta** (orig. *sentencia*) judicial .
- en medio de la somnolencia de toda una sala de críticos , donde de vez en cuando resonaba la butaca que se **closo** (orig. *cierra*) bruscamente ante la partida precipitada de su ocupante , bellísimos paisajes de montaña , diálogos limitados al estricto mínimo y un grupo de guardabosques encargados de custodiar un polovorín .
- respecto a la NORAD , una **structura** (orig. *estructura*) creada por ambos países durante la guerra fría para vigilar los movimientos aeroespaciales en América del Norte , se **posa** (orig. *plantea*) la cuestión [COLLOCATE CREATION] de su prórroga , en 1996 .
- en avisos en todos los diarios , la cemento se quejó de que el municipio adujo reclamos tributarios no contemplados en la ley para clausurar la cantera que la preveía de la **mattera** (orig. *materia*) prima industrial .
- el **selector** (orig. *seleccionador*) nacional ruso , Pavel Sadyrin , comunicó este lunes en Moscú la lista de los 22 jugadores para el mundial de fútbol 1994 que se disputará del 17 de junio al 17 de julio en Estados Unidos .

Collocate creation

- Cárdenas , ante cerca de 500 indígenas reunidos en un rudimentario albergue , en el que las moscas y la insalubridad acompañan a un numeroso grupo de expulsados , achacó el problema a “ la fuerza y acciones de los caciques que toman como pretexto razones **religiousas** (orig. *religiosas*) ” , pero en realidad ejercen posiciones de explotación y despojo de tierras .

- con la reducción de la toxicidad podrán **apply** (orig. *aplicarse*) *dosis* mayores de medicina en busca de mayor eficacia clínica , sostienen .
- la tenista argentina gabriela sabatini soportó hoy jueves otra derrota en una **longa** (orig. *larga*) *serie* de frustraciones , al ser eliminada en tercera ronda del torneo de berlín – válido por el circuito y con 750.000 dólares en premios – por la estadounidense ann grossman que la venció 6-3 y 6-4 .
- estimó que al momento el **higher** (orig. *mayor*) *enemigo* de los sistemas democráticos sin lugar a dudas es la corrupción , que afecta la credibilidad de los pueblos , los que no han podido tener un armónico desarrollo social y económico por causa de este mal .
- figueres , que sucedió en el poder al ex presidente rafael calderón , enfatizó que los costarricenses “ aspiramos a que el conflicto cubano se resuelva por sí solo , ya que creemos en la autodeterminación de los pueblos ” e insistió en emplear *vías* **peacefulas** (orig. *pacíficas*) para *lograr retorno* [DETERMINER DELETION] de cuba al sistema democrático .

Determiner base deletion

- yan peizhei y xu zhihe , dos campesinos de shandong de 35 y 50 años , también *pertenecían a* _ *congregación* protestante cuando *fueron condenados* [COLLOCATE CREATION] a tres años de reeducación por el trabajo en diciembre de 1992 .
- _ *acuerdos fueron firmados* por arafa y por el vice-presidente de la comisión europea , manuel marín .
- “ hemos venido para *ayudar a* _ *pueblo* ” , afirma el policía hassan al zahar , 25 años , que partió siendo niño del campamento vecino de chatti hacia sudán , mientras busca con la mirada familiares y amigos .
- uno de los niños *sufre* _ *fracturas* y heridas en el rostro .
- _ *personas murieron* y los 6.000 habitantes del barrio afectado tuvieron que ser evacuados .

Determiner base insertion

- los soldados israelíes *abren* **un** *fuego* y matan a un joven palestino .
- “ tengo la sensación que es una manera de *dar* **una** *marcha* atrás porque no se siente seguro de sí mismo ” , sentenció .
- los combates no *han alejado del* **un** *peligro* a quienes creían estar refugiados en seguridad y varios de los ruandeses que están en las mil colinas afirman haber recibido informaciones según las cuales serán “ *masacrados* ” antes del fin de semana .

- el cineasta sólo podrá *tener un derecho* de visitar , vigilado , a satchel , de 6 años , el único hijo biológico de la pareja .
- la gente compra por precaución , pero eso no quiere decir que *habrá unos problemas* ” , declaró un gerente de un supermercado ubicado en el centro de la capital .

Government base deletion

- los dos encuentros atrajeron *gran cantidad* _ espectadores , mucho más de lo que reunieron en buenos aires .
- tras recomendar una *mayor vigilancia y defensa* _ los “ valores en los que nosotros creemos ” , morin se refirió a la crisis de representatividad de los partidos políticos que se encuentran “ en estado de esclerosis ” , advirtió .
- los automóviles se mostraron deprimidos por temor de que una suba de la tasas a corto plazo tengan un *efecto negativo* _ las ventas .
- la fia , acosada por la serie de tragedias que *sacuden al mundo* _ la f1 en los últimos 15 dias , decidió comenzar a trabajar para reducir el rendimiento de los autos y aumentar la seguridad de sus pilotos , aunque para ello haya puesto en la picota a muchos de los avances tecnológicos logrados en los últimos años .
- el entrenador carlos alberto parreira *resaltó la importancia* _ los exámenes que los jugadores del equipo auriverde serán sometidos en el icaf , ya que revelarán la condición de cada uno y determinarán el ritmo de entrenamiento del seleccionado .

Government base substitution

- negó que hubiera querido intervenir en el debate , reiterando que únicamente buscó *enfatizar el derecho de* la igualdad consagrado en la nueva constitución .
- *el ministro a* información del gobierno nordista , ahmed al lozi , *anunció* por su parte que , tras varios días de combates , las fuerzas fieles al presidente saleh habían tomado daleh .
- fappiano dictaminó ante la corte suprema de justicia a favor de acceder al pedido de la justicia uruguaya para la extradición de moreira abreu , a pesar de que en argentina *sufre un proceso a* supuesto “ lavado ” de dinero del narcotráfico .
- entretanto , el presidente joaquin balaguer , que no *había hecho referencia* directa *de* las manifestaciones de peña , dijo anoche que el no creía que se pudiera producir un fraude , dijo la tarde del lunes que el no creía que nadie impugnara los comicios “ porque todo el mundo quedará satisfecho de la forma en que se ha celebrado el proceso ” .

- el presidente de funam , raúl montenegro , declaró a la *agencia privada para* noticias dyn que “ el producto de promocionará a través de un lobby verde , como si proviniese de bosques manejados racionalmente , aunque en realidad *se trataría a la simple explotación* [GOVERNMENT COLLOCATE SUBSTITUTION] de un bosque ” .

Government collocata deletion

- el sendero “ rojo ” *se opone _ las negociaciones* de paz que efectúa la dirigencia encarcelada de la organización , encabezada por su jefe abimael guzmán , con funcionarios del servicio de inteligencia nacional .
- las declaraciones de goldenberg y espichán tienen el propósito de *responder _ objeciones* formuladas por la cancillería de chile inmediatamente después de la sentencia .
- los sirios exigen , por su parte , que el estado hebreo *se comprometa _ una retirada* de todos los territorios árabes ocupados en 1967 , especialmente el golán .
- el presidente de brasil , itamar franco , realizará una visita oficial a ecuador a mediados de junio próximo , *accediendo _ una invitación* de su homólogo ecuatoriano , sexto durán ballén , anunció el jueves el canciller ecuatoriano , diego paredes .
- los cambios incesantes de reglamento habían roto el frágil equilibrio de *jugar _ la ruleta* rusa .

Government collocata insertion

- el hombre se plantea entonces buscar y hallar a aquella persona que *escribió en las cartas* y que mantuvo vivo el recuerdo del hermano .
- perez esquivel sostuvo que “ es necesario *hacer a todo un trabajo* de fortalecimiento de las instituciones , pero que existe un punto crítico en la situación democrática que son los procesos de impunidad y las situaciones de corrupción que se notan en muchos países .
- las negociaciones entre el gobierno y los huelguistas para *poner con fin* al movimiento de fuerza prosiguieron este jueves en brasilia .
- el último error , ocurrido el 14 de abril en cielo del kurdistán iraquí , *costó con la vida* a los 26 ocupantes de los dos helicópteros blackhawk , a los que los pilotos de los cazas f-16 confundieron con hind iraquíes de fabricación soviética .
- la justicia le *ordenó entregar de* su pasaporte .

Government collocate substitution

- yo elegí dónde quiero estar y esa es una manera de *participar a la construcción* de nuestras instituciones ” , agregó ashraui .
- por su parte , los dirigentes musulmanes bosnios *reaccionaron* airadamente *en la declaración terminal* [COLLOCATE SUBSTITUTION] de la reunión sobre bosnia de las grandes potencias , que propone atribuir 49 % del territorio a la minoría serbia .
- *duran* “ sólidas *pruebas* [COLLOCATE SUBSTITUTION] de que el crimen organizado intenta sistemáticamente *acceder en los depósitos* de armas nucleares ” en rusia , amenazando la seguridad y el control de las 15.000 ojivas nucleares tácticas de ese país , precisa hersh .
- centenares de habitantes de daleh y sus alrededores van a adén para *huir a los combates* .
- “ nos oponemos a *convertirnos por siervos* modernizados y retornar a la explotación indiscriminada del campesinos ” , agregó la pacari .

Pronoun deletion

- “ los correctivos serán aplicados con la más estricta precisión , porque no podemos *dar_ el lujo* de perder un solo punto más ” , agregó al elogiar el despliegue físico y técnico de los jugadores a quienes observó el domingo cuando barcelona empató sin goles en su propio estadio , el más moderno de sudamérica , ante el deportivo cuenca de la austral ciudad del mismo nombre .
- lo que tenemos que hacer es *rebelar_ contra nuestros dirigentes* ” , comenta ahmed , indignado .
- lo más importante es *dar_ una oportunidad* de que sí queremos entendernos y ayudar a nicaragua ” , subrayó ortega , cerrando así aparentemente el capítulo de confrontación que sostuvo el ejército con el gobierno para llegar a estos acuerdos .
- finalmente , al *propiciar_ la furia* [+ COLLOCATE SUBSTITUTION] con el correr de los meses , lentamente fue tomando forma la idea de un espectáculo musical , donde las canciones fueran el vehículo de la protesta , de la denuncia y , de paso , de la ironía y las “ púas ” con que los intelectuales cariocas primero , secundados luego por paulistas y de las otras metrópolis luego , descargaban las armas de su oposición .
- *el hecho* comenzó a *crear_* [+ COLLOCATE SUBSTITUTION] a las 18,30 gmt cuando el cielo comenzó a oscurecerse provocando alarma en la población .

Pronoun insertion

- y sin titubear , mosley comenzó a enumerarles una por una las medidas de urgencia que hoy iban a ser anunciadas , mientras los dueños de los equipos *se guardaban silencio* como niños que escuchan una reprimenda .
- integrada por alemania , Bélgica , dinamarca , francia , holanda , italia , suecia y suiza , la organización astronómica europea *se consideró la posibilidad* de trasladar el telescopio a namibia cuando surgieron problemas en chile , *admitieron unas fuentes* [DETERMINER INSERTION] de la eso .
- *el índice* de precios mayoristas *se bajó* 0,1 % el mes pasado y el de los precios minoristas subió 0,1 % .
- suárez , quien fue una de las revelaciones de este certamen , *se jugó un* muy buen primer *set* , al punto que coetzer – no en vano la primera favorita del torneo – debió esforzarse a fondo para ganar la manga .
- “ cuando *un avión se aparece* en mi pantalla de radar , mi trabajo consiste en *portar a buen puerto* [COLLOCATE SUBSTITUTION].

Number base

- *fuentes confiables* cifran la fuga de divisas en los primeros cuatro meses del año en más 3.000 millones de dólares , dejando las reservas operativas en 2.700 millones de dólares , aunque sosa asegura que “ superan ampliamente ” los 3.000 millones .
- el 9 de diciembre de 1987 , 24 horas después de la muerte de cuatro palestinos por un camión conducido por un *civiles israelí* cerca del punto de paso de erez , el campamento de jabalia , uno de los más pobres de la franja de gaza , se inflama .
- “ sentimos que instituciones como la oea se han quedado rezagadas ” , recaló naranjo , quien manifestó esperanzas de que el organismo se renueve con la gestión del presidente colombiano César Gaviria , elegido nuevo *secretarios general* de la organización .
- según informó el sábado la agencia azerí turán , grachev anunciará la *próxima semanas* una nueva tentativa de conciliación con sus homólogos de armenia y azerbaiyán , así como con las autoridades armenias del territorio en disputa .
- según ellos , Peña Gómez pretende abrir campos de refugiados en la república dominicana para “ congraciarse ” con potencias como *estado unidos* .

Number determiner

- whyte recibe entonces el *apoyo incondicionales* de la deutsche oper y de su intendente , friedrich goetz .

- tanto el cuerpo técnico que dirige el entrenador alfredo basile como el plantel mundialista argentino , se solidarizaron con maradona y acordaron no viajar a japon , lo que obligó a la afa a reprogramar el calendario de *partidos preparatorio* .
- este jueves , efectivos militares ocuparon el puesto de control de la policia federal en el puente sobre el río paraná , en la ciudad de foz de iguazú , y normalizaron el tránsito de vehículos y de personas entre brasil y paraguay retrasado durante las últimas semanas por la huelga de los *agentes federal* .
- *las policía chilena* detuvo a un centenar de estudiantes , que el jueves se apoderaron del edificio central de la universidad tecnológica metropolitana , al oriente de santiago , informaron fuentes universitarias .
- colamarco era señalado por el gobierno como obstructor del proceso de reestructuración de la institución , centro de *dura críticas* de los afiliados por denuncias de corrupción , mal manejo de fondos y desatención en cuanto a la entrega de *benefitos económicos* [BASE CREATION] y atención médica .

Gender

- no solamente porque es su película sino también porque , junto a la presencia de otras tres cintas latinoamericanas , la mexicana “ la reina de la noche ” , la peruana “ sin compasión ” y la uruguaya “ el dirigible ” , le parece *una excelente augurio* .
- el pronóstico global del piloto sigue siendo “ comprometido , pero su juventud es *una factor favorable* ” añadió el portavoz quien de esa forma racticamente repitió , sin ningún cambio , el informe que había sido suministrado pocas horas antes .
- se trata de la primera expulsión de *refugiados haitianas* desde que el presidente bill clinton anunciara una política menor rigurosa en la rechazo de refugiados haitianos .
- ciento nueve proyectos *conforman la plan* , revelado por el departamento nacional de planeación , que deben desarrollar prioritariamente los 14 ministerios .
- guardaparques y militares ecuatorianos iniciaron el sábado *un decidida acción* para extinguir un incendio que desde hace 33 días arrecia en el sur de la isla isabela , la más grande del archipiélago de galápagos , informó la defensa civil .

Order

- el ex-presidente dominicano juan bosch , un octogenario lider socialdemócrata y a quien las encuestas ubican en un tercer puesto de las preferencias electorales , se encuentra seguro de que será la sorpresa en los *generales comicios* del próximo lunes .
- “ pienso que los mayores violadores de los *humanos derechos* en colombia son los guerrilleros , terroristas y miembros del crimen organizado , los que *se oponen las fuerzas* [GOVERNMENT COLLOCATE DELETION] armadas ” , dijo pardo .
- “ me parece injusta la apreciación de ponernos como país de alto riesgo para la inversión , a la par de bosnia y serbia que están en guerra y otras naciones proclives a *bélicos conflictos* ” , declaró a la prensa local el dirigente de colegio hondureño de economistas , cecilio zelaya .
- ayer domingo , en el principado , williams renault conoció un *premio gran* a imagen de este comienzo de temporada , calamitoso , con la eliminación , tan pronto arrancaron , de su *representante único* , el británico damon hill .
- según la *oficial versión* , sólo hubo una “ suspensión ” de las conversaciones para que las delegaciones abandonaran moscú hacia sus respectivas capitales para consultar a sus gobiernos .