Verb and *–mente* Adverb Collocations in Portuguese

Extraction from Corpora and Automatic Translation into English

A dissertation submitted as part of a programme of study for the award of M.A. in Natural Language Processing and Human Language Technology

LUCAS NUNES VIEIRA

Supervisors:

Prof Jorge Baptista
Dr Izabella Thomas

Universidade do Algarve
Faculty of Human
and Social Sciences

Université de Franche-Comté
UFR Sciences du Langage,
de L'Homme et de la Societé

This project was supported by the European Commission,
Education &Training, Erasmus Mundus: EMMC 2008-0083

Besançon, France
June 2012

Acknowledgments

My warm thanks to my parents for being so supportive of all my endeavours and my decision of studying abroad. I dedicate this dissertation to them.
I would like to thank my supervisor, Prof Jorge Baptista, for his constant support, guidance, his attentive and at times almost immediate feedback and his admirable perspicacity in making suggestions that have been crucial for the development of this project. My sincere thanks also to my co-supervisor, Dr Izabella Thomas, for her valuable comments, her careful reading and the new perspective she always brought to the topics being addressed.
I am deeply thankful to Prof Nuno Mamede and Mr Cláudio Diniz for their immense contribution to this project in all aspects involving the computational processing of Portuguese.
Thanks also to Dr Lucia Specia who has been a constant source of support in this past year and who has introduced Machine Translation to me in a totally new light.
My gratitude to those who, either in Brazil or in Portugal, have so kindly participated in the annotation task.
Thanks also to my Erasmus Mundus classmates for the many moments of laughter that have been so important in striking a balance with all the hard work.
Last but not least, I would like to thank the European Commission for funding this study.

Resumo

Esta dissertação tem por objetivo a investigação do padrão colocacional formado por verbo e advérbio terminado em *–mente* em português em vista de sua extração de *corpora* e sua tradução automática para o inglês. O trabalho envolve o processamento computacional de um corpus do português; o desenvolvimento de um conjunto de regras que permitam um melhor processamento desse padrão, sobretudo resolvendo o problema de coordenação adverbial; um teste da intuição de falantes nativos do português em vista da identificação do valor colocacional do padrão linguístico estudado; uma avaliação da sensibilidade de medidas de associação para a identificação de colocações com este padrão; o desenvolvimento de um classificador automático de colocações com base em métodos de aprendizagem supervisionada; a construção de um léxico bilíngue deste tipo de colocações; e a avaliação da tradução automática deste padrão para o inglês.

Na primeira fase do estudo, um corpus do português de grande porte, o *CETEMPúlico*, composto por 191 milhões de palavras de textos jornalísticos, foi processado computacionalmente por meio da cadeia de processamento STRING, que faz desde a segmentação do texto até sua análise sintática. Nesta fase, uma série de regras com vistas a um melhor processamento de casos de coordenação adverbial em português foram criadas e incorporadas na STRING. Os resultados obtidos para desambiguação de partes do discurso consistem em uma medida-f de 0.724, já para chunking e extração de dependências, uma medida-f de 0.810 foi obtida.

Uma vez processado o corpus, 65.535 dependências sintáticas entre verbo e advérbio terminado em *–mente* foram extraídas. Em seguida, uma série de filtros foram aplicados ao resultado da extração para que fossem excluídos desde o início casos que não apresentavam potencial para formar colocações. Primeiramente, um filtro de frequência que excluía pares que ocorrem menos de 5 vezes no corpus foi adotado. Também foram excluídos bigramas que incluíam verbos de ligação, assim como bigramas que incluíam classes adverbiais que apresentam pouco ou nenhum potencial colocacional. Uma classificação previamente existente de advérbios terminados em *–mente* em português foi utilizada para este fim. Esta classificação foi estendida em aproximadamente 500 advérbios e em seguida incorporada na cadeia de processamento STRING como parte do presente estudo. Uma série de critérios propostos para a classificação de advérbios terminados em *–ment*, em francês, foi tomada como o conjunto de princípios linguísticos que serviram de base para a classificação dos advérbios em português.

Após a fase de filtragem, 5.793 pares de verbo e advérbio terminado em *–mente* restaram da extração. Para que se chegasse a uma lista de colocações deste padrão em português, esses 5.793 pares, considerados como o conjunto de pares-candidatos, passaram por uma classificação manual que etiquetava os pares como "colocação" ou como "não colocação". Uma série de testes linguísticos foram desenvolvidos para a classificação dos pares. O objetivo desses testes era facilitar a identificação deste tipo

de colocação por meio de princípios sintático-semânticos que discutivelmente refletem a existência de um caráter colocacional em um par ou grupo de palavras.

Como resultado da classificação manual, 501 bigramas foram considerados colocações dos 5.793 candidatos. Pôde-se notar que a frequência dos pares no corpus estava de certa forma ligada ao seu caráter colocacional, uma vez que 60 por cento dos pares mais frequentes, contra 8.6 por cento do total de candidatos, foram considerados como casos de colocação.

Para averiguar a intuição de falantes nativos do português a respeito desse padrão colocacional, uma tarefa de classificação foi desempenhada com uma amostra de 30 pares selecionados aleatoriamente da lista de candidatos – 15 tendo sido previamente classificados como colocações, e 15 como não colocações. Vinte e um falantes nativos do português foram recrutados para a tarefa de classificação, dos quais 13 eram falantes nativos do português europeu, e 8 do português brasileiro. Foi possível concluir com o resultado dessa experiência que o padrão colocacional tratado é extremamente problemático no que diz respeito a sua identificação. A medida Kappa de acordo entre anotadores para a amostra de 30 pares foi de 0.06, o que, embora possa ser interpretado como "leve acordo", é ainda discutivelmente um valor consideravelmente baixo. A dificuldade de se explicar o próprio conceito de colocação assim como o tamanho reduzido da amostra seriam algumas das razões para o baixo nível de convergência alcançado.

Haja vista a baixa qualidade dos resultados alcançados com a tarefa de classificação envolvendo falantes nativos do português, uma série de medidas de associação foram testadas em vista do padrão colocacional tratado. Primeiramente, constatou-se que o limiar de referência existente para a análise das medidas "*t* test" e "chi-quadrado" não apresenta resultados satisfatórios na identificação do tipo de colocação tratado. Em seguida, a sensibilidade dessas mesmas medidas, e também de "Informação Mútua", "Log-Likelihood Ratio", "Coeficiente Dice", e "Unigram Subtuples", foi testada com base em sua correlação com a classificação manual dos pares-candidatos. Constatou-se que "Informação Mútua", "Log-Likelihood Ratio", e "Unigram Subtuples" são as medidas de associação com maior correlação com a classificação manual, o que representa um desempenho satisfatório dessas medidas para a identificação do padrão colocacional sob estudo.

Em seguida, técnicas de aprendizagem de máquina supervisionada foram utilizadas para que, a partir do conjunto de pares-candidatos classificados manualmente e seus respectivos valores de medidas associação, fosse possível treinar um classificador automático de colocações. Os resultados alcançados com esta experiência são extremamente promissores. O desempenho de quarenta e cinco classificadores disponíveis na ferramenta de aprendizagem de máquina WEKA foi testado com base em validação cruzada. O classificador que apresentou o melhor resultado foi "RotationForest", que alcançou uma medida-f de 0.816 em um corpus de treino balanceado composto pelos 501 bigramas classificados como colocação, mais outros 501 bigramas classificados como não colocação. A estratégia que consiste em combinar diferentes classificadores por meio do algoritmo "Vote", disponível na ferramenta WEKA, provou ser capaz de melhorar ainda mais os resultados. O desempenho de uma série de combinações foi testado, e o melhor resultado foi alcançado com a combinação "Rotation Forest" e "LMT". Para validar os resultados obtidos, o classificador proveniente da combinação desses dois algoritmos foi testado em um corpus não visto, o *NILC/São Carlos*, consideravelmente menor que o corpus de treino. Considerando os casos de colocação que ocorrem nos dois corpora e excluindo-se casos de *hápax legomena* no *NILC/São Carlos*, o classificador alcançou

uma medida-f de 0.733 para o corpus não visto, o que pode considerado bastante promissor devido à considerável diferença de tamanho entre os dois corpora.

Após os testes com os diferentes métodos mencionados acima, compilou-se um léxico bilíngue português-inglês contendo o padrão colocacional tratado. Três corpora paralelos e um dicionário de colocações foram utilizados como fontes de referência para que versões equivalentes das colocações em inglês fossem estabelecidas. O dicionário adotado, o *Oxoford Collocations Dictionary*, foi considerado como fonte principal já que, mais que apenas ocorrências em um corpus paralelo, entradas em um dicionário de colocações atestam o verdadeiro valor colocacional das combinações em inglês.

Uma vez construído o léxico, as equivalências deste tipo de colocação entre português e inglês foram utilizadas como referencia para a avaliação de três sistemas de tradução automática disponíveis gratuitamente na rede: *Google Translate*, *Systranet*, e *Reverso*. Exemplos do contexto de ocorrência dos pares em português foram extraídos do corpus *CETEMPúblico* e então traduzidos automaticamente para o inglês com esses três sistemas. Foi constatado que a tradução da maioria dos pares é correta no sentido de não infringir regras gramaticais da língua, mas, em contrapartida, a tradução sugerida para a maioria dos pares não reflete uma escolha lexical fluente em inglês. A avaliação da fluência das traduções foi feita tomando-se como referência medidas de associação calculadas para os pares com base em dados de frequência do corpus do inglês *Collins Wordbanks*.

De modo geral, os resultados obtidos com este trabalho demonstram que o padrão linguístico formado por verbo e advérbio terminado em *–mente* impõe uma série de obstáculos a diversos níveis de processamento de linguagem natural, desde desambiguação de partes do discurso até tradução automática. A identificação do valor colocacional deste padrão também mostrou-se problemática, sobretudo quando a classificação de diversos anotares, ainda que falantes nativos do português, é considerada. Por fim, espera-se que os métodos testados no decorrer desta pesquisa possam não somente servir a um melhor tratamento computacional do padrão estudado em português, mas que possam também ser replicados a outros problemas linguísticos, sobretudo àqueles relacionados a termos compostos e expressões multipalavra em geral.

Palavras-chave

Colocações, Processamento de Linguagem Natural, Advérbios terminados em *-mente*, Medidas de Associação, Tradução Automática

Abstract

This dissertation aims at investigating verb and *–mente* (‘-ly’) adverb collocations in Portuguese (e.g. *vencer confortavelmente*, 'win comfortably') in view of their extraction from corpora and their automatic translation into English.

The main objectives of the study are to exploit a syntax-based approach to collocation extraction in order to assess the performance of different association measures in capturing collocations, as well as evaluate the performance of Machine Translation systems in view of the linguistic pattern dealt with.

To this aim, an existing syntactic-semantic classification of Portuguese *-mente* adverbs was substantially extended; a set of disambiguating, chunking and parsing rules were developed and integrated in an operating rule-based natural language processing chain; these rules were particularly aimed at dealing with the complex phenomenon of adverb coordination and reduction in Portuguese; an automatic collocation classifier was built, using Machine Learning techniques; and a bilingual PT>EN lexicon was compiled.

Results from this investigation show that the sparsity of the phenomenon makes it difficult to retrieve, even from large sized corpora. It also showed the subtle nature of this collocational pattern, which constitutes a serious challenge for existing MT systems, still unable to capture the fluency of natural language.

Keywords

Collocations, Natural Language Processing, *mente* (‘ly’) adverbs, Association Measures, Machine Translation

Résumé

Ce mémoire a pour objectif d'examiner les collocations portugaises formées par des verbes et des adverbes terminés en *–mente* ('-ment') (par exemple *vencer confortavelmente*, 'gagner confortablement') en vue de leur extraction des corpus et de leur traduction automatique vers l'anglais.

Les objectifs principaux de cette étude consistent à exploiter une approche d'extraction des collocations basée sur les critères syntaxiques afin d'évaluer la performance des différentes mesures d'association, ainsi que la performance des différents traducteurs automatiques pour le type de combinaisons étudiées.

Pour arriver à ces objectifs, une classification syntaxiquo-sémantique existante des adverbes terminés en *–mente* ('-ment') en Portugais a été reprise et amplement étendue ; un ensemble de règles de désambiguïsation, de 'chunking', et de 'parsing' a été intégré à une chaîne de traitement automatique du portugais déjà existante, basée sur des règles ; ces règles ont eu pour but de traiter le phénomène complexe de réduction et de coordination des adverbes en portugais ; un classificateur automatique de collocations a été construit en s'appuyant sur les techniques d'apprentissage automatique et un lexique bilingue portugais-anglais a été compilé.

Les résultats de cette investigation montrent que la rareté du phénomène le rend difficile à extraire, même d'un grand corpus. Il a été montré aussi que la subtilité de ce type de collocations constitue un défi sérieux pour les traducteurs automatiques existants, qui sont encore incapables de saisir la fluidité de la langue naturelle.

Mots Clés

Collocations, Traitement Automatique des Langues, adverbes en *-mente* ('ment'), Mesures d'Association, Traduction Automatique

Table of Contents

List of Abbreviations

Adv-mente	Adverbs ending in *–mente* (*-ly*)
AM	Association measure
EMMC	Erasmus Mundus Masters Course
NLP	Natural Language Processing
HLT	Human Language Technology
CETEM	Corpus de Extractos de Textos Electrónicos MCT (Ministério da Ciência e Tecnologia)
MI	Pointwise Mutual Information
STRING	Statistical and Rule-based Natural Language Processing Chain
INESC-ID	Instituto de Engenharia de Sistemas e Computadores – Investigação e Desenvolvimento
L2F	Laboratório de Sistemas de Língua Falada
POS	Part of-speech
XIP	Xerox Incremental Parser
MOD	Modifier (syntactic dependency)
PT	Portuguese
EN	English
ES	Effect size
MT	Machine Translation
MI	Pointwise mutual information
UAP	Uninterpolated Average Precision
SMT	Statistical Machine Translation
LEXMAN	Lexical Morphological Analyser
LLR	Log-likelihood ratio
MARv	Morphossyntactic Ambiguity Resolver
ADVP	Adverbial phrase
NP	Noun phrase
SC	Sub-clause
QUANTD	Quantifier (syntactic dependency)
S1	Subset of collocation candidates with frequency higher than 100
S2	Subset of collocation candidates with frequency between 99 and 10
S3	Subset of collocation candidates with frequency lower than 10
SUBJ	Subject (syntactic dependency)
CDIR	Direct object (syntactic dependency)
DETD	Determinant (syntactic dependency)
NE	Named entity
MAIN	Main element of the sentence
UnigSub	Unigram subtuples
VDOMAIN	Verb domain (syntactic dependency)
PA	Disjunctive adverbs of attitude
PC	Conjunctive adverbs
PS	Disjunctive adverbs of style
MF	Focus adverbs

List of Tables

Chapter 1. Introduction

Collocations started to be a target of research in the twentieth century after Firth (1957) coined the term and called attention to the fact that the way we combine words in natural language is far from being unconstrained.

Certain verb-adverb combinations have collocational status in the sense of Firth, namely *chorar copiosamente* ('to cry copiously'), *dizer textualmente* ('to say textually'), *criticar duramente* ('to criticise harshly'), to cite a few. In the example:

(1) *O professor criticou duramente o aluno*
'The teacher criticised the student hard-ly'

the combination of the verb *criticar* ('to criticise') with the adverb *duramente* ('hard-ly') is considered a collocation in the sense of Firth (1957) since the frequency of the two words together is relevant to establish their collocational status. In this line of reasoning, the probability for the co-occurrence of this pair significantly exceeds chance levels. In the sense of Mel'čuk (2003), the adverb functions as a modifier of *criticar* ('to criticise'), but its choice is not arbitrary and depends on the main verb. From this perspective, the modifying value attributed to the adverb can be seen as a lexical function of the verb, and its collocational status must have a distributional counterpart that should be empirically measurable in large-sized corpora. The theoretical ground of this study profits from both these senses, since at different stages of the research both frequency of distribution and purely linguistic principles are used to extract and classify collocations.

Concerning the verb-adverb pair in (1), when looking for the distribution of *duramente* in a European Portuguese news corpus of 197,2M words, the *CETEMPúblico*[1], one finds a total of 481 occurrences of this adverb accompanied by a verb. Out of this total of occurrences, 111 are with the verb *criticar* ('to criticise'). This seems to corroborate the idea that this pair holds collocational status in the corpus.

Church and Hanks (1989) are among the first authors to develop statistical tools to help lexicographers in the task of collecting collocational patterns based on

[1] http://www.linguateca.pt/cetempublico/ [Accessed 15 May 2012]

distributional data derived from large-sized corpora. Manning and Schütze (2003: 151-189) present and compare statistical association measures to assess the degree of fixedness of word combinations. Among these measures are, for example, the Student's *t* test (Fisher, 1925), Pearson's Chi-Square (χ^2) (Pearson, 1900), and Mutual Information (MI) (Fano, 1961). Table 1.1 shows results obtained by applying the measures just mentioned to a selection of the top ten verb-adverb occurrences including *duramente* ('harshly') within a three-word window in the *CETEMPúblico* corpus. The verbs *ser* ('to be'), *ter* ('to have'), *estar* ('to be'), *poder* ('can'), and *fazer* ('to do') were disregarded in the search due to their little semantic content and consequent slim potential of forming collocations.

Verb (count)	Bigram count (*duramente*: 1126)	*t* test	χ^2	MI
criticar (18581)	111	*10.525*	*112432.3*	9.986
trabalhar (46984)	38	*6.119*	*5146.443*	7.102
atacar (13372)	14	*3.720*	*2462.186*	7.474
atingir (2189)	12	*3.460*	*11151.08*	9.863
ir (43875)	12	*3.389*	*533.917*	5.537
lutar (12845)	7	*2.617*	*634.141*	6.532
penalizar (3531)	5	2.226	*1192.779*	7.910
condenar(19033)	5	2.185	*213.264*	5.479
reprimir (1089)	4	2.233	*3889.812*	9.607
combater(10746)	4	1.968	*245.007*	5.982

Table 1.1 Verb-adverb distribution of *duramente* in the *CETEMPúblico* corpus along with statistical association measures

The *t* test and χ^2 are hypothesis testing statistical measures that have a pre-established threshold serving as a parameter to the statistical relevance of the results. MI, on the other hand, relies mostly on ranking and is subject to a more case-specific interpretation.

At a probability level of $\propto = 0.005$, the critical value for the *t* test is 2.576. As for the χ^2, considering a probability level of $\propto = 0.05$, its critical value is 3.841.

It can be observed in Table 1.1 that *criticar duramente* ('to criticise hard-ly'), a pair that can be considered to hold collocation status in Portuguese, has crossed the

statistical relevance threshold for both the *t* test and the χ^2. It has also reached the highest MI value in comparison with the other bigrams in the table.

Nevertheless, the *t* test is known for yielding less reliable results in some situations due to the fact that is assumes a normal distribution of probabilities (Church and Mercer, 1993: 20). The χ^2 has been reported in the literature as a more appropriate measure in that respect (Manning and Schütze 1999: 158). However, it is also known that this measure overemphasises low-frequency events (Kilgarriff, 1996: 35), which results of this brief experiment would suggest. As it can be seen in the Table, all bigrams reached the critical value of the χ^2.

As to MI, it can be noticed that its highest values were in fact associated with pairs that can be considered interesting with respect to their collocational value, namely pairs including the verbs *criticar* ('to criticise'), *atingir* ('to hit'), and *repreender* ('to reprimand'). This seems to be indicative, in some degree, of the promising potential of this measure in capturing the collocational pattern this study addresses.

1.1 Objectives and Methods

The first objective of this investigation is to automatically acquire statistically relevant verb-adverb combinations to build a Portuguese-English collocation dictionary. The *CETEMPúblico* corpus is used as the source of distributional data. This is, to the best of our knowledge, the largest publicly available and freely distributed corpus of Portuguese. The scope of this project was limited to morphologically derived, *-mente* ('-ly') ending adverbs, henceforth *Adv-mente*. Combinations of {V, *Adv-mente*} could have been considered understudied in Portuguese hitherto, especially in respect to their collocational potential. This has highly motivated this choice of topic.

Albeit constituting just over 10% of all simple adverb occurrences in the corpus, *Adv-mente* represent in fact the majority of the simple-word lemmas of this grammatical class. Table 1.2 shows details of the frequency of adverbs in the corpus.

CETEMPúblico	
lemmas (l)	1,2M
words (w)	191,6M
Adv (l)	5,361
Adv (w)	9,1M
Adv-mente (l)	4,654
Adv-mente (w)	1,0M

Table 1.2 Adverbs in the *CETEMPúblico* corpus: (l) lemmas, (w) words.

Relevant verb-adverb combinations are based on the syntactic relation these words have in a sentence and not on their mere co-occurrence or adjacency. For example:

(2) *O Pedro leu o livro atentamente e resumiu-o.*
'Peter read the book attentively and summarised it'

In this sentence, a correct syntactic relation should be established between the adverb *atentamente* ('attentively') and the verb *ler* ('to read'), a combination that could be deemed to have collocational status in Portuguese. There is no direct relation, in this case, between *resumir* ('to summarise') and the adverb *atentamente*, a pair that could erroneously come up in a 3-word window search for surface bigrams in the corpus.

Furthermore, many *Adv-mente* are not, in any context, directly connected to a verb, e.g.:

(3) *A biblioteca era composta principalmente por livros de História*
'The library was composed mainly of books of History'

In this case, the adverb functions as a focus determiner (Molinier and Levrier, 2000: 273-292, Baptista and Català, 2009) on the prepositional phrase, therefore the co-occurrence of the verb and the adverb in the same sentence is irrelevant in view of the discovery of collocational patterns.

Adverbs with scope on the entire proposition (or sentence) rather than on the main verb (or predicate) of a sentence should also be noted, e.g.:

(4) *Curiosamente, o Pedro disse isso*
'Curiously, Peter said this'

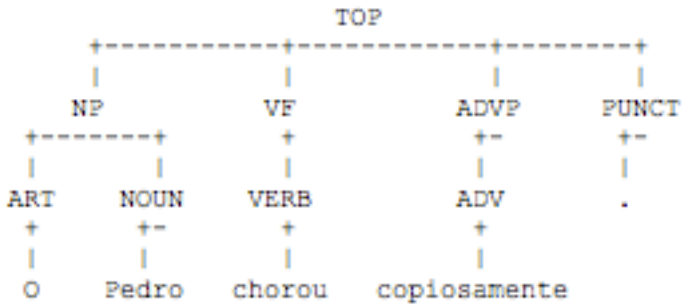
In this case, even if for parsing purposes the adverb can be said to modify the main verb, a more linguistically appropriate representation would have it operate on the sentence as a whole. The proposition in sentence (4) would be the equivalent of *Eu acho curioso que o Pedro tenha dito isso* ('I find it curious that Peter has said this'). In cases of this kind where adverbs are sentential modifiers, {V, *Adv-mente*} combinations are also irrelevant for an assessment of collocational status.

Because of cases such as the ones just described, a more sophisticated process for extracting {V, *Adv-mente*} combinations from corpora is required, based on the correct syntactic parsing of the text and on the extraction of bigrams that actually hold a dependency relation.

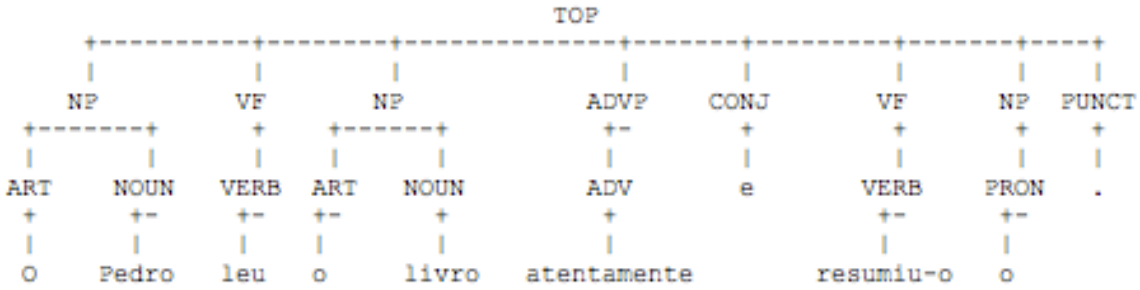
STRING (Mamede et al., 2012) is a text processing chain developed at L2F-INESC ID Lisboa that is able to process large-sized corpora in a robust way that has been adopted in this study. In broad terms, the chain comprises three main stages: pre-processing, disambiguation, and syntactic analysis, respectively. The pre-processing stage is responsible for text segmentation, for part of-speech (POS) tagging and for the chunking of the input into sentences. In the POS disambiguation stage, a rule-driven and a statistical tool perform the disambiguation of tokens. In the last stage, the syntactic parsing of the text is performed by XIP (Xerox Incremental Parser) (Aït Mokhtar et al., 2002), a rule-based parser that establishes syntactic dependencies between words.

In this framework, a dependency relation (called MOD[fier]) is extracted for (1) and (2), with the correct pair of {V, *Adv-mente*}, whereas a determinative focus relation is obtained for (3). For example:

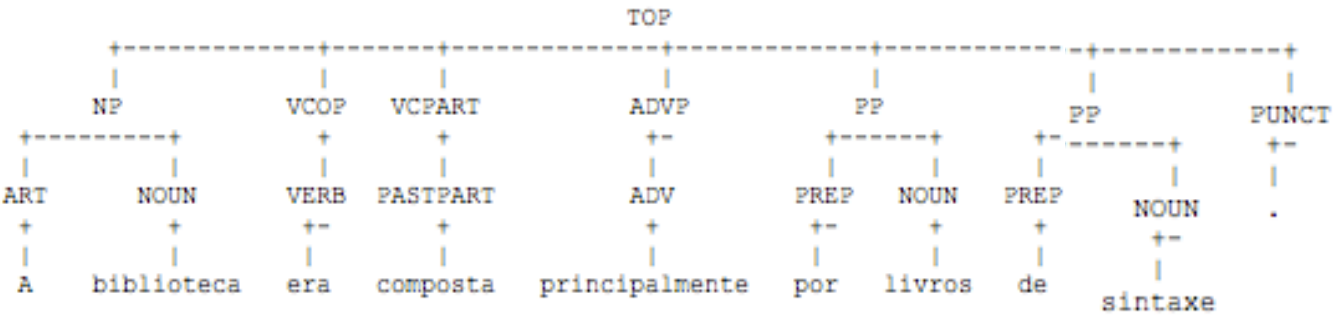
(1) MOD_POST (chorou, copiosamente)



(2) (MOD_POST (leu, atentamente)



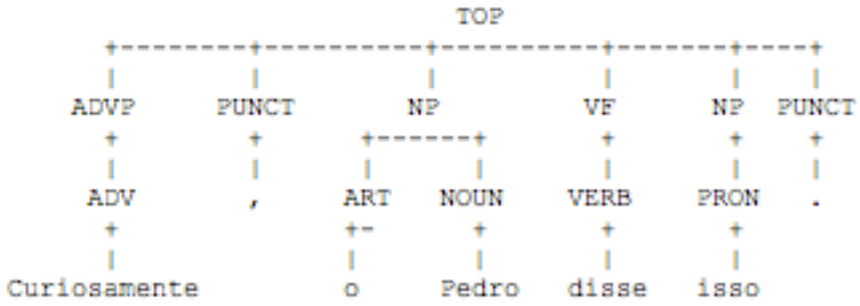
(3) MOD_PRE_FOCUS (livros, principalmente)



Preliminary observations, however, have shown that the rule-based grammar, in some cases, is still unable to correctly establish all the MOD verb-adverb dependencies. Therefore, another objective of this study was to improve the rules of the XIP-L2F grammar.

In the case of sentence (4), for instance, where the adverb has scope on the entire sentence, the resulting dependency provided by XIP is not entirely adequate, since it is represented as a relation between the main verb and the adverb:

(4) MOD_PRE (disse, Curiosamente)



Similarly, cases of adverbial coordination also pose a problem. For example:

(5) *O Pedro fez isso lenta e cuidadosamente*
'Peter did this slowly and carefully'

In (5), the adverb *cuidadosamente* ('carefully') is in coordination with the term *lenta* ('slowly'), which is an adverbial form reduced of the suffix *–mente* ('-ly'). This consists in fact of two adverbs modifying a single verb. The resulting dependency provided by XIP for cases of this kind has been improved as a result of this investigation. In addition, an existing syntactic-semantic classification of *Adv-mente* for Portuguese (Fernandes, 2011) has been substantially extended and incorporated in the STRING chain. This enables the identification of sentence modifying adverbs as in (4).

After processing the corpus and extracting {V, *Adv-mente*} combinations that are syntactically connected, a manual classification of collocation candidates is carried out, and the intuition of native speakers on the collocational value of this pattern is tested through a small-scale annotation task conducted with native speakers of either Brazilian or European Portuguese.

Even though preliminary results of statistical association measures presented in Table 1.1 point to MI as a promising choice in the task of capturing the {V, *Adv-mente*} collocation pattern, more extensive experiments are required in this

respect so that more decisive conclusions are drawn. In an attempt to respond to this necessity, we also assess the performance of different statistical association measures in capturing the {V, *Adv-mente*} collocation pattern. Based on results of these measures, we further experiment with training an automatic collocation classifier using Machine Learning techniques.

Finally, because collocations often pose difficulties to translation, we set out to investigate if there is any correlation between the collocational status of {V, *Adv-mente*} combinations and the English translations provided for them by commercial Machine Translation (MT) engines available to the general public.

In the remainder of this dissertation, we discuss related work in Chapter 2. In Chapter 3, we present the corpus processing stage and the process for developing and testing a set of disambiguating, chunking and parsing rules that have been integrated in the STRING chain. We describe the experiment aimed at building an automatic collocation classifier in Chapter 5. In Chapter 6, we explain how the bilingual Portuguese-English collocation lexicon was built and also present the methodology for evaluating MT engines as well as results of the evaluation. And finally, in Chapter 7, we conclude by overviewing the findings and general contributions of the research and proposing future work in the field.

Chapter 2. Related Work

This Chapter is devoted to a brief review of what has been discussed in the literature concerning four main topics related to this study: the notion of collocation, *Adv-mente*, extraction of collocations from corpora and, finally, automatic translation of collocations.

2.1 The Notion of Collocation

Since Firth (1935) coined the term collocation, this subject has received considerable attention in the field of Linguistics, being a constant topic for discussion and research. The definition of a collocation, however, is even nowadays far from getting to a consensus between specialists in the area. Looking up the term in the Oxford Concise Dictionary of Linguistics (Matthews, 2007: 63), one finds: "a relation within a syntactic unit between individual lexical elements [...] used specially where words specifically or habitually go together". Sinclair (1991: 170) affirms that a collocation is "the occurrence of two or more words within a short space of each other in a text". Even though these definitions may seem to suffice, when dealing with specific cases there is still disagreement concerning what binds these elements together, as well as which cases should or should not receive the label of collocation.

Nevertheless, the acknowledgement of collocations as an extremely important notion for a number of purposes related to an adequate use of natural language is common ground. The knowledge of how words are combined in a way that sounds natural and smooth to the ears is an artefact sub-areas of linguistics highly profit from, such as foreign language teaching, and Natural Language Processing (NLP). NLP, specifically, has a number of sub-fields that would be more directly beneficed by information on collocations. Natural Language Generation, parsing, and corpus linguistic research would be among them, for instance (Manning and Schütze, 1999).

Due to the wide spectrum of applicability the knowledge of collocations presents, specialists and research teams around the world have been experimenting with different ways of retrieving word combinations from corpora in an attempt to compile lists of collocations, or collocation lexicons, and devise strategies to

incorporate them in NLP engines in order to improve the quality of the results. Examples of such compilations are the *Oxford Collocations Dictionary* (Oxford, 2009), the *Macmillan Collocations Dictionary Book* (Macmillan, 2010), and the multilingual collocation dictionary (Cardey et al, 2006) developed in the framework of the *MultiCoDiCT* project[2]. However, before engaging in a study like the one here proposed, it is important to present the different notions concerning the concept of collocation and how it has been understood in previous related work.

Manning and Schütze (1999: 151-189), in their description of methods for extracting collocations from corpora, reiterate three criteria that have been commonly taken into account when defining a collocation. The first one is the *non-compositionality* criterion, according to which the meaning of a collocation would not directly derive from the meaning of its components, ranging from stricter cases – where the meaning of the combination is totally distant from the meaning of its individual words – to less strict ones – where the meaning attributed to the combination does not differ completely from the meaning of the words isolated, but still fails to be their sum. The second criterion, *non-substitutability*, states that it is not possible to substitute any of the components of the collocation, not even by words that would have an equivalent meaning in other contexts. The third and last criterion, *non-modifiability*, states that the collocation would not be able to be modified, either structurally or with the insertion of lexical elements.

Albeit very recurrent in the literature, these criteria do not comprise all cases that could be considered to have collocational value. Evert (2005: 15-18) mentions two different approaches to the notion of collocation: the *distributional* approach and the *intensional* approach. The former would be more closely related to Firth's notion of collocation (Firth, 1957), inherited and further developed by his successors, forming what is commonly referred to in the literature as the Neo-Firthian school (Evert, 2008). This group would regard collocations as word combinations that are recurrent in the language, words that are frequently used together. The intensional approach, in turn, would take into account more than just co-occurrence. It is based on the assumption that a collocation is, in fact, a lexical phenomenon in which a word "collocates" another. More specifically, there would be a free choice element in the combination, called the *base* word, and another element that would be lexically

[2] http://tesniere.univ-fcomte.fr/multicodict.html [Accessed 15 May 2012]

determined by the *base,* called the *collocate*. This notion has been elaborated by Mel'čuk (2003), who also explains how these combinations can be analysed in terms of lexical functions. In broad terms, lexical functions consist of a type of formalism that expresses how words can be combined with other words based on lexical properties they possess. These functions would allow word combinations to be formalised, and processed by computers.

Mel'čuk (2003) establishes that instances in which the meaning of the combination cannot be directly derived from any of its components isolated is, in fact, an *idiom*, and also that combinations whose meaning is possible to be derived from its components and yet none of them is dominant in the combination are cases of *quasi-idioms*. For this last category, the author provides the example in French of *bande dessinée* ('comic strip'), whose meaning is both related to *bande* ('strip') and *dessinée* ('drawn'), but still none of these elements alone is capable of conveying the specific meaning of *bande dessinée*, a sequence of *drawings* arranged in *strips* displaying some type of narrative that is often humorous. Mel'čuk (2010) defines still the concept of *cliché*, which would be a compositional expression whose elements are chosen non-arbitrarily, forming what could be regarded as a single textual entity (Mel'čuk 2010: 4). The author also proposes the more general dichotomy between *syntagme libre* ('free phrase') and *syntagme non libre* ('non-free phrase') (Mel'čuk, 2010). The former would include utterances that are entirely arbitrary, whereas the latter consists of expressions in which the choice of at least one of its components is constrained. Collocations would be included in this last group.

In this way, the type of word combinations explored in this study can be deemed to be very close to what Mel'čuk (2003, 2010) defines as a collocation. In pairs of verbs and derived *Adv-mente*, the verb would be the freely chosen *base* of the combination, with certain types of *Adv-mente* possibly playing the role of its *collocate*. In the pair *chorar copiosamente* ('to cry copiously'), for example, *chorar* ('to cry') is a verb chosen by the speaker to express the act of "crying". In regard to the adverb *copiosamente* ('copiously'), it can be argued that its choice is not arbitrary, from all the adverbs that could modify *chorar* ('to cry'). Conversely, its choice would be controlled by the verb, forming a pair that is common in Portuguese to express the act of crying abundantly. That is not to say, however, that other adverbs could not convey this same meaning. Adverbs transmitting the idea of "large amounts" or "excess" in a

flux could be correctly employed with the verb *chorar* ('to cry'), but the choice of *copiosamente* ('copiously') seems to be one of the best in terms of fluency and naturalness. It is among the options that better *collocate* with this verb.

This project profits both from Firth's and Mel'čuk's notions of collocation, since, at different stages, it relies both on frequency of distribution and on meaning-oriented human annotations. Whilst using statistical measures to assess how frequent a word combination is in the language, based on a sufficiently large-sized corpus, one would be presupposing the Firthian notion of collocation as words that appear in the lexicon together more often than by chance. Notwithstanding, cases covered by Mel'čuk's explanation would arguably be still expected to be found in corpora. In other words, the line of reasoning proposed here bases on the assumption that pairs that can be linguistically classified as collocations in the sense of Mel'čuk (2003), albeit not necessarily frequent by definition, tend to be used frequently in the language. That does not mean, however, that a linguistic analysis should be discarded. On the contrary, such an analysis would be in charge exactly of validating (or not) the statistical results.

A view that profits both from Mel'čuk's and Firth's senses of collocation is very close to what is discussed by McKeown and Radev (2000: 508). They establish that collocations would stand at an intermediary point in a spectrum that has free-word combinations at one extremity and idioms at the other, i.e. the least and most constrained possibilities within the range, respectively. In effect, they give credit to the frequency of co-occurrence in a definition of collocation, but reinforce that isolated words with a high overall frequency should not be taken into account, making it clear that both linguistic-dependent and linguistic-independent factors should be considered in a definition of collocations for NLP. Words with high overall frequency may simply happen to be frequent due to their functionality, as in grammatical classes such as prepositions and conjunctions.

Moreover, McKeown and Radev (2000: 511) draw a distinction between two specific types of collocation: grammatical and semantic. Grammatical collocations would contain closed-class words in their composition, often including syntactic pairs, such as verb + preposition (*get off*, *pull over,* etc.). Semantic collocations would be word combinations only lexically restricted, as in *running commentary*, *commit*

treason, etc. The pattern here under study, verb-adverb pairs, would be very closely related to the second type.

In view of the array of definitions and classifications exposed above, one can easily note that the concept of collocation is still rather loose and non-straightforward in Linguistics. In an attempt to address this lack of specificity, Choueka (1988), reiterated by Evert (2005), affirms, after all, that a good parameter to classify a combination as a collocation is to ask oneself if it should deserve an entry in a lexicon or dictionary. If it does, then it could be called a collocation. From this perspective, pairs of verb and *Adv-mente* could be regarded as collocations in that their presence in a lexicon can be considered of great relevance for a number of applications. MT, speech generation, word sense disambiguation, and other NLP tasks of the like should be able to account for the fixedness between these pairs in order to provide results that are closer to real utterances in natural language.

2.2 *Adv-mente*

In view of the collocation pattern that this project addresses, what follows is a brief overview of how adverbs in general, and more precisely *Adv-mente*, are regarded in terms of their possible morphological and syntactic classification.

The characterisation of adverbs in general is far from being evident and clear-cut. Molinier and Levrier (2000: 23), in French, affirm that adverbs are in fact a residual class, defined as non-prepositions, non-conjunctions and non-interjections, and sharing with these the property of being morphologically invariable.

Bechara (2003), in Portuguese, also highlights the lack of specificity that underlies the grammatical class of adverbs. He calls attention to the fact that much of what accounts for this vagueness is the virtually unconstrained mobility adverbs have in the speech, which would be closely related to the different functions and syntactic roles the adverb can fulfil in a sentence.

In English, the unspecific character of adverbs is also taken into account, as it can be seen in Quirk et al. (1985: 438), who describe them as a "puzzling" and "nebulous" class. The authors acknowledge as tempting the posture of simply affirming that adverbs consist of everything that does not fit into any other grammatical class.

Concerning their syntactic behaviour, a common first distinction that is made in relation to adverbs is the one between *adverbs* proper and *objects*. The former would play the role of accessorising a verb, by modifying its meaning, whereas the latter would be in fact a syntactic notion, having the function of an argument. This classification can be observed in Molinier and Levrier (2000: 25), and Gross (1986: 13), who states that the distinction between *adverbes généralisés* and *objets*, i.e., adverbs and arguments, lies in the fact that arguments have a straighter relation to the verb than adverbs, being in fact more dependant on the verb or even selected by it.

With regard to *Adv-mente,* specifically, Molinier and Levrier (2000) have extensively compiled a repertoire of these word forms in French based on three respected French dictionaries, namely *le Trésor de la Langue Française*, *le Grand Larousse*, and *le Grand Robert*. In doing so, they classified such forms in nine main syntactic-semantic classes, establishing the most important linguistic traces that account for their distinction. In fact, they first group the adverbial forms into two main classes: adverbs with scope on the entire sentence, and adverbs that are an integrated part of a clause. The former is further subdivided into three subcategories, while the latter is subdivided into six, resulting in a total of nine subcategories altogether. Based on this classification in French, Fernandes (2011) has carried out an equivalent classification of *Adv-mente* for Portuguese. The list of classified adverbs produced by Fernandes has been substantially extended as part of the present study.

Portuguese and English grammars also seem to account for the difference between adverbs that are part of the proposition and adverbs that modify entire sentences. Bechara (2003: 292) refers to the phenomenon that allows adverbs to function on the sentential level using the Portuguese terms *hipertaxe* or *superordenação*. These would be a type of grammatical structuring that make it possible for a term that belongs to a lower syntactic level to perform an autonomous role in upper levels. He specifically addresses *Adv-mente* in this respect, remarking that they can even work as an entire sentence, as in the example below:

(6) *Certamente!*
'Certainly!'

Bechara (id: ibid.) also mentions that this phenomenon would be related to the concept of *antitaxe,* in Portuguese, which concerns the reference or substitution of units that are already present in an utterance, even if implicitly. In (6), for example, despite the fact that the adverb *certamente* ('certainly') is itself the entire sentence, if regarded in context, it would make reference to previously mentioned linguistic units that need not be repeated in a second utterance but that are, in any way, implicitly present.

Concerning *Adv-mente* of manner, i.e. those with scope on the verb, Gross (1986) highlights the different roles of the noun *façon* ('manner') and its modifying adjectives, and also the possibility of the verb being nominalised, as in (9) below:

(7) *Max se conduit ignoblement*
'Max behaves ignobly'

(8) *Max se conduit de façon ignoble*
'Max behaves in an ignoble manner'

(9) *Max a une conduite ignoble*
'Max has an ignoble behaviour'

In English, Quirk et al. (1985: 438) divide adverbs into three main morphological groups: simple adverbs, compound adverbs, and derivational adverbs. *Adv-mente* would fall into the third group, which comprises adverbs deriving mainly from adjectives. In regard to their syntactic function, Quirk et al. (1985: 439-440) highlight two main categories for adverbs: premodifiers and what is described as "clause element adverbials", which would be equivalent to the more autonomous adverbs with scope on the entire sentence. Basing on this general division, Quirk et al. (1985: 440) establish four grammatical functions for the second group: adjuncts, subjuncts, disjuncts, and conjuncts. Adjuncts and subjuncts have a closer relation to the clause, without losing the status of "clause element adverbials". Disjuncts and conjuncts, in turn, play a more peripheral role in the sentence, the former expressing an evaluation of the speaker about what is being uttered, and the latter expressing an assessment of a connection between two distinct units. Making use of Quirk et al.'s

examples, adjuncts (10), subjuncts (11), disjuncts (12) and conjuncts (13) would be, respectively:

(10) *Slowly* they walked back home.
(11) We haven't *yet* finished.
(12) *Frankly*, I'm tired.
(13) If they open all the windows, *then* I'm leaving.

According to Palma (2009: 24), the first Portuguese grammar to group adverbs in these two main categories was Cunha and Cintra (1984). However, it is noteworthy that although grammarians in the three languages referred to address the property adverbs have of modifying the entire sentence, it is possible to note that, in some grammars of Portuguese, this could perhaps be regarded as a somewhat secondary role of the adverb. In Cunha and Cintra (2000: 537), the first statement concerning adverbs is that they are "fundamentally" verb modifiers. Bechara (2003: 293) also mentions that "canonical" adverbial characteristics do not apply to adverbs that modify sentences.

As previously mentioned, for the purpose of this study, unambiguous adverbs of the sentence-modifying type are not going to be analysed since they do not seem to have the potential to form verb-adverb collocations. That is simply due to the fact that adverbs that modify the sentence have no straight connection with the verb itself.

Molinier and Levrier (2000) have more extensively investigated the specific category of *Adv-mente*. The description they make of these forms and the categories established for their classification are going to be used as a central reference to address the problems this study deals with. These categories are going to be regarded as a guiding parameter as to what should be considered and what should be discarded in a search for the collocational status of {V, *Adv-mente*} pairs and also as to the syntactic relation of relevant forms with other terms in the sentence.

In view of their broad classification of adverbs in the two groups previously mentioned, Molinier and Levrier (2000: 44) establish that *adverbes de phrase* 'sentence-modifying adverbs' can be identified by two main linguistic properties:

a. The possibility of occupying a peripheral position in negative sentences;

b. Impossibility of being "extracted" by making use of the structure "It's… that" (*C'est…que*, in French);

These principles could also be applied to Portuguese and English. For example:

(14) *Honestamente, este não é um bom filme*
'Honestly, this is not a good film'

Sentence (14) above has an *Adv-mente* in a peripheral position of a negative construction. It is not possible to extract the adverb from the sentence by means of saying:

(15) **É honestamente que este não é um bom filme*
*'It is honestly that this is not a good film'

Because of this, *honestamente* ('honestly') is classified as a sentence-modifying adverb.

Molinier and Levrier (2000) establish three categories for sentence-modifying *Adv-mente*: *les conjonctifs* ('conjuncts'), *les disjonctifs de style* ('disjuncts of style'), and *les disjonctifs d'attitude* ('disjuncts of attitude'). The first group would be characterised by their conjunctional property of linking two clauses; the second would express the enunciator's posture before the interlocutor; and the third would complement the second, being possible to be subdivided into adverbs of habit, adverbs of evaluation, adverbs of manner, and adverbs of attitude oriented to the subject.

Verb-modifying *Adv-mente* were classified by Molinier and Levrier (2000: 50-52) into six categories, which can be found bellow with examples provided by the authors, accompanied by a translation into English.

Adverbs of manner oriented to the subject:

(16) *Max regarde anxieusement l'horizon*
'Max looks anxiously into the horizon'

Adverbs of manner oriented to the verb:

(17) *Max regarde fixement l'horizon*
'Max looks fixedly at the horizon'

Quantifying adverbs of manner:

(18) *Max aime exagérément ce tableau*
'Max likes this painting exaggeratedly'

Adverbs of point of view:

(19) *Légalement, Max est responsable*
'Legally, Max is responsible'

Adverbs of time:

(20) *Max est venu ici récemment*
'Max came here recently'

Focus adverbs:

(21) *Max écrit principalement des poèmes*
'Max writes mainly poems'

This last category will also not be taken into account in a search for {V, *Adv-mente*} collocations for the simple reason that focus adverbs do not hold a straight connection with the verb, as previously pointed out (Baptista and Català, 2009). That also arguably applies to adverbs of time and of point of view, which, due to their looser connection to the verb, are not considered worthy of exploration in view of their collocational value.

The classification just shown includes single adverbs that may fall into more than one subcategory. It is the case of syntactically homonymous adverbs that, depending on the context in which they appear, can be either deemed adverbs modifiers of the sentence or adverbs that are an integrant part of the clause. In fact, Molinier and Levrier (2000) make a distinction between what they call *item lexical*, ('lexical item'), and the adverb itself. The lexical item is the form *per se*, which is able to play the role of what would be different adverbial forms, therefore belonging

to different subcategories. Among the cases cited by Molinier and Levrier (2000), is the lexical item *gracieusement*. It could either be an adverbial form of the category of adverbs of manner related to the subject, as in:

(22) *Marie danse gracieusement*
'Mary dances graciously'

Or an adverbial form falling into the category of manner adverbs oriented to the verb, as in:

(23) *Elle lui a envoyé gracieusement la brochure*
'She sent him the brochure free of charge'

In view of this brief overview of how *Adv-mente* are regarded in the literature, one can note that linguists tend to agree that the grammatical class of adverbs in general is rather blurred and unspecific. Albeit this non-specificity, it seems to be common ground in a comparison of two relevant grammars of Portuguese, one with another and also both in relation to other reference grammars of English and French, that adverbs should be syntactically divided into two main categories: those that modify sentences as a whole and those that constitute an integrant part of the clause. This division and the further stratification proposed by Molinier and Levrier (2000) are going to be of high importance for this study in setting syntactic filters for the extraction of {V, *Adv-mente*} collocation candidates from corpora.

2.3 Extraction of Collocations from Corpora

As the rich applicability of collocations came to the attention of linguists and language professionals in general, extensive efforts have been made within the field of NLP to automatically or semi-automatically extract such combinations from corpora. Statistical measures capable of gauging the degree of association between two or more terms have been proven extremely useful for this task. The extraction of {V, *Adv-mente*} pairs specifically is likely to require the corpus to be parsed since the mere adjacency of words is often not enough to make potential collocations of this

pattern surface. As seen in Pecina (2010: 139), some approaches can be based merely on the search of "surface bigrams", i.e. pairs of adjacent words. These approaches do not require the corpus to be processed and are sometimes justifiable by the assumption that "the majority of bigram collocations cannot be modified by insertion of another word". This, however, does not apply to the pattern investigated in this study, since {V, *Adv-mente*} pairs consist of a strictly syntactic relation that is not necessarily reflected by the adjacency of the terms. In this way, approaches for collocation extraction that do not include a corpus processing stage and deal with surface combinations only will not be described here.

Seretan (2011) has run an experiment that compares the sliding window method based on adjacency and a syntax-based approach to collocation extraction. The experiment was carried out with French data retrieved from the Hansard corpus (Roukos et al., 1995), composed of Canadian parliamentary proceedings. The top 500 collocation candidates yielded by each method were manually classified with respect to their grammatical correctness and collocational strength. Three French-speaking annotators trained for the task were recruited for the classification. In terms of collocational strength, results obtained with the experiment show that the syntax-based method achieved an uninterpolated average precision (UAP) (Manning and Schütze, 1999: 536) of 70.7, against 67.3 achieved with the sliding window method. The syntax-based approach also outperformed its counterpart in relation to the grammaticality of the candidate pairs obtained. A similar experiment is then replicated in four different languages with data taken from the *Europarl* parallel corpus[3] (Koehn, 2005), which is 3.1 times bigger than the corpus used in the first experiment. Results for the second experiment are consistent with those obtained in the first. The method based on parsing outperforms the sliding window method in the four languages dealt with – English, French, Spanish, and Italian. Even though the sliding window method has been largely adopted for collocation extraction in previous research, experiments of this kind show that parsing source corpora has indeed a great potential of improving final results.

Portela (2011), dealing with the identification of compound terms in Portuguese, has described a pipeline for the extraction of these terms from corpora that included both the processing of the corpus and the use of statistical measures. He established

[3] http://www.statmt.org/europarl/ [Accessed 15 May 2012]

that after the corpus has been processed, undesired syntactical structures should be filtered out, followed by the application of statistical measures and then by the use of algorithms aimed at automatically identifying the compounds. However, the task has proven extremely challenging, since a manual sub-sampling validation of the results has shown the methods employed had a precision of 25% in identifying noun-adjective compounds, and a precision of 10% identifying noun-preposition-noun compounds, results that fall short of being satisfactory.

Tools and algorithms aimed at identifying collocations in corpora have been frequently devised in NLP. Each tool tends to focus on one specific type of pattern, as differences in the syntactic relation between the terms may influence the strategy adopted for their retrieval. Evert (2005) lists some of the most important initiatives in that respect, highlighting automatic and semi-automatic pipelines designed for English, French, German, and Estonian. A tool that appears to be a target of constant attention amongst researchers addressing this topic is the XTRACT tool (Smadja 1993), which combines the use of association measures, heuristics, syntactic patterns and filters, and is, according to Evert (2005: 26), "the most well-documented collocation extraction system so far".

Manning and Schütze (1999), after briefing the reader with some important concepts of statistics, describe the most widely used association measures for the purpose of extracting collocations from corpora. Some of these measures have already been applied in this study in preliminary experiments. What follows is a succinct explanation of those deemed more relevant amongst them.

One of the most basic and widely known is the Student's *t* test (Fisher, 1925), explained by Manning and Schütze (1999: 163) in view of the collocation extraction task. It is a measure that shows how probable or improbable a combination is of occurring. The *t* test should be employed based on a threshold that establishes the limit between statistically relevant and non-relevant cases. The *t* value that corresponds to a confidence level of $\alpha = 0{,}005$ is 2.576, which is a pre-established fixed value in statistics for the *t* test and can be found in Manning and Schütze (1999: 609). In this way, whenever the *t* value of a combination is lower than 2.576, considering $\alpha = 0{,}005$, this combination does not receive the status of a collocation according to this measure. The *t* test has received a considerable dose of criticism from specialists in the area, because it is claimed to wrongly assume a normal

distribution of probabilities. This is seen in Manning and Schütze (1999: 158), who point this out based on Church and Mercer (1993: 20).

As a potential alternative to the *t* test, there is the Pearson's chi-square test (χ^2) (Pearson, 1900). The chi-square test, χ^2, is based on a comparison between the observed frequency of the combination with the expected frequency with which its terms appear separately in the corpus. It is generally applied to two-by-two tables, considering the frequency of the bigram, the frequency of each word of the bigram separately and the frequency of bigrams that do not contain any of the words of the pair whose collocation status is being assessed. Similarly to the *t* test, the χ^2 has a reference value that functions as a relevance threshold indicating which cases could be considered a collocation and which could not. As seen in the table of critical values in Manning and Schütze (1999: 610), a confidence level of $\alpha = 0{,}05$ would be acceptable for the χ^2, resulting in a value of 3.841. Hence, all combinations analysed with the χ^2 test that stand below this limit are not to be considered relevant in terms of their collocation status.

Another statistical measure described by Manning and Schütze (1999) is *Mutual Information* (MI) (Fano, 1961). This measure takes into account the type of relation that exists between the terms of a combination. Roughly speaking, it considers information about one word and uses it to assess the influence the occurrence of this first word has on the occurrence of the second. This measure is different from the other two previously mentioned in that it does not have a pre-established reference value that discards irrelevant cases. When MI is applied, one has to freely evaluate and interpret results based on their ranking.

While the use of the referred measures can be deemed extremely recurrent in collocation extraction tasks in general, there is a vast body of literature on other statistical measures that can potentially point to conclusions concerning the collocation status of word combinations, each one with its own particularities and best applicability environments. Pecina (2010) has run a series of tests to evaluate the performance of 82 different association measures, contrasting results with a reference set of manually annotated collocations extracted from a corpus. The author runs tests in three different contexts: collocations extracted as syntactic dependencies from an annotated corpus of 1.5 million words, collocations extracted as surface bigrams from the same corpus, and collocations extracted from a considerably larger corpus of 242

million words, considering the instances of the same surface bigrams from the previous corpus. Results have shown that MI, χ^2, and, surprisingly, the *t* test, were amongst the measures that presented the best results in the experiment with the first corpus of 1.5 million words. However, the best method observed for the classification of extracted dependencies in the experiment with this corpus was *Cosine context similarity in Boolean vector space*, whose formula is provided in Pecina (2010: 156). Results for the large corpus of 242 million words suggest that the best two methods to be applied to large data sets are *Unigram subtuples* and MI (Pecina, 2010).

Similarly, Pearce (2002) calls attention to a number of different collocation extraction techniques, running a series of tests aimed at comparing and evaluating them. The author discusses the achievements of researchers in the NLP field and their experiments in extracting collocations from corpora, including what, he affirms, is the earliest attempt in this respect, the technique devised by Berry-Rogghe (1973). He also described the more recent experiments of Church and Hanks (1989), Kita et al. (1994), Shimohata et al. (1997), Blaheta and Johnson (2001), and Pearce (2001). The technique developed by Pearce (2001) could be considered particularly interesting because it relies on synonymic substitution as an indication of collocational potential. As in an example provided by the author, the collocation *emotional baggage* loses its collocation status if the word *baggage* is substituted by its synonym *luggage*, which denotes that *emotional baggage* is in fact a collocation. This principle is also adopted in this study for the classification of {V, *Adv-mente*} collocations. Pearce (2002) concludes that the lack of consensus concerning the linguistic notion of collocations poses a problem to any comparison of extraction techniques, since each technique may presuppose a different notion, resulting in biased results.

Pecina and Schlesinger (2006) addresses this problem by means of dealing with statistical scores isolated, instead of complete techniques as the ones exposed by Pearce (2002), not only comparing different measures but also attempting to combine them. The experiment showed that the combination of different measures could present a considerable potential of enhancing the task of extracting collocations from corpora. The approach adopted consists in combining all 82 association scores analysed, yielding one result that will indicate if the bigram is a collocation or not. It was observed, however, that 82 was perhaps too large a number, making the task considerably more complex. An algorithm capable of optimising the use of

association scores was proposed to overcome this obstacle. The assumed principle is that some scores are too alike and therefore their inclusion in the combination would be redundant. This gives rise to what the author calls *reduced models*, which would be intelligent combination models capable of selecting the most relevant scores for the combination. All in all, it can be apprehended from Pecina (id: ibid.) that combining different measures may be an extremely promising strategy in the task of collocation extraction from corpora. The underlying idea is that electing one optimum measure would be a limited approach, highly dependant on the notion of collocation presupposed in the search. Profiting from particular advantages presented by different measures may be a more effective option instead. In that way, an experiment that combines measure results to train an automatic collocation classifier in described in Chapter 5. We have also experimented to combine different classifiers themselves based on these measures, which has shown to be a fairly promising strategy.

2.4 Automatic Translation of Collocations

Automatically translating collocations is commonly seen as a problematic task in NLP due to the fact that the translation cannot be performed on a word-by-word basis. Even though, as previously seen, the concept of collocation is not consensual; it is frequently assumed that the meaning of a collocation does not necessarily have an evident relation with the meaning of its constituents. This poses a problem to MT, since an equivalent construction for the source language has to be found in the target language, and the two combinations can have words that are different parts-of-speech and whose literal meaning may differ. This problem is usually referred to as *lexical transfer*.

It can be argued, however, that the pattern investigated in this study poses a subtler level of difficulty to MT. The core issue of translating {V, *Adv-mente*} pairs would reside in the fact that, from the various options of adverbs that can be employed with a given verb, there could be one that proves to be the best in terms of fluency and adequacy.

This idea is very closely connected with the MT concept of *fluent output*, seen in Koehn (2010: 94). It consists of the premise that the context surrounding any word to be translated should be taken into account by the MT engine. As in an example

provided by Koehn (id: ibid.), both *small* and *little* would be correct translations for the word *klein*, in German. Hence, if the next word is, for instance, *step* both *small step* and *little step* should be correct options. However, there probably is one alternative that accounts for a higher level of fluency of the output. A search performed by the author in the Google index has shown that *small step* has 2,070,000 occurrences, as opposed to 257,000 occurrences of *little step*. The higher frequency of *small step* would serve as an indicator that this pair is, in fact, the best option from the two possible translations.

What Koehn (id: ibid.) suggests as a way to ensure fluent output in MT is the use of n-gram language models. The use of n-gram models would make it possible to compute the probability of longer strings, task that Google is not able to perform successfully, as the author indicates.

Smadja et al. (1996) have developed an MT system named *Champollion*. Having a parallel bilingual corpus as database, the system is reported to be able to automatically translate collocations from a source language into a target language. Briefly put, it works by means of progressively electing in the target language words that correlate to the ones in the source language. After a group of highly correlated words in the target language is selected, these are combined among themselves, first forming pairs, and then triples, with a third highly correlated word being added to the pair, and so forth. In the final stage, it analyses the corpus and provides the adequate word ordering. It also labels the produced combination as *flexible* or *rigid*. Flexible combinations would be those that allow for the insertion of other words, whereas rigid combinations would stand for those that can only appear consecutively, without other terms in-between. The measure they considered the best to establish the correlation between words in the source and target language is Dice coefficient (Dice, 1945; Sörensen, 1948). This measure was chosen because it ignores cases in which correlated words do not appear together in any of the aligned sentences, which perfectly meets their criteria.

The program was evaluated by means of compiling lists of collocations with the XTRACT tool (Smadja, 1993) and then translating these collocations into French through *Champollion*. Results were submitted to the judgment of fluent speakers of English and French, and a range of accuracy that goes from 65% to 78% was achieved. The authors consider that this result can be further improved with the use of

a larger corpus as database. Perhaps a relevant differential of this model is the fact it is able to perform the translation of combinations, regardless of the already mentioned problems of having a different number of words or different parts-of-speech in what would be equivalent collocations in the two languages.

The approach adopted by Smadja et al. (1996) seems to be a practical proof that MT involves problems that are related to linguistic particularities of both languages of the translation pair. This notion was already in vogue ten years prior to their work, as it can be seen in Tsujii (1986), who affirms that problems related to MT cannot be addressed under a merely monolingual perspective, and that "certain 'understanding processes' are target language dependent" (Tsujii, 1986: 662).

However, it appears that the use of parallel bilingual corpora was subjected to a considerable dose of criticism when it started to be suggested as a way to address MT problems. Church and Gale (1991) make mention of this criticism, defending that the use of parallel corpora in MT present enough advantages to make it an avenue worth taking. They point out that one of the reasons behind the criticism received by the use of parallel corpora in the past lied in the many difficulties that once made this approach unfeasible, contributing to the popularity of monolingual corpora instead. Nevertheless, Church and Gale (id: ibid.) claim, back at the beginning of the 1990's, that bilingual corpora were already a reality due to the considerable improvement the task of text alignment had undergone. In fact, they present and discuss a number of different tools for the specific purpose of text alignment, envisaging its application in MT.

As to the relation between collocations and parallel corpora, recent studies report the identification of collocations as a way to improve bilingual multi-word alignment and the phrase-based approach to Statistical Machine Translation (SMT). This approach is described in Koehn et al. (2003), and was further implemented in the Moses system (Koehn et al. 2007). In broad terms, the system first segments sentences into word chunks – phrases – and then performs the translation of these chunks based on a phrase translation table. Liu et al. (2010) have reported significant results in their attempt to use collocation information as a way to improve SMT. A similar experiment was carried out by Costa-jussà et al. (2010), who attempted to use the collocation segmentation method developed by Daudaravicius (2010) as a way to improve the phrase translation table for SMT. In this last experiment, a baseline

phrase translation table is combined with a phrase table extracted based on collocation equivalents between English and Spanish, resulting in a more accurate translation of collocation patterns in these two languages.

As these studies suggest, information of collocation patterns presents a relevant potential of improving MT, perhaps more so in regard to SMT. In view of the reported results, the work here envisaged aims at following a similar trend, relying on the belief that the knowledge of collocational {V, *Adv-mente*} pairs in Portuguese and their relation to English can lead to further improvement guidelines for the automatic translation of this pattern between these two languages. In that way, we have devised an evaluation method to assess the automatic translation of this pattern in the direction Portuguese-English, taking into account three commercial MT engines available to the general public free of charge. This experiment is described in Chapter 6.

Chapter 3. Corpus Processing and Dependency Extraction

The extraction of verb-adverb collocation candidate pairs from the corpus was addressed in the context of the development of the STRING system (Mamede, 2011), a Portuguese NLP chain developed at L2F-INESC ID Lisboa. The system is composed of several modules, including a tokeniser, a morphological analyser LEXMAN (Diniz, 2010, Diniz and Mamede, 2011) a statistical POS tagger MARV (Ribeiro, 2003), and a syntactical parser XIP (Xerox Incremental Parser) (Aït Mokhta et al., 2002). XIP is a cascade, finite-state, rule-based parser that analyses sentences into chunks, extracts syntactic dependencies between chunks and is also used for named entity recognition (Hagège et al., 2010, Oliveira, 2010) and (partially) for co-reference resolution (Nobre, 2011) and relation extraction (Santos, 2010).

3.1 Coordination of *Adv-mente*

As already mentioned in Chapter 1, this project also aims at improving the STRING text processing chain for a more appropriate analysis of *Adv-mente*. This task was addressed in view of the problem that the coordination of *Adv-mente* poses to the correct computational analysis of these adverbs in Portuguese.

When coordinated, Portuguese *Adv-mente* lose the suffix and appear in the feminine-singular (*fs*) form of the base adjective:

(24) *O Pedro leu isso lenta e atentamente*
'Peter read this slow_*fs* and attentively'

=

(24a) *O Pedro leu isso lenta[mente] e [O Pedro leu isso] atentamente*
'Peter read this slow(ly)_*fs* and [Peter read that] attentively'

If there is a feminine-singular noun before the reduced adverb, it is very likely that the adverb would be considered as an adjective instead, and treated as a modifier of that noun, e.g. *a revista lenta*, ('the magazine slow') in the example below:

(25) *Pedro leu a revista lenta e atentamente*
'Peter read the magazine_*fs* slow_*fs* and attentively'

Finally, as coordination can be iterated, longer chains of reduced adverb forms can be found:

(26) *O Pedro leu isso lenta, pausada e atentamente*
'Peter read that slow_*fs*, pausing_*fs* and attentively'

Because the reduced form of the adverb and the feminine-singular form of its base adjective are homographs, the POS of the word has to be disambiguated. However, without semantic (distributional) information on noun-adjective combinations, adverb combinations, or even verb-adverb pairs, any solution to this non-trivial problem is just an approximation.

On the other hand, it would be useless (and eventually hampering to a system) to consider that all feminine-singular adjectives could be adverbs in every context. So this particular type of strictly local ambiguity should be solved prior to general parsing rules or statistical models be applied to the text.

The performance of statistical POS taggers depends on the granularity of the tag set used by the learning algorithms, and since many systems only use a coarse tag set, i.e., considering only the major POS category, but discarding the inflection, it is very difficult to train models sensitive to this particular phenomenon.

Finally, the coordination of adverbs, while a relatively common phenomenon in Portuguese, occurs very infrequently in texts. For the system here used, the statistical POS tagger (Ribeiro, 2003), based on the Viterbi algorithm, uses a manually annotated corpus of 250K words. In this corpus only 10 instances occur of the pattern corresponding to the coordination of *Adv-mente* but only 4 are in fact coordinated *Adv-mente*. The sparsity of the phenomenon makes it an interesting challenge to NLP systems, difficult to tackle by a purely machine-learning approach.

An alternative solution has been proposed in the context of this study. In the following Sections, the modules that compose the STRING system are explained in view of this solution. Results obtained are presented in Section 3.6.

3.2 The Lexicon

In view of the verb-adverb dependency extraction task, the existing lexicon of the system has been systematically completed by adding all *Adv-mente* entries found in an orthographic vocabulary (Casteleiro, 2009). These correspond to 3,614 entries. Then, all valid *-mente* ending forms found in the European Portuguese corpus were manually perused and the adverbs selected. Duplicates from the first list were removed, thus yielding 3,636 new entries.

For each entry, the feminine-singular form of the base adjective was automatically generated, which consist of part of the strategy to disambiguate coordinated *Adv-mente* reduced of the suffix, described in the following Section. The list was then manually revised for errors and for the insertion of orthographic variants, resulting from the new, unified Portuguese orthography.

The final list consists of 7,250 *Adv-mente*. For example, the entry for *abstratamente* ('abstractly') is associated with the orthographic variant *abstractamente* ('abstractly'), and to the reduced forms *abstrata* and *abstracta* (abstract_*fs*). This reduced form is then given the feature 'r' (for 'reduced').

When analysing a sentence where *abstracta* appears, at this morphologic stage, the system produces the following tags (format adapted for clarity):

```
abstracta: abstratamente Adv_r; abstrata Adj_fs
```

It has been previously noted by Afonso (2002) that compound adverbs (or collocational combinations), such as *única e exclusivamente* ('uniquely and exclusively'), and *única e simplesmente* ('uniquely and simply') occurred quite often in the corpus. Besides these forms, the lexicon was completed with other similar ones, such as *pura e simplesmente* ('purely and simply'), *dire(c)ta ou indire(c)tamente* ('directly or indirectly'), *explícita ou implicitamente* ('implicitly or explicitly'), and *total ou parcialmente* ('totally or partially'). These combinations occur 3,074 times in the *CETEMPúblico* corpus.

3.3 Rule-Based Disambiguation

The next step in the system processing chain is a rule-based disambiguation module (Diniz, 2010, Diniz and Mamede, 2011). The linguistically motivated disambiguation rules produced consist of regular expressions that take the general form:

<left context>|<pattern>|<right context> := <result>

where <pattern> corresponds to the ambiguous target word and the different categories it may be associated with; <result> consists in selecting (+) or discarding (-) a given category; the left and right contexts are facultative.

Considering the disambiguation of coordinated reduced adverbs, for example, the general rule below selects the reduced adverb form when it appears coordinated with an *Adv-mente*:

```
0> [CAT='adv',SYN='red'][CAT='adj'] |
[surface='e'];[surface='ou'];[surface='mas'],
[surfaceRegex='.+mente',CAT='adv'] |
:= [CAT='adv']+
```

This rule reads as follows: the left context is empty; the <pattern> consists of the ambiguous form adverb/adjective; the adverbial form must present the feature SYN with the value 'red'; then follows the right context, where the coordinative conjunctions and the *Adv-mente* are explicit; for the conjunctions, the surface form is sufficient; to define the adverb, a regular expression is used along with its POS.

Most rules have to be duplicated in order to deal with the feminine-singular form of past participles. This is the purpose of the rule below:

```
 0>  [CAT=`adv',SYN=`red'][MOD=`par',GEN=`f',NUM=`s'] |
[surface=`e'];[surface=`ou'];[surface=`mas'],
[surfaceRegex='.+mente',CAT=`adv'] |
:= [CAT=`adv']+.
```

Rule-order application is fixed, so more specific rules are stated before more general ones. For example, the pattern of coordinated adjectives, each modified by an

adverb is more constraint than the previous patterns and it is thus stated before the general rules above:

```
0>  [CAT='adv'] \textbar \\
[CAT='adv',SYN='red'][CAT='adj',GEN='f',NUM='s'] |
[CAT='con',SCT='coo'],[surfaceRegex='.+mente',CAT='adv'],
[CAT='adj',GEN='f',NUM='s'][MOD='par',GEN='f',NUM='s'] |
:= [CAT='adv']-.
```

Some rules require lists of words to be spelled out, such as the next one, where a negation adverb in front of an ambiguous adjective is the context that allows to discard the reduced adverbial form; the negation adverb is provided by a list of words (at later stages, namely in the parser, this information is encoded by way of feature-value pairs):

```
0> | [surface='não'];[surface='nem'];[surface='nunca';
[surface='jamais'];[surface='nada'] |
[CAT='adv',SYN='red'][CAT='adj'] |
[surface='e'];[surface='ou'];[surface='mas'],
[surfaceRegex='.+mente',CAT='adv'] |
:= [CAT='adv']-.
```

Finally, at the last stage of the process and for the remaining ambiguous forms, the tag corresponding to the reduced adverb form is discarded by a general “cleaning" rule:

```
0> [CAT='adv',SYN='red'][SYN=~'red']
:= [SYN='red']-.
```

3.4 Chunking

In the chunking stage, the XIP parser analyses the sentence by splitting it into elementary constituents (or chunks).
Ordinarily, a stand-alone adverb construes an adverbial phrase (ADVP). Chunks are formed according to chunking rules, such as the following, allowing up to three consecutive adverbs to form an ADVP:

```
ADVP @= (adv), (adv), adv.
```

At this stage, the system can make use of a rich set of lexicons, featuring syntactic and semantic information, as well as the information derived from the morphological analyser. In the coordination of *Adv-mente*, an ADVP is construed. For example, for the sentence *O Pedro leu isso lenta e atentamente* ('Peter read this slowly and attentively') the following chunking is produced:

```
0> TOP{NP{O Pedro} VF{leu} NP{isso}
1> ADVP{lenta e atentamente} .}
```

The ADVP results from the application of the following rule:

```
18> ADVP @= ~ | ?[noun,fem,sg] |
(adv[advquant];adv[advcomp];adv[neg])*,
adv[reducedmorph],
conj[lemma:e];conj[lemma:ou];conj[lemma:mas],
(adv[advquant];adv[advcomp];adv[neg])*,
adv[surface:"\%c+mente"].
```

The chunking rule reads: an ADVP chunk is built with two coordinated adverbs, the first is a reduced form, indicated by the feature [reducemorph], and the second an Adv-mente; only conjunctions the *e* ('and'), *ou* ('or'), and *mas* ('but') are allowed; both adverbs can be further modified by a quantifying adverb, a comparative adverb or a negation adverb; these adverbs have been given the features

[advquant], [advcomp] and [neg], respectively, in the lexicon; this chunking is not made if there is a feminine-singular noun in the left context of the pattern. A similar rule is used for coordination of three (or more) *Adv-mente*.

3.5 Dependency Extraction

Finally, the parser extracts the syntactic relations between the chunks. Dependency extraction rules have the general format:

```
<left context> | <pattern> |  <right context>
if <conditions> <dependencies>
```

Particularly relevant for this study is the *modifier* (MOD) dependency, which is now very briefly presented.

The modifier dependency holds between two chunks. For *Adv-mente*, most of them modify a verb or an adjective. One of the basic rules for extracting the adverbial, right modifier of a verb is given below:

```
|#1[verb];sc#1, ?[verb: ~{ } ,scfeat: ~{ } ],
(AP;PP), (PUNCT[comma]), ADVP#2 |
if ( HEAD(#3,#1) and HEAD(#4,#2) and ~{ } MOD(?,#4)
and ~{ } QUANTD(#3,#4))
MOD[post=+](#3,#4)
```

Briefly, this rule reads: For a verb (or a sub-clause SC) #1 and an adverbial phrase #2, eventually admitting an adjectival or prepositional phrase, or a comma, in-between; if no modifier MOD has been extracted for the head of #2, nor a quantifier QUANTD dependency has been extracted between the heads of #1 and #2; then build the MOD dependency between the heads of the verb and the adverb phrases.

The result of the dependency extraction process for sentence *O Pedro leu isso lenta e atentamente* ('Peter read this slowly and attentively') is the following:

```
MAIN(leu)                 MOD_POST(leu,atentamente)
DETD(Pedro,O)             MOD_C-MENTE POST(leu,lenta)
COORD_C-MENTE(e,lenta)    SUBJ_PRE(leu,Pedro)
COORD(e,atentamente)      CDIR_POST(leu,isso)
VDOMAIN(leu,leu)          NE_PEOPLE_INDIVIDUAL(Pedro)
```

Briefly, the dependencies above include the subject (SUBJ) and direct object (CDIR); the determinant (DETD) and the named entity (NE); the main (MAIN) element of the sentence; the verb domain (VDOMAIN), for dealing with auxiliary verbal chains (not relevant in this example); and, finally, the two coordination dependencies involving the adverbs, and the corresponding modifier dependencies. Features _PRE and _POST indicate if the dependent is to the left or to the right of the dependency head.

3.6 Results for *Adv-mente* Coordination

3.6.1 The Evaluation Corpus

For the evaluation, a corpus, with 1,132 sentences, was retrieved from the *CETEMPúblico*. It consists of sentences presenting an adjective or past participle, one of the three main coordinating conjunctions – *e* ('and'), *ou* ('or'), or *mas* ('but'), and an *Adv-mente*. Sentences were obtained from the concordances retrieved using the AC/DC search system of *Linguateca* webpage[4].

The corpus was then parsed by the system and the dependencies were manually corrected, each sentence being independently checked at least twice, by two linguists. The chunking was also corrected, when appropriate. For this paper, only the COORD and MOD dependencies involving *Adv-mente* or their reduced forms were kept from the system's output.

Table 3.1 shows the breakdown of each dependency in the corpus. The difference between COORD and COORD_C-MENTE is due to the cases of multiple

[4] http://www.linguateca.pt/ [Accessed 15 May 2012]

coordination – i.e., more than two adverbs coordinated together. The large difference between MOD and MOD_C-MENTE consist of *Adv-mente* that, although occurring next to a conjunction and after a reduced form, are not coordinated with it, and modify some other constituent in the sentence.

Dependency	#
COORD	438
COORD_C-MENTE	462
MOD	1403
MOD_C-MENTE	462

Table 3.1 Dependencies in reference corpus

3.6.2 Results for the Disambiguation Rules

This step consists in assessing the impact of the disambiguation rules in selecting or discarding the POS tags corresponding to the adjective or the reduced adverbial form. Table 3.2 shows the results of the rule-based disambiguation module. From the 462 adverb reduced forms, the system fails to spot 21, while it incorrectly accords this tag to 316, therefore yielding a relatively low precision but high recall, contributing to the interesting F-measure result. This means that in spite of the conservative approach in devising the disambiguation rules and the final, "cleaning" rule that eliminates all remaining reduced forms not previously captured, the system still fails to recognize the cases where there is no coordination of adverbs.

Precision	Recall	*F*-Measure
0.583	0.955	0.724

Table 3.2 Results for disambiguation rules

3.6.3 Results for the Dependency Extraction

The next figures are a combined result of the chunking and of the dependency extraction modules. The purpose of parsing a text is to retrieve the syntactic-semantic

relations between constituents, which (partially) express the text meanings. Table 3.3 shows the results for the dependency extraction module. In order to obtain a better perception of the system performance, a set of experiments was carried out. The first line presents the overall performance of the system. In the next lines, each dependency is evaluated separately. Finally, the two coordination and modifier dependencies are evaluated in pairs.

Experiment	Precision	Recall	*F*-Measure
All dependencies	0.754	*0.875*	0.810
MOD	*0.921*	*0.852*	*0.886*
MOD_C-MENTE	0.608	0.719	0.659
COORD	0.642	0.777	0.703
COORD_C-MENTE	0.646	0.805	0.717
2MOD	0.822	0.849	0.834
2COORD	0.644	0.858	0.736

Table 3.3 Results for dependency extraction

The overall performance of the system in the dependency extraction is promising. In general, the system is able to extract most of the modifier dependencies (92%), and only 39% of reduced adverbial forms are not adequately related to the element they modify. The system shows suboptimal performance in the extraction of coordination dependencies. There is a clear relation between the low precision in the MOD C-MENTE and the low precision on COORD dependencies. When the system fails the coordination, it also (partially) fails to extract the modifiers. The reason for this is to be found in the previous module of disambiguation rules, which often and inadequately selects the reduced adverb form instead of recognizing the coordination of adjectives

3.7 Extracting {V, *Adv-mente*} Pairs from the Corpus

Once the corpus was processed and syntactically analysed, all the syntactic modifying dependencies between verbs and *Adv-mente* were extracted from the corpus through computing techniques as the ones used in Portela (2011). In total, 65.535 pairs of {V, *Adv-mente*} were extracted, along with details of the frequency of

the pairs and of their components isolated in the corpus, resulting in a text file that had the following format:

Adv-mente	Verb	V-Adv_Frq	Adv_Frq	Adv_Class	V_Frq
abertamente	*advogar*	6	1659	MV	1997
abertamente	*combater*	6	1659	MV	8880
abertamente	*atacar*	7	1659	MV	10995
abertamente	*confrontar*	5	1659	MV	7577

Table 3.4 Examples of {V, *Adv-mente*} pairs extracted from corpus

Table 3.4 shows an example of the resulting pairs of modifying dependencies between verbs and *Adv-mente* extracted from the corpus. The first column has the *Adv-mente*, whilst the third has verbs. The fifth, sixth and ninth columns, respectively, have the frequency of the pair together in the corpus, the frequency of the *Adv-mente*, and the frequency of the verb. Information regarding the classification of the adverbs is also present, which will prove extremely important for the filtering process that takes process prior to the classification of collocations, which is going to be explained in the next Chapter.

Chapter 4. Classification of Collocation Candidates

4.1 Filtering the Extraction Output

As mentioned in the previous Chapter, the total number of {V, *Adv-mente*} modifying syntactic dependencies extracted from the corpus was of 65,535, whose frequency in the corpus exceeds 290,000 occurrences altogether. In order to narrow down the search for the collocation pattern investigated, a number of filtering strategies have been applied to the results so as to eliminate from the outset cases that did not present any potential for being classified as collocations.

With regard to *Adv-mente*, we have augmented the adverbial classification carried out by Fernandes (2011) for *Adv-mente* in Portuguese, which initially covered approximately 520 adverbs. This number has now been increased to nearly 1,000, including adverbs whose frequency is equal to or higher than 3 ($f \geq 3$) in the *NILC São Carlos* corpus of Brazilian Portuguese[5] (Pinheiro and Aluísio, 2003). This corpus is mostly composed of news texts and has approximately 32,3M words. For the most part, the classification can also be used for the processing of European Portuguese and has been incorporated in the lexicon of the STRING text processing chain. The classification of adverbs carried out as part of this study can be found in Appendix D.

Having knowledge of the class or classes a given *Adv-mente* belongs to played an important role in filtering out adverb categories that do not hold a straight connection with the verb, which consequently impedes the formation of a verb-adverb collocation. That would be the case of adverbs that play the single role of modifying a sentence, i.e. sentence-modifying *Adv-mente*, namely conjunctive adverbs (PC), disjunctive adverbs of style (PS), and disjunctive adverbs of attitude (PA) (Molinier and Levrier, 2000). Focus adverbs (MF), albeit being commonly integrated in the clause, were also filtered out due to their low potential of receiving collocation status, since their sole purpose in an utterance is to emphasise a sentence constituent.

Certain verbs with little semantic content were also filtered out at this stage. So-called *light verbs* (Jespersen, 1965) or *support verbs* (Gross, 1981), such as *fazer* ('to do'), *dar* ('to give'), *ter* ('to have'), and *haver* ('to exist'/'there is/are') were among

[5] http://www.linguateca.pt/acesso/NILCsaocarlos.html [Accessed 15 May 2012]

the verbs to be discarded, as well as copula verbs such as *ser* ('to be'), *estar* ('to be'), *permanecer* ('to remain'), *ficar* ('to stay'), and *parecer* ('to seem'). At the moment of the syntactical parsing, however, if these verbs were part of a verb phrase containing a participle form, the modifying relation would be established between the participle, as head of the phrase, and the adverb. Participles used as adjectives were also ignored.

Due to the high frequency with which the verbs just mentioned are used in the language, they were present in the vast majority of verb-adverb combinations extracted from the corpus. After the filtering process, the remaining number of verb-adverb bigrams was of 5,973, which was then considered the set of collocation candidates on which manual annotations would be made.

A frequency threshold of five ($f \geq 5$) was established for the consideration of pairs as collocation candidates. This is a threshold that can be deemed considerably low, given the fact the total number of words in the corpus is 192M. The reason for such a low frequency threshold lies in the fact that it would potentially enable the coverage of a more significant amount of collocation pairs, whose low frequency is associated with the specific nature of the linguistic phenomenon investigated.

4.2 Establishing Empiric Classification Criteria

A linguist, native speaker of Portuguese, manually classified the 5,973 collocation candidates as to their collocational status. The classification at this stage was binary, i.e. a given candidate pair could be given either the tag of 'collocation', or the tag of 'non-collocation'.

As previously mentioned, even though frequency of distribution is taken into account in this study as an influencing factor for the classification of collocations – since a frequency threshold was applied to the output of the extraction – a linguistic definition of collocations was adopted as the guiding parameter for the classification of candidates. The broad linguistic notion used is based on Mel'čuk's formalisations (2003, 2010). However, since the pattern {V, *Adv-mente*}, to the best of our knowledge, has not been treated to date in the literature in view of its collocational potential, empiric and more precise linguistic criteria had to be devised for the classification of candidate pairs.

In broad terms, the collocational phenomenon that this study addresses is one that holds a straight connection with the fluency of the combination in a natural language-production context. In other words, combinations that sound more fluent than others, but that not necessarily represent the only linguistic option available when producing an utterance. Hence, from this point of view, the fact that a given verb-adverb pair is considered a collocation relies less on the ungrammaticality or unacceptability of other potentially equivalent combinations, and more on the fluency this specific combination accords to the speech. The ungrammaticality and/or unacceptability of equivalent constructions might be the case, however, which in fact makes it easier to identify cases of collocations.

The criteria devised to establish the collocational status of candidate pairs were based on what Greenbaum (1970: 10) calls an 'integrated' study of collocations, i.e. a study that considers both syntax and semantics, taking into account the relationship of a given term with its possible collocates as well as the meaning of the words involved in the combination.

In the analysis of the {V, *Adv-mente*} pairs extracted from the corpus, it has been observed that adverbs that – either themselves or in the form or their base adjective – represent more than one lexical item tend to present a higher potential to form collocations. That would be the case of the following examples, in which the adverb's base adjective has more than one possible meaning:

(27) *A professora **criticou duramente** o aluno*
'The teacher **criticised** the student **hard(ly)**[6]'

(28) *Ele **mostrou claramente** a solução para o problema*
'He **showed** the solution to the problem **clearly**'

(29) *O politico **defendeu abertamente** sua opinião*
'The politician **openly defended** his opinion'

In (27), the adverb *duramente* ('hard-ly') derives from the adjective *duro* ('hard'), which can either have a more canonical, universal meaning of *something that is hard to the touch*, or the alternative meaning of something that is *difficult* or *poses physical or mental effort*. The same applies for the adverbs in the other two examples:

[6] The adverb would be substituted by *harshly* in an equivalent construction in English.

the adjective *claro* ('clear') has both the meanings of *illuminated by some source of light* and *not difficult to understand*; the adjective *aberto* ('open'), in turn, can either mean *not obstructed by a physical barrier* or *conspicuous*, *not hidden*. In the examples above, the adverb assumes exactly the non-canonical semantic construction of its base adjective. This is a pattern that has been observed for cases that were deemed to be collocations, and in that way it could be considered an overall guiding parameter for their identification.

In the same line of reasoning, *Adv-mente* of time (MT) (Molinier and Levier, 2000) would arguably also be indicative of the higher collocational potential of non-canonical meanings. Even though *Adv-mente* of time are not completely destitute of the potential for forming collocations, it has been noticed that their vast majority do not seem to form interesting verb-adverb pairs from the collocational point of view. It could be argued that this aspect is related to the fact that these adverbs have a less varied semantic charge if compared with supposedly richer categories in that respect, such as manner *Adv-mente* (MV) and *Adv-mente* oriented to the subject (MS). *Imediatamente* ('immediately'), for example, albeit occurring 525 times among the candidate pairs, has been found to form collocations in four instances only, namely with the verbs *reagir* ('to react'), *parar* ('to stop'), *suspender* ('to suspend'), and *iniciar* ('to start'). These verbs are themselves connected somehow to the notion of time, and together with the adverb *imediatamente*, they express the notion of *abruptly* starting or finishing something, while this adverb, in its vast majority of occurrences, did not seem to assume this meaning. This seems to corroborate the idea that the adverb *imediatamente* in fact represents a different, but homonymic, lexical item when combined with the verbs just mentioned. It is on these terms that we have observed that semantically rich words had a higher potential to form {V, *Adv-mente*} collocations.

This has been one of the principles behind the list of criteria devised for the classification of the collocational pattern addressed in this study. An explanation of these criteria, along with illustrating examples, is presented bellow.

1. The adverb has a hyperbolic meaning in the combination, e.g.:

(30) *Ele **esperou eternamente** pelo telefonema*
'He waited eternally for the phone call'

2. The adverb holds a non-literal meaning in the combination, e.g.:

(31) *O time **venceu confortavelmente** a partida*
'The team **won** the match **comfortably**'

≠ *O time estava confortável*
'The team was comfortable'

(32) *Ele **deitou-se confortavelmente** na cama*
'He **lay comfortably** in bed'

= *Ele estava confortável*
'He was comfortable'

While in (32) the adverb *confortavelmente* ('comfortably') holds its literal meaning, connected to the idea of physical comfort, in (31) it assumes a figurative meaning adopted to express the idea that the match was won *effortlessly* or by a large scoring difference. The non-literal meaning in this case attributes a unique character to this combination that accounts for its classification as a collocation in Portuguese. In (32), the adverb in the combination, a manner adverb with scope on the action itself and on the subject of the verb, can be paraphrased by its equivalent base adjective operating on the same subject. In (31) this transformation is not possible, which would denote the non-literal construction of the adverb in the context of this sentence.

In another example, the adverb modifies the verb by according a quantifying/intensive value to it, such as *perdidamente* ('lost ly'), below:

(33) *Ele **apaixonou-se perdidamente** por ela*
'He **fell lost(ly)** in love for her'

≠ Ele estava perdido
'He was lost'

In (33), the adjective *perdido* (lost), which corresponds to the adverb *perdidamente* ('lost ly')[7] in Portuguese, requires a specific context in order to be able to modify the subject of the sentence and maintain the same meaning of the adverb. Even though the construction *Ele está perdido em seu amor por ela* ('He is lost in his love for her') would arguably be possible, the more canonical meaning of *lost*, in the

[7] The adverb would be substituted by *madly* in an equivalent construction in English.

sense of *one who does not know or is unable to find his/her whereabouts*, is not possible to be applied to the subject of the sentence in this context.

3. The combination belongs to the specific vocabulary of a scientific or technical area of expertise, e.g.:

(34) *Ele **respondeu civilmente** pelo crime que cometeu*
'He **responded civically** for the crime he committed'

In (34), the {V, *Adv-mente*} pair is part of the vocabulary commonly employed in the domain of law, which accounts for the fixedness of the expression.

4. Synonymic relations between adverbs are broken in the collocational context, e.g.:

(35) *Ela chorava **copiosamente***
'She cried **copiously**[8]'

(36) *?Ela chorava **abundantemente***
'She cried **abundantly**'

Even though the adverbs *copiosamente* ('copiously') and *abundantemente* ('abundantly'), in (35) and (36) respectively, could be considered synonymous, only the adverb *copiosamente* holds a collocational value in this context, since the use of *abundantemente* renders the construction unnatural in Portuguese. We thus say that the synonymic relation between these adverbs is broken.

5. In a collocation context, the adverb holding collocational status cannot be combined with the antonymous of the verb in question, e.g.:

(37) *O time **venceu** a partida confortavelmente*
'The team **won** the match comfortably'

(38) **O time **perdeu** a partida confortavelmente*
'The team comfortably **lost** the match'

[8] The adverb would be substituted by *bitterly* or *uncontrollably* in equivalent collocations in English.

While the {V, *Adv-mente*} combination in (37) can be considered a collocation in Portuguese, the antonymous of the verb seems to impede a coherent construction in (38), which would denote the collocational value of the pair in (37). Naturally, this criterion only holds true if an antonymous form of the verb exists in the language. Equally noteworthy is the fact the simple use of the negative form of the verb does not function as a deciding parameter, as both the collocation status and coherence of the combination would be maintained in this case:

(39) *O time **não** venceu a partida confortavelmente*
'The team **did not** win the match comfortably'

6. The adverb can be combined with often only one subset of the possible meanings of the verb, e.g.:

(40) *A secretária **reproduziu fielmente** os documentos*
'The secretary **reproduced** the documents **faithfully**'

(41) **Coelhos **reproduzem-se fielmente***
'Rabbits **reproduce faithfully**'

While the adverb *fielmente* ('faithfully') can be combined with the verb *reproduzir* ('to reproduce') in (40), the combination is not possible in (41), as the verb in this sentence, albeit having the same form as in (40), has a different meaning and syntactic construction.

7. As a general guiding parameter, it was also established that when different verbs of similar meaning are possible to be combined with the same adverb, the classification should be as permissive as possible towards classifying the {V, Adv} pairs as collocations. For example:

(42) *A professora **criticou** duramente o aluno*
'The teacher **criticised** the student hard-ly'

(43) *A professora **repreendeu** duramente o aluno*
'The teacher **reprimanded** the student hard-ly'

The verbs in (42) and (43) have very close meanings in Portuguese and both can be combined with the adverb *duramente* ('hard-ly'). In situations like this, it was

established that the collocation status would be applied to all synonym or quasi-synonym combinations that could be formed with a single adverb.

Concerning verb meaning, the verb classes described by Baptista (2010) and Levin (1993) were taken into account as a guiding parameter, denoting groups of verbs that share the same or similar syntactic (Batista, 2010) and semantic (Levin, 1993) traits.

Even though the criteria described above proved extremely valuable in guiding the classification task, by no means they exhaust all linguistic contexts that would denote the presence of a collocational pattern. As previously pointed out, the definition of collocation is an extremely challenging linguistic concept that has not yet reached a consensus in the literature.

4. 3 Assessing Native Speakers' Intuitions

In order to test the intuition of native speakers of Portuguese with regard to the collocational status of the linguistic pattern investigated, a sample classification task was carried out with 21 subjects, of which 13 were native speakers of European Portuguese and 8 of Brazilian Portuguese. The dataset to be classified was composed of 30 collocation candidates randomly selected, 15 having been previously classified as collocations, and 15 as non-collocations.

The candidate pairs were presented to the subjects in the contexts where they actually occurred in the corpus, with the {V, *Adv-mente*} pairs being highlighted in each sentence.

Prior to making a decision on the collocational status of the pairs, annotators were asked to attentively consider a set of guiding criteria that should be taken into account for the classification, which is the list of linguistic criteria that figures in Section 4.2. For the purpose of the task, the criteria have been presented to the subjects in a simplified version that did not include much theoretical reasoning, which could undesirably pose a higher level of complexity to the task. The full version of the questionnaire used in the experiment, including the candidate pairs to be classified, along with a summary of responses, can be found in Appendix B.

Results of Cohen's κ statistic chance-corrected inter-annotator agreement (Cohen, 1960) for the entire set of 30 pairs randomly selected for the experiment are

presented in Table 4.1. Results based solely on the 15 pairs that had been previously classified as collocations are presented in Table 4.2.

κ for 30 randomly selected candidates	
Percent of overall agreement	0.57
Fixed-marginal kappa	0.06

Table 4.1 κ for 30 randomly selected pairs of collocation candidates

κ for 15 cases of collocation in the sample	
Percent of overall agreement	0.62
Fixed-marginal kappa	0.10

Table 4.2 κ for 15 pairs among random selection previously classified as collocations

Cohen's κ values can vary from -1.0 to 1.0, where 0 would represent chance agreement. The κ results for the entire set of randomly selected collocation candidates and just for the cases previously classified as collocations were of 0.06 and 0.10 respectively, which can be considered to stand in the range of slight agreement according to the scale used to interpret κ values proposed by Landis and Koch (1977).

Even though these results are above what could be considered agreement by chance, they can be arguably deemed low. The most likely reason for this lies in the fact that the sample used in the experiment was too small, requiring an extremely high raw agreement percentage in order for the κ value to reach higher levels of significance. Because of this, κ values achieved in the experiment do not allow for definitive conclusions to be taken with respect to the agreement of the recruited linguists on the collocational status of the pairs that figure in the sample. The limited size of the sample was due to the foreseen resistance that a larger sample would most likely find among potential voluntary annotators, and to the risk of losing consistency if a larger list of examples had been presented to them.

Other reasons that would account for the low κ value lie in the random selection of cases to be classified and/or in the difficulty of the task itself. With regard to the first alternative, even though the selection was entirely random, it included candidate pairs that arguably stand in the fringes of what can be considered a collocation. One example of such pair is *decidir conjuntamente* ('to decide collectively'), which

despite the fact of being considered a collocation[9] can be deemed to stand in the borderline of this classification, having been classified as a collocation by 13 linguists, and as a non-collocation by 8, denoting an extremely low agreement for this pair in particular. The fuzziness of this case is further corroborated by the low values it presented for statistical association measures such as Mutual Information, Log-Likelihood Ratio and Dice Coefficient. Values of these measures for the referred case have been of 3.89, 25.4, and 0.0001, respectively, which places the pair roughly in the bottom 25% of collocation candidates if the list is ranked according to these measures.

As to the difficulty posed by the classification task, it would lie in the elusive nature of the very concept of collocation, which has not yet reached a consensus in the literature, as pointed out in Chapter 2. The low agreement obtained and the elusiveness of this concept suggest that annotators should undergo extensive and rigorous training before engaging in the classification task, which poses a number of operational difficulties to experiments of this kind.

Still in respect to the elusive nature of the concept of collocation, it can be seen in Tables 4.1 and 4.2 that the agreement achieved among cases that had been previously classified as collocations was higher than the overall agreement. This denotes that identifying negative cases poses more difficulty than identifying positive ones, which only confirms that the limit between both is far from being clear-cut. Considering just the positive cases, it can be noted that a raw agreement of 62% has been reached, which, despite the low κ value, could be considered to be indicative in some degree of the collocational phenomenon dealt with.

4. 4 Correlation of Results with Statistical Association Measures

A number of statistical association measures have already been tested for capturing the linguistic phenomenon of collocations. As seen in Chapter 2, Pecina (2010) provides an extensive account in this respect, remarking the particularly good performance of Unigram Subtuples (UnigSub) (Pecina, 2010) and Mutual Information (MI) (Fano, 1961) for large-sized corpora. Seretan (2011), in turn, mentions the

[9] This pair has an English equivalent as an entry in the Oxford Collocations Dictionary (Oxford, 2009)

appropriateness of Log-likelihood Ratio (LLR) (Dunning, 1993) for capturing low-frequency word combinations. In this Section, the manual classification of the collocation candidates will be contrasted with association measures that have received significant attention in previous research. The aim of this comparison is to unveil the measures that are most sensitive to the specific collocational pattern investigated, i.e. {V, *Adv-mente*} pairs in Portuguese.

The following measures were chosen for the experiment: *t* test, Pearson's chi-square (χ^2), Mutual Information (MI), Log-likelihood Ratio (LLR), Dice Coefficient (Dice), Unigram Subtuples (UnigSub). The formula of each measure has been provided in Appendix 1.

The entire set composed of 5,973 collocation candidates, already classified by a linguist as to their collocational status, was stratified into three subsets according to the frequency of the bigrams in the corpus. The first subset (S1) included word pairs with a frequency higher than one hundred; the second subset (S2) included pairs with frequency between one hundred and ten; and the third subset (S3) included pairs with frequency between ten and five. S1, S2, and S3 represent, respectively, the top, middle, and bottom of the frequency range of the collocation candidates, and include both cases that were classified as collocations and as non-collocations. The number of collocation candidates in each subset is presented in Table 4.3.

	Frequency Range	# Candidate bigrams	# Collocations
S1	> 100	65	39
S2	100 - 10	2700	334
S3	5 - 10	3208	128

Table 4.3 Number of collocation candidates per frequency

The *t* test and χ^2 are both measures that have pre-established statistical significance thresholds for the analysis of results. The performance of these two measures was analysed in terms of precision, recall, and *F*-measure, taking into account a threshold value of 2.576 for the *t* test, and 3.841 for χ^2, values that correspond to a confidence level of $\alpha = 0{,}005$ and $\alpha = 0{,}05$, which have been previously adopted in similar contexts aimed at identifying collocations (Manning and

Schütze, 1999: 153; 159). Results of these two measures for S1, S2, and S3 separately as well as for the set considered altogether are shown in tables 4.4 and 4.5.

	t test		
	Precision	Recall	*F*-measure
S1	0.603	0.974	0.745
S2	0.129	0.937	0.227
S3	0.082	0.460	0.140
All	0.128	0.818	0.222

Table 4.4. *t* test results on collocation candidates

	χ^2		
	Precision	Recall	*F*-measure
S1	0.609	1	0.757
S2	0.123	0.964	0.218
S3	0.041	1	0.079
All	0.084	0.976	0.156

Table 4.5. χ^2 results on collocation candidates

Figures in Tables 4.4 and 4.5 clearly denote that the *t* test and χ^2 fell far short of identifying the collocation pattern investigated. The reason behind the poor performance of these measures is most likely connected to the low frequency of the linguistic phenomenon dealt with, a fact that has already been reported in the literature with regard to the *t* test (Dunning, 1993; Seretan, 2011). The reason why this measure can be considered inappropriate for capturing low-frequency candidates lies in the fact that it assumes a normal distribution of events, which renders it unreliable for rare occurrences (Dunning, 1993). Despite the fact that the χ^2 makes up for the assumption of normal distribution (Manning and Schütze, 1999: 158; Seretan, 2011: 43) and is usually deemed to provide more reliable results in comparison with the *t* test in the task of extracting collocations from corpora (Manning and Schütze, 1999), the empiric experiments carried out in this study have shown that this measure would also be considered inappropriate for extracting the pattern {V, *Adv*-mente}. Both association measures presented similar values for Precision, Recall, and *F*-measure for the most frequent case, with the *t* test presenting a slightly better

F-measure for infrequent pairs. One known disadvantage of the χ^2 is the fact that it overemphasises low-frequency events (Kilgarriff, 1996: 35), which is in fact corroborated by the high number of false positive cases it presented in this experiment. It can also be observed in Tables 4.4 and 4.5 that the higher the frequency of the collocation candidates in the corpus, the more satisfactory the performance of the *t* test and χ^2 are in identifying the phenomenon. The *F*-measure of both tests increases from S3 to S1.

The other association measures applied to the collocation candidates extracted from the corpus – namely MI, LLR, Dice, and UnigSub – do not have a pre-established threshold for filtering results[10]. The correlation of these measures with the binary classification of collocation candidates was assessed based on the Pearson product moment correlation coefficient, *r* (Pearson, 1896)[11], which measures the linear relationship between two variables – in this case, the referred measures and the classification of bigrams as (non-)collocations. Pearson's *r* values for the aforementioned measures, considering S1, S2, and S3 and the set altogether, are presented in Table 4.6.

	Pearson Correlation Coefficient (*r*)						
	t test	χ^2	MI	LLR	Dice	UnigSub	# Instances
S1	0.0321	0.2358	*0.4562*	*0.3610*	*0.3831*	*0.3469*	65
S2	0.0759	0.0633	0.2876	0.2403	0.1711	*0.3816*	2700
S3	0.1126	0.0447	*0.3137*	*0.3312*	0.1144	0.1707	3208
All	0.1519	0.0528	*0.3093*	*0.3109*	0.2287	*0.3453*	5,973

Table 4.6. Pearson results for *t* test, χ^2, MI, LLR, Dice, and UnigSub for considering the classification of collocation candidates

Values for *r* can range from -1.0 to 1.0. According to Cohen (1988), an *r* of .10 could be considered to have a small *effect size* (ES), while an *r* of ± .30 would have a medium ES, and an *r* equal to or above .50 ($r \geq .50$), a large ES. In other words, the furthest the *r* value is from zero, the stronger the relationship between the two

10 Certainly a decision can always be made with regard to a threshold value to be applied to results based on specific circumstances of the problem dealt with.

variables analysed should be. While the sign of *r* can be established as either positive or negative in advance, both positive and negative values of *r* can be considered to assess the strength of correlations. Even though the sign of *r* has not been previously established in this experiment, the results obtained were all positive

In Table 4.6, it can be observed that the four association measures presented a medium ES for S1, the subset including collocation candidates with higher frequency in the corpus. Concerning S3, the *r* value of Dice and UnigSub presented a considerably small ES, which stood at approximately 0.1 for both measures. The small ES of *r* for Dice and UnigSub seems to suggest that these two measures are not appropriate to capture the collocation pattern investigated when it occurs infrequently. LLR, on the other hand, has maintained *r* values from 0.24 to 0.36 across the three subsets. This corroborates findings of previous research that affirm this measure could be deemed reliable for the task of collocation extraction in general (Daille, 1994; Evert, 2005; Orliac, 2006; Seretan, 2011), since it would be sensitive to both high and low-frequency phenomena (Dunning, 1993: 62). MI showed a similar trend in this respect, with *r* values ranging from 0.28 to 0.45, where the lowest value corresponds to S2, the subset including pairs of medium frequency in the corpus. This was also the case with LLR, whose lowest *r* value was also the one corresponding to S2.

Considering the entire set of collocation candidates, UnigSub, LLR, and MI were, in descending order, the measures to present the highest correlation with human annotations on the collocation status of the pairs. The *t* and χ^2 tests presented a notably low correlation with the annotations, which seems to confirm the poor Precision, Recall and *F*-measure results of these two measures.

The strategy of combining different association measures to enhance the extraction of collocations from corpora has already been reported in previous research (Pecina and Schlesinger, 2006; Portela, 2011). The advantage of this strategy would lie in the fact that different measures might have different levels of sensitivity in respect to a given collocational pattern. In order to reveal how the measures adopted in this study correlate with each other in view of {V, *Adv-mente*} pairs, the *r* coefficient between these measures has been calculated. Results can be seen in Table 4.7.

	Pearson Correlation Coefficient (*r*)				
	χ^2	MI	Dice	UnigSub	LLR
t test	-0.0002	*0.2355*	*0.2968*	0.0480	*0.3823*
χ^2		0.0091	-0.0060	0.0251	-0.0041
MI			*0.2331*	*0.2456*	0.2215
Dice				0.0081	*0.6377*
UnigSub					0.0081

Table 4.7. *r* values between association measures

Results in Table 4.7 reflect how correlate the association measures are between themselves in terms of *r*, where the higher the correlation between two given measures, the more overlapping there would be in the performance of these measures in identifying {V, *Adv-mente*} collocations in the corpus. As it can be seen in the table, Dice and LLR were the measures that presented the highest degree of correlation, with an *r* value of 0.637, which can be considered to have a large ES (Cohen, 1988). This result could arguably lead to the conclusion that these two measures have very similar sensitivity to the collocational pattern under study, and would capture similar sets of collocational bigrams.

LLR also correlates well with the *t* test, the *r* value between these measures being 0.38, which denotes a medium ES. Since the *t* test has a pre-established significance threshold for the analysis of results, the precision of the *t* test for the instances above and below this threshold is compared with the precision of an equivalent LLR threshold for the same instances. The threshold value considered for the *t* test is 2.576, which corresponds to a confidence level of $\alpha = 0{,}005$. If all collocation candidates are ranked according to LLR, the LLR value that occupies the same position as the *t* test threshold in the list is 49.109. This has been the value considered to assess the precision of this measure. Results of the comparison between *t* test and LLR precisions are shown in Table 4.8.

	t test Precision	LLR Precision
$t \geq \rho = 2.576$ (# 3157)	0.129	0.126
$t < \rho = 2.576$ (# 2813)	0.033	0.019

Table 4.8 *t* and LLR Precision for instances above and below the *t* threshold of 2.576

Results in Table 4.8 denote that, even though LLR presented a stronger correlation with the human classification of candidates in comparison with the *t* test, if a threshold value equivalent of the *t* test's is applied to LLR, the precision of the latter also falls far short of satisfactory. Its precision was in fact slightly worse than that of the *t* test for the same instances.

In view of the medium correlation between the LLR results and the human classification of collocation candidates, it is reasonable to assume that this measure is sensitive, to a certain degree, to the collocational pattern dealt with. However, when a threshold is applied to the results, the precision achieved is considerably poor, which leads to the assumption that the threshold in question, established based on the *t* test's, is the reason for the poor precision obtained. In this line of reasoning, we have experimented with a higher LLR threshold, and checked to see if any improvement could be observed.

Considering the distribution table of the *t* test, provided in Manning and Schütze (1999: Appendix), the most rigorous *t* threshold would be 3.905, which corresponds to a confidence level of $\alpha = 0{,}0005$ for an infinite degree of freedom. The equivalent threshold for LLR considering the list of collocation candidates would be 93.013 – which is the value that, if the list is ranked according to the LLR, occupies the same position as the *t* test threshold. The performance of the more rigorous threshold for *t* compared with its equivalent value estimated for LLR is presented in Table 4.9.

	t test Precision	LLR Precision
$t \geq \rho = 3.905$ (# 1313)	0.181	0.272
$t < \rho = 3.905$ (# 4657)	0.056	0.030

Table 4.9 *t* and LLR Precision for instances above and below the *t* threshold of 3.905

As shown in Table 4.9, the precision of both the *t* test and LLR improve in a small degree if a more rigorous threshold is applied to the results, with LLR presenting an improvement slightly more pronounced in that respect for cases that cross the threshold. However, the results obtained are still unsatisfactory. The fact that LLR presents an unsatisfactory threshold-based performance, albeit having a medium correlation with the classification of collocation candidates, suggests that considering results in view of significance thresholds might not be the most appropriate to identify the collocational pattern investigated. This assumption is corroborated by the lack of a clear-cut division in the results of LLR, and also the other measures, that would be able to separate positive and negative cases. It seems that any value that is chosen as a threshold based on the human classification of candidates would either leave out too many positive cases or include too many negative ones.

In that way, we have attempted to train an automatic collocation classifier for {V, *Adv-mente*} pairs by applying machine learning techniques to the table of collocation candidates and their respective human classification and association measure results. This approach disregards any decision based on critical values, and, instead, takes into account results from all association as being potentially useful for the classification task. This experiment is described in detail in Chapter 5.

Chapter 5. Training an Automatic Collocation Classifier

5.1 Using All Classified Collocation Candidates as a Training Set

Given the difficulty in identifying a pattern that reflects the type of collocation dealt with in the results of the association measures, we have experimented to train an automatic collocation classifier by applying machine learning algorithms to the results of these measures. The 3-6-6 version of the WEKA Toolkit[12] (Witten et al., 2011) has been used for that purpose. The performance of the different supervised machine learning algorithms that compose the tool has been tested based on the manual classification of collocation candidates performed by a linguist. The training set consists of the manually annotated list of collocation candidates, accompanied by results of the association measures used in this study hitherto, namely *t* test, Chi-Square (χ^2), Log-Likelihood Ratio (LLR), Mutual Information (MI), Dice Coefficient (Dice), and Unigram Subtuples (UnigSub).

A number of classifiers grouped according to different algorithmic methods are available on the WEKA Toolkit. First, we have assessed the performance of WEKA classifiers in an attempt to identify the one in each algorithmic group that would achieve the best results in the classification. Multi-instance classifiers have been disregarded in this experiment since the nature of classification dealt with does not match the type of classification problems multi-instance classifiers usually address[13]. Classifiers that presented too poor or insignificant results, potentially denoting an incompatibility with the task, were also not considered. Altogether, the performance of 45 classifiers was tested. Table 5.1 shows all the classifiers considered in the experiment. Table 5.2 shows results based on a ten-fold cross-validation for the best classifier of each type, ranked in descending order according to F-measure values.

[12] http://www.cs.waikato.ac.nz/ml/weka/ [Accessed 15 May 2012]

[13] Examples of typical classification problems for which multi-instance classifiers are adopted can be seen in Xu (2003).

Bayesian Classifiers	Bagging	**Trees**
BayesNetwork	ClassificationViaClustering	ADTree
NaïveBayes	ClassificationViaRegression	BFTree
NaiveBayesSimple	Dagging	J48
Functions	Decorate	LADTree
LibSVM	LogitBoost	LMT
Logistic	RacedIncrementalLogitBoost	NBTree
RBFNetwork	RandomCommittee	RandomForest
SMO	RandomSubSpace	RandomTree
SPegasos	RotationForest	FT
VotedPerceptron	**Rules**	REPTree
MultilayerPerceptron	DecisionTable	SimpleCart
SimpleLogistic	DTNB	**Lazy**
Miscellaneous Classifiers	JRip	IB1
HiperPipes	NNge	KStar
VFI	OneR	LWL
Meta Classifiers	PART	
AdaBootsM1	Ridor	

Table 5.1 Classifiers whose performance was tested

			Weighted Average Results			
Type	Classifiers	# Collocations (out of 501)	Precision	Recall	*F*-Measure	Overall Ranking
Tree	LADTree	172	*0.918*	*0.929*	*0.919*	1
Rule	JRip	184	0.915	0.926	0.918	3
Meta	LogitBoost	158	0.917	*0.929*	0.917	5
Function	Logistic	123	0.919	*0.929*	0.913	15
Lazy	KStar	177	0.905	0.916	0.909	28
Bayesian	NaïveBayes	92	0.895	0.916	0.898	34
Miscellaneous	VF1	160	0.851	0.685	0.752	43

Table 5.2 Performance of classifiers with best results among each method

As seen in Table 5.2, LADTree has been, of all classifiers tested, the one that presented the best results. JRip and LogitBoost, in turn, have been the best classifiers among the group of rules and Meta classifiers, respectively.

Considering the overall ranking, out of the top ten classifiers, 6 were Trees, 2 were Meta, and 2 were Rules, which seems to indicate that these three groups are particularly adequate for the classification task at hand.

We have also experimented to exclude results of the *t* test and χ^2 from the training set, since these two measures have not presented a good correlation with the human classification of collocation candidates. However, it has been observed that the absence of these measures incurred, in fact, in poorer final results. This seems to imply that WEKA classifiers are sensitive to these measures in some degree, and for that reason they have been maintained in the training set.

5.2 Experimenting with a Balanced Training Set

It is noteworthy that the training set used in this experiment was considerably disproportional with regard to its number of positive and negative cases, since 501 bigrams had been manually classified as collocations out of the 5,973 that compose the set. Bearing in mind that a classifier of this kind should be expected to capture as many true cases of collocations as possible, this discrepancy might set too favourable a condition for the performance of the classifiers if results are considered in terms of weighted averages. In an attempt to compensate this discrepancy, we have also tested the performance of the classifiers on a balanced training set, in which the number of negative cases has been reduced to 501, which equals the number of positive ones. Results of the best classifier of each type, ranked in descending order according to F-measure values, are presented in Table 5.3.

			Weighted Average Results			
Type	Classifiers	# Collocations (out of 501)	Precision	Recall	*F*-Measure	Overall Ranking
Meta	RotationForest	411	*0.816*	*0.816*	*0.816*	1
Tree	LMT	387	0.807	0.806	0.806	3
Rule	JRip	413	0.805	0.804	0.804	5
Function	Logistic	412	0.803	0.802	0.802	7
Bayesian	BayesNetwork	389	0.793	0.792	0.792	17
Lazy	IB1	384	0.75	0.755	0.755	31
Miscellaneous	VF1	495	0.752	0.602	0.532	42

Table 5.3 Results of best classifiers on balanced training set

As seen in Table 5.3, results based on the balanced training set are not as good – as it was already expected – but the performance of most classifiers could arguably be considered satisfactory nonetheless, with RotationForest presenting the best overall result. It is interesting to notice that results for some of the classifiers changed considerably between the balanced and unbalanced training sets, with SMO, ClassificationViaRegression, OneR, and NBTree being outperformed by Logistic, RotationForest, JRip, and LMT, respectively. Also noteworthy is the fact that the number of true cases of collocations retrieved has increased with the use of the balanced set, which, contrary to what has been observed previously, is not necessarily associated with a particularly poor precision in the classification of the positive cases, as results in that respect range between 55 and 81 per cent.

5.3 Combining Classifiers

Since each algorithm is based on a different classifying method, we have also attempted to combine classifiers of different types and checked to see if any improvement could be observed in the results. The combination strategy adopted consists in taking into account the vote of each classifier with regard to a given class and in the end utilise the class that has received the largest number of votes. This technique can be implemented by making use of the Vote algorithm, which is part of the WEKA toolkit.

We have first attempted to combine the classifiers of each algorithmic group that presented the best performance, but results achieved were unable to outperform those obtained for RotationForest alone, the classifier with the best overall performance on the balanced training set. In view of that, new attempts have been made with other combinations in order to verify which classifiers seemed to contribute the most to the performance of the group as a whole. The most significant results of this experiment are presented in Table 5.4.

Combinations	# Collocations (out of 501)	Weighted Averages		
		Precision	Recall	*F*-Measure
RotationForest LMT	419	*0.824*	*0.823*	*0.823*
Logistic RotationForest LMT	416	0.819	0.818	0.818
Best of each type	409	0.806	0.805	0.805

Table 5.4 Results of combined classifiers on balanced training set

As previously mentioned, the combination that included the best classifier of each algorithmic group was not able to improve results of the best overall classifier isolated. However, more selective combinations that include fewer classifiers have proven to present more promising results, outperforming the best classifier alone. As it can be observed in Tables 5.3 and 5.4, the *F*-Measure achieved by the RotationForest algorithm isolated, a Meta classifier, has increased from 81.6 per cent to 81.8 if it were combined with the Logistic and LMT algorithms – which are function and rule-based classifiers, respectively –, and to 82 per cent, if combined just with LMT. The number of true collocations retrieved, in turn, has increased in 8 cases considering the combination RotationForest and LMT.

5.4 Results for a Different Evaluation Set

The strategy employed in this study takes results of statistical association measures as an indication of the collocational status of word pairs based on a reference manual classification. Since the results of these measures are known to be influenced by the frequency of the events whose association is being assessed – in our case, word pairs – we have attempted to evaluate the performance of *RForestLMT* with data from a different, smaller corpus.

The *NILC/São Carlos* corpus of Brazilian Portuguese, with 31,2M words from journalistic texts, was used for this experiment. Out of the 501 collocations originally retrieved from the *CETEMPúblico* corpus, 297 also occur in the *NILC/São Carlos*. In that way, these 297 bigrams have been used to evaluate the degree of influence that the

frequency of the pairs in the corpus would exert in the classification. The search for the bigrams in the *NILC/São Carlos* was based on the adjacent co-occurrence of the terms within a sliding window of up to three words[14].

We have tested the *RForestLMT* model, trained on the balanced set mentioned in Section 5.2, to see how consistent its classification would be for the same data retrieved from different corpora. First, the 297 pairs with the previously seen association measure results based on the *CETEMPúblico* corpus were used as evaluation set. Then we used a second evaluation set composed of the same bigrams, but with unseen association measure results based on the considerably smaller *NILC/São Carlos* corpus. Out of the 297 cases that occur in both corpora, 82 are *hapax legomena* in the *NILC/São Carlos*, i.e. they occur only once in the corpus. In that way we have also experimented to exclude these cases from the set and check if any improvement would be observed in the performance of the model. Results of these evaluations are presented in Table 5.5.

Evaluation sets	Precision	Recall	*F*-Measure
CETEMPúblico (297 bigrams)	1	1	1
NILC/São Carlos (297 bigrams)	1	0.5	0.667
NILC/São Carlos without *hapax* (215 bigrams)	1	0.579	0.733

Table 5.5 Performance of *RForestLMT* on data from a different corpus

As it would be expected, the *RForestLMT* model had an F-measure of 100 per cent for the 297 bigrams retrieved from the *CETEMPúblico* corpus, whose association measure results had been previously seen at the moment of training. However, considering the same 297 pairs but with statistical association measures from a smaller corpus, the model had a recall of 50 per cent. This result is improved, however, if cases of *hapax legomena* are excluded from the set, as it can be seen from the table.

[14] This degree of separation has been previously used in Manning and Schütze (1999: 148) to extract collocations from corpora.

Even though the results obtained with a different corpus seem to fall short of satisfactory, the adverse conditions set by the difference in size between the two corpora should perhaps be taken into account as too challenging to be overcome by the model. The *NILC/São Carlos* corpus is approximately six times smaller than the *CETEMPúblico*. This certainly contributes for the infrequency of certain pairs in the former, resulting in too discrepant association measure results. The pair *adoecer gravemente* ('to fall gravely ill'), for example, is a *hapax legomenon* in the *NILC/São Carlos* corpus, with a *t* test result of 0.99. The same pair occurs 25 times in the *CETEMPúblico*, with a *t* of 4.92 – and yet, despite this difference, it should be considered a collocation in Portuguese.

In that way, given the extreme size difference between the corpora, it appears that a recall of 0.5 could perhaps be considered suggestive that, if applied to data retrieved from a corpus that is closer to the *CETEMPúblico* in size, the model should be expected to yield more promising results. Arguably indicative of this is the fact that an F-measure of 0.73 was reached when all *hapax legomena* were excluded from the evaluation set with association measures based on the *NILC/São Carlos*. This is a result that is reasonably close to the ones obtained with the ten-fold cross-validation based on the balanced training set.

5.5 Comparing Human and Machine Classifications

Considering the set given to native speakers of Portuguese for classification, we have also compared the performance of the 21 linguists who took part in the experiment with the performance of *RForestLMT* in classifying the same set. For this comparison, it has been established that the classification carried out by the linguists would be considered based on the vote of the majority. Results for this comparison are shown in Table 5.6.

			Weighted Average Results		
Classifiers	# Collocations (out of 15)	Precision Positive Cases	Precision	Recall	*F*-Measure
RForestLMT	12	*0.706*	*0.738*	*0.733*	*0.732*
21 Linguists	12	0.521	0.393	0.685	0.497

Table 5.6 Performances of *RForestLMT* and linguists based on reference classification

As seen in Table 5.6, *RForestLMT* has outperformed the 21 linguists based on the reference classification. As already mentioned in Chapter 4, there are a number of reasons for the poor agreement of the linguists with the reference classification, including the small scale of the set, and the complexity of the task itself, which in ideal circumstances should be preceded by extensive training. Particularly noteworthy, however, is the reasonably good recall of positive cases in the classification performed by linguists. Out of the 15 cases that had been previously tagged as collocations, 12 were captured by the majority of the linguists who took part in the task - which equals the recall achieved by *RForestLMT* in that respect. This confirms even further that more training would be necessary to achieve more significant weighted average results. As already mentioned in Chapter 4, it seems that the critical problem in the classification carried out by linguists lies in the correct identification of negative cases, i.e. what is not a collocation.

The following Chapter describes the process in which the 501 manually classified cases of {V, *Adv-mente*} collocations extracted from the *CETEMPúblico* corpus are aligned with their equivalents in English in order to build a bilingual lexicon of this pattern. The lexicon is then used as reference for the evaluation of MT engines with regard to the correct PT>EN translation of this type of collocation.

Chapter 6. A Bilingual PT>EN Collocation Lexicon and MT Evaluation

6.1 Building the Lexicon

Having information on word collocational patterns is important for a number of areas, including second language learning and NLP. Equally important is to have information on collocational equivalents between languages, since the translation of a collocation is not necessarily done on a word-by-word basis.

In that way, we have attempted to build a bilingual collocation lexicon containing the pattern {V, *Adv-mente*}, having Portuguese and English as source and target language, respectively.

Even though techniques for automatically extracting translation equivalents from comparable corpora are already available – through initiatives such as the *Terminology Extraction, Translation Tools and Comparable Corpora Project*[15] (*TTC*), for example – the approach here adopted is focused on a more linguistic/comparative analysis, and does not have as its main aim the provision of large-scale resources, but rather a research-oriented investigation of a linguistic pattern that can be considered understudied with respect to its collocational potential.

The {V, *Adv-mente*} pattern could be deemed to pose a rather subtler problem to translation since more than one possible combination is often available to express the same, or very close meanings either in the source or in the target language. That would be the case of the bigrams *chorar convulsivamente* ('cry convulsively') and *chorar copiosamente* ('cry copiously'), for example, where both can be used to convey the idea of *crying in excess*. In English, these pairs would be translated into *cry bitterly* or *cry uncontrollably*, pairs that could also be arguably deemed roughly interchangeable in the meaning they convey. In that way, we have attempted to find equivalents in English for the collocations extracted from the Portuguese corpus, also grouping pairs that, as the ones mentioned above, could be placed together with respect to the meaning conveyed. The question as to whether there are contexts that render the use of a given collocation preferable to the use of a supposedly equivalent one is beyond the

[15] http://www.ttc-project.eu/ [Accessed 10 May 2012]

investigation here undertaken. The main aim of the PT>EN lexicon compiled is to make available a range of collocation options that can be adopted to convey a given meaning in either one of the languages.

English equivalents for the 501 Portuguese collocations were retrieved both from a collocation dictionary and from parallel corpora. The *Oxford Collocations Dictionary* (Oxford, 2009) was the source that provided the bulk of equivalent combinations, followed by the English-Portuguese *Europarl* parallel corpus[16] (Koehn, 2005), the *COMPARA* Portuguese-English parallel corpus[17] (Frankenberg-Garcia and Santos, 2002), and the journalistic version of the Portuguese-English *CorTrad* parallel corpus[18] (Tagnin, 2010), developed in the framework of the *COMET* Project[19]. The breakdown of combinations found in each one of these sources is presented in Table 6.1.

	# Equivalents
Oxford Collocations Dictionary	427
Europarl	18
Compara	11
CorTrad	5
Not found in any of the sources	40
Total of equivalents found (out of 501)	461

Table 6.1 Sources of translation equivalents

The *Oxford Collocations Dictionary* was the base source for the establishment of equivalences, where priority was given to this dictionary over the other sources. That means that not necessarily an equivalent that was found in the *Oxford Collocations Dictionary* did not exist in the parallel corpora. The dictionary was prioritised because, more than just translation equivalents, it provides combinations that are assuredly considered to be *collocations* in the target language, which, in turn, also serves as extra validation for the classification of the Portuguese pairs.

16 http://www.statmt.org/europarl/ [Accessed 10 May 2012]

17 http://www.linguateca.pt/COMPARA/ [Accessed 10 May 2012]

18 http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html [Accessed 10 May 2012]

19 http://www.fflch.usp.br/dlm/comet/ [Accessed 10 May 2012]

As it can be seen from Table 6.1, out of the total of 501 bigrams classified as collocations in Portuguese, there were 40 for which translation equivalents in English were not found in any of the sources consulted. Reasons for this gap lie most likely in the fact that these expressions are essentially typical of Portuguese usage contexts, which requires the use of paraphrase or adaptation in the translation. That would be the case, for example, of *responder criminalmente* ('to respond criminally'), which is a combination typical of the law jargon in Portuguese and for which no direct equivalent in English has been found in the *Oxford Collocations Dictionary* or in the parallel corpora. Even though constructions such as *to be held responsible for a crime* could be arguably mentioned as an English equivalent, phrases of this type denote a change in structure that differ from the verb-adverb pattern under study.

The methodology adopted for the dictionary search was based on the literal translation of the verb. Considering the pair *analisar detalhadamente* ('to analyse detail-ly'), for example, an equivalent combination was searched in the dictionary based on the literal translation of the verb into English, namely *analyse*. Then, from the adverbs that could be collocated with this verb according to the dictionary, pairs that have an equivalent meaning to the combination in Portuguese were chosen as target equivalent collocations. In the case of *analisar detalhadamente*, possible equivalent collocations in English, according to the *Oxford Collocations Dictionary*, would be *analyse in detail* and *analyse in depth*. As it can be noted in this example, equivalent combinations in English are not necessarily composed of *–ly* ending adverbs, but might include another type of adverb or even by an adverbial phrase, as in the case of *detalhadamente* (PT) > *in detail*, *in depth* (EN). Also noteworthy is the fact that equivalent combinations without *Adv-mente* might exist in Portuguese. It can be argued, for example, that *analisar a fundo* ('to analyse in depth') is also a collocation in Portuguese. However, the compilation of the bilingual lexicon was carried out in the direction Portuguese-English at this stage, having the {V, *Adv-mente*} collocations in Portuguese as a starting point.

The search for equivalents in the parallel corpora was based on surface bigrams on both languages – Portuguese and English –, within a window of up to three words[20] between them.

[20] This degree of separation has been previously used in Manning and Schütze (1999:148) to extract collocations from corpora.

Based on the equivalent pairs in the lexicon, a MT evaluation task was carried out in view of the {V, *Adv-mente*} pattern. The evaluation process is described in the following section.

6.2 Evaluating MT Systems in View of the {V, *Adv-mente*} Pattern

6.2.1 The Evaluation Set

For the purpose of evaluating the performance of MT systems in translating the {V, *Adv-mente*} pattern, only combinations that were deemed to be problematic to MT were considered in the evaluation. The criterion to establish which pairs to consider was based on how morphologically/etymologically different the combinations in Portuguese were from their equivalents in English. The pair *sustentar financeiramente* ('to support financially'), for example, was in the group that was considered to pose low difficulty to translation, since its equivalent in English, *support financially* – found in the *Oxford Collocations Dictionary* – does not differ considerably from the combination in Portuguese, from the morphological/etymological point of view. On the other hand, pairs such as *mentir descaradamente* ('to lie shamelessly') would have equivalents such as *lie blatantly* – found in the *Oxford Collocations Dictionary* –, a combination where the adverb differs morphologically/etymologically from the one in Portuguese. Only cases such as the latter were taken into account whilst evaluating the MT systems.

However, English equivalents that incurred in a deviation from the syntactic pattern verb-adverb were not considered for the evaluation. That would be the case of the example *responsabilizar criminalmente* (PT) > *to be held responsible for a crime* (EN). In that way, in order to be taken into account for the evaluation, equivalents should not represent a deep syntactical change as the one exemplified above, and yet be morphologically/etymologically different from their version in Portuguese.

Out of the 461 pairs for which translation equivalents were found, 79 pairs were selected to compose the evaluation set. Original contexts of occurrence of these pairs were retrieved from the *CETEMPúblico* corpus for the evaluation, where three sentences were randomly extracted for each pair in order to provide different contexts

for each bigram in the evaluation of the MT systems (this process will be explained in detail in Section 6.2.3).

Albeit randomly selected, a number of criteria have been established to filter the sentences where the pairs occurred in the corpus. These criteria are outlined below:

i) Sentences should be not too short neither too long, ranging between 90 and 240 characters (with spaces);
ii) The verb-adverb pair should not occur at the end nor at the beginning of the sentence;
iii) The verb-adverb pair should not be in the immediate vicinity of punctuation marks, proper nouns, or abbreviations;

The objective of these criteria is to avoid patterns that are known to pose difficulties to MT. Such patterns have been referred to in previous research as Negative Translatability Indicators (Underwood and Jongejan, 2001; Bernth and Gdaniec, 2000; and Gdaniec, 1994).

After the retrieval and selection process, three different sentences for each verb-adverb pair were retrieved from the corpus to form the evaluation set. The set was composed of 237 sentences in total, consisting of three subsets of 79 sentences each, all of which conforming to the criteria described above.

6.2.2 The Criteria for Evaluation

Despite the existence of automatic metrics for the evaluation of MT, we have opted to carry out a manual evaluation in this study, since most automatic MT quality metrics available nowadays require the existence of reference translations, which are not available for the sentences extracted from the Portuguese corpus. In addition, a manual approach provides a higher degree of freedom in the evaluation, which makes a difference in our case since it is the automatic translation of a specific linguistic pattern that is being assessed, and not MT in general.

In that way, a set of empiric criteria have been established to assess the translation of the 79 {*V, Adv-mente*} pairs that composed the evaluation set. A three-point scale has been devised for that purpose, ranging from 0 to 2. The meaning of each point in the scale is outlined below.

0 – The system does not translate both or one of the words in the pair, either by maintaining the terms in Portuguese or by supressing them; OR the translation is inaccurate to the extent of changing the meaning of the original combination.

1 – The translation is accurate but does not match with the English equivalent found in the sources consulted. Grammatical errors that do not impede comprehension and do not incur in a change of meaning are allowed into this class.

2 - The translation matches the equivalents found in the sources consulted, even if with some minor grammatical errors[21].

As it can be noted in the scale, the grammatical correctness of the translation was not the focus of the evaluation. The meaning conveyed was significantly more central, instead.

6.2.3 Results

Google Translate™ [22], Systranet™ [23] and Reverso™ [24] were the MT engines chosen for the evaluation. The performance of these three systems was tested for the 237 sentences that composed the evaluation set based on the three-point evaluation scale presented in Section 6.2.2. The number of instances translated by the MT systems that fell into categories 0, 1, and 2 for each system is presented in Table 6.2.

Class	Google (out of 237)	Systranet (out of 237)	Reverso (out of 237)	Total (out of 711)
0	28	41	103	172 (24.1%)
1	135	191	116	*442 (62.1%)*
2	74	5	18	97 (13.6%)

Table 6.2 Evaluation of MT outputs

[21] In three exceptional occasions, pairs that were not an exact match with the reference were considered to fall into class 2 due to their evident proximity with the reference bigrams. This was the case of the pairs *bang loudly*, *weep uncontrollably*, and *say with conviction*, whose reference bigrams were *beat loudly*, *cry uncontrollably*, and *say with confidence*, respectively.

[22] http://translate.google.com/ [Accessed 10 May 2012]

[23] http://www.systranet.com/translate [Accessed 10 May 2012]

[24] http://www.reverso.net/ [Accessed 10 May 2012]

As it can be seen in the Table, the majority of MT outputs fall into the class 1, where translations were accurate but did not match the reference English equivalents found in the sources consulted. Reverso™ was the system that, by a large margin, presented the largest number of cases that fell into group 0. The system that presented the largest number of translations that matched the reference was Google Translate™, followed by Reverso™ and Systranet™, respectively. The considerable percentage of pairs that fell into class 0 (24.1%) may be related to the coverage of the systems' lexicons, since it is known that *Adv-mente* are not systematically registered in the dictionaries (Fernandes, 2011), and may have been overlooked by lexicographers. The number of bigrams that fell into class 1 may confirm the difficulty of the task at hand, namely the retrieval of this collocation type, due to their lexical variation, and also due to the problems it poses to MT. These aspects are going to be further investigated below.

In order to analyse the influence of linguistic context in the automatic translation of the bigrams we have also assessed how consistent the machine translations yielded for the pairs were based on three different contexts retrieved from the corpus.

Google Translate™ would be expected to present some degree of variation in that respect, since this system is based on Statistical Machine Translation (SMT) techniques (Och, 2006). Because SMT makes use of previously translated training data, this translation strategy could be arguably deemed more susceptible to be influenced by context. As to the translation strategy used by the other MT engines, Systran™ – the system that Systranet™ is connected to – makes use of both linguistic technology and statistical techniques, being in that way a hybrid system (Systran, 2012). As to Reverso™, no precise information published by its developers has been found regarding the strategies used by this engine. Previous studies, however, claim that it would be a rule-based system (Forcada, 2000; Way and Gough, 2003).

Three situations were possible in the analysis: the three outputs falling into the same class, two of the outputs falling into one class with the third output falling into a different one, and each of the outputs falling into a different class. These three possibilities have been considered to represent *consistency*, *half-consistency*, and *inconsistency*, respectively, and the number of cases in each group is presented in Table 6.3.

Class	Google	Systranet	Reverso	Total (out of 237)
000	1	10	31	42
111	30	59	35	124
222	16	1	6	23
Consistent (out of 79)	47 (59.4%)	70 (88.6%)	*72* (91.1%)	*189* (79.7%)
002	1	1	0	2
001	5	1	2	8
110	8	6	5	19
112	8	0	0	8
220	2	0	0	2
221	5	0	0	5
Half-consistent (out of 79)	*29* (36.7%)	8 (10.5%)	*7* (8.8%)	44 (18.5%)
012	3	1	0	0
Inconsistent (out of 79)	*3* (3.7%)	1 (1.3%)	0 (0.0%)	4 (1.6%)

Table 6.3 Evaluation of the influence of context in MT

Concerning the degree of influence the systems suffer from the context environing the pairs, it can be observed in Table 6.3 that the largest number of machine-translated bigrams were consistent as to the class they fell into, comparing three different contexts of occurrence. Particularly noteworthy is the fact that the majority of consistent cases correspond to pairs that do not match the reference. Reverso™ was the system with the largest number of consistent MT outputs, with 72 bigrams, out of 79, falling into the same class. The largest set of half-consistent outputs (class 110, with 19 instances) is the case where two outputs are deemed correct but do not match the reference. Google Translate™ was the system that presented the largest number of such cases.

As it can be seen from this Table, results on MT consistency are in line with expectations, since Google Translate™ was, by a large margin, the system with the largest number of half-consistent cases. Even though the number of inconsistent cases was in general very low (4), Google Translate™ was also the system that presented the majority of cases in this group, with 3 occurrences of inconsistent translations.

6.2.4 Assessing the Fluency of MT Outputs

It is noteworthy about the three-point scale used to evaluate the systems that, while the notion of translation quality is at stake if degree 0 is compared with the other two degrees, this notion is not necessarily present in a comparison between degrees 1 and 2, since what distinguishes these two degrees is not the quality of the translation itself but how close the MT output is to the reference. In other words, the fact that the translation provided by any of MT systems differed from the versions found in the reference sources consulted does not necessarily mean that the MT output is not fluent or of poor quality.

The difference between classes 1 and 2 in the scale would thus be related to the concept of *fluent output*, formulated by Koehn (2010:94) and which has already been mentioned in Chapter 2. This concept is connected to the notion that, ideally, MT outputs should be not only *correct* but also *fluent* in the target language, with fluency being arguably connected to frequency of use, which is a criterion the author himself adopts.

In order to address this problem and shed light on the question of how fluent the machine-translated bigrams in class 1 are as opposed to their respective reference combinations, we have applied statistical association measures to the machine-translated bigrams classified as 1 and also to their reference versions as found in the sources consulted. The *Collins WordBanks Online*[25] corpus of English (HarperCollins, 2008 was used as source of distributional data. It is composed of approximately 455M words, and texts of newspapers, books, magazines, and speech, among others.

The analysis was carried out based on the different class-1 bigrams yielded by each system. Bigrams that did not occur in the *Collins WordBanks Online* corpus were excluded from the analysis. Whenever more than one English equivalent was available, the most frequent pair was selected. In the case of Google Translate™ , four cases that fell into class 1 were also excluded because they were not a verb-adverb combination. This was the case of *stare*, *misuse*, *scrutinise*, and *soar*, translations that have been yielded for the Portuguese pairs *olhar fixamente* ('to look intently'), *usar indevidamente* ('to use improperly'), *analisar exaustivamente* ('analyse exhaustively')

[25] http://www.collinslanguage.com/content-solutions/wordbanks [Accessed 10 May 2012]

and *subir acentuadamente* ('to rise steeply'). The fact that these translations are composed of one word only impedes an evaluation of how fluent they are in terms of association measures, reason for which they were not taken into account for the analysis. The resulting number of different class-1 bigrams used for the fluency evaluation of each system is presented in Table 6.4.

Google	Systranet	Reverso	# bigrams that overlap across the 3 systems (∩)	# total different bigrams (Δ)
54	24	22	16	67

Table 6.4 # different class-1 bigrams

The higher number of different bigrams yielded by Google Translate™ is related to the fact that this system presented a high degree of variation based on context, as seen in Table 6.3, which consequently results in a larger number of different translations for the same original bigram in Portuguese. The other two systems were more consistent in this respect, which explains their smaller number of different bigrams. As seen in Table 6.4, 16 class-1 bigrams overlap across the outputs yielded by the three systems (intersection), while 67 is the total of different class-1 bigrams (symmetric difference), considering the systems altogether.

The association measures applied to the pairs are the same ones that have been adopted in this study hitherto, namely *t* test, Chi-Square (χ^2), Log-Likelihood Ratio (LLR), Mutual Information (MI), Dice Coefficient (Dice), and Unigram Subtuples (UnigSub), whose formulas can be found in Appendix A.

Since the *t* test and χ^2 are measures that have pre-established significance threshold values, we have first checked to see how many machine-translated bigrams have reached these values for each MT system in comparison with the number of reference versions that also cross the threshold. The threshold values considered were 2.576 for the *t* test, and 3.841 for χ^2, which correspond to a confidence level of 0,005 and 0,05, respectively. Results are presented in Table 6.5.

# Bigrams	Systems	*t* test	χ^2
54	**Google**	36	54
	Ref.	*37*	54
24	**Systranet**	10	24
	Ref.	*19*	24
22	**Reverso**	11	22
	Ref.	*19*	22

Table 6.5 Bigrams that are equal to or above the *t* test and χ^2 threshold values

As seen in Table 6.5, all pairs, both those in the MT outputs and the ones deemed as reference, have reached the χ^2 critical value, which renders results of this measure inconclusive. The overly high results of the χ^2 would be connected to the tendency of this test in overemphasizing low-frequency events (Kilgarriff, 1996: 35), as already mentioned in Chapter 4. Results for the *t* test, on the other hand, show that, in the case of Systranet™ and Reverso™, a larger number of reference bigrams among the totals of 24 and 22, respectively, have reached the threshold value. In the case of Google Translate™, the number of reference bigrams that reach significance is almost the same as the machine-translated outputs. Thus, based on the *t* test threshold, these results are indicative that both the reference sources and the MT system output bigrams can be considered for the most part significantly fluent, considering the distributional data taken from the corpus.

Nevertheless, it is also desirable to look into how fluent a given machine-translated bigram is in comparison with its respective reference version. In order to carry out an evaluation of this kind, for the sake clarity and simplicity, we have considered association results in terms of the difference between values obtained for the reference versions and for their corresponding machine-translated bigrams. This difference is able to represent how distant the MT bigrams are from the reference in terms of association measure results, where a positive difference denotes a higher result for the reference, and a negative result a higher result for the MT output. The number of positive and negative difference values of each measure for the three MT systems is presented in Table 6.6. Tables with the raw values of the association measures for all the instances analysed can be found in Appendix E.

		t test	χ^2	LLR	MI (diff = 0)	Dice	UnigSub
Google (out of 54)	+	*43*	*29*	*29*	10	*37*	*34*
	-	11	25	25	*10 (34)*	14	20
Systranet (out of 24)	+	*20*	*15*	*15*	1	*19*	*16*
	-	4	9	9	*3 (20)*	5	8
Reverso (out of 22)	+	*19*	*14*	*14*	2	*18*	*15*
	-	3	8	8	*2 (18)*	4	7

Table 6.6 Number of positive and negative results in the difference between reference and class-1 MT bigrams (Ref – MT)

As it can be seen in Table 6.6, results of most measures were higher for the reference pairs than for their machine-translated counterparts, since a larger number of positive results can be observed. MI is the only measure that presents a trend in the opposite direction. Some of the MI results were the same for both the machine-translated and the reference bigrams, resulting in difference values that were equal to 0 (zero). Since the MT outputs have equalled the reference MI values in these cases, they were included in the group of negative results, which denote a high fluency of machine-translated bigrams in the corpus in terms of mutual association.

Figures shown in Table 6.6 seem to confirm that the versions provided by the reference sources tend to be more *fluent* than the MT outputs, according to distributional data of a large-sized English corpus.

From the three systems evaluated, Google Translate™ has proven to be the one that provided the most fluent output, since it was the system with the largest number of translations that fell into class 2, as it can be seen in Table 6.2. As to class 1, results in Table 6.6 seem to confirm that outputs that fell into this class tend to be non-fluent in English, with Systranet™ being the system that presented the largest number of outputs in this class. This result justifies this study in the sense that, for these particular combinations, MT systems do not conform to the subtle collocation pattern. The high number of incorrect outputs (class 0) also raises the issue of MT inaccurateness, which may not be due to the collocational nature of the combinations, but to other causes that are beyond the scope of this study.

In order to illustrate the comparison between machine-translated pairs that fell into class 1 and their respective reference versions, a selection of bigrams along with association measures is presented in Table 6.7. The original collocations in Portuguese,

whose association measure values are based on the *CETEMPúblico* corpus, are also presented, just for reference. The values of the association measures for the English word pairs are based on the *Collins WordBanks Online* corpus, as in previous examples.

	bigram	*t* test	χ^2	LLR	MI	Dice	UnigSub
Portuguese	*mentir descaradamente*	5.476	9445000	6145	17.712	0.0257	35.294
Reference	lie blatantly	2.587	*2902144*	*20155*	12.587	0.0001	36.609
MT (3 systems)	lie shamelessly (class 1)	*2.795*	1752444	12080	12.587	*0.0002*	*37.628*
Portuguese	*falar francamente*	4.859	1709068	28415	12.365	0.0005	37.304
Reference	speak earnestly	6.038	1489399	17313	11.688	0.0005	*40.753*
MT (3 systems)	speak frankly (class 1)	*11.250*	*9119156*	*110164*	11.688	*0.0018*	40.122
Portuguese	*chorar convulsivamente*	5.291	1642101	1175	16.958	0.0155	39.388
Reference	cry hysterically	5.472	3393737	10321	13.911	0.0020	39.246
	cry uncontrollably	*6.474*	*4663791*	*14246*	13.911	*0.0027*	*39.363*
MT (Google)	weep convulsively (class 1)	0.998	3564152	3029	*16.026*	0.0002	31.449

Table 6.7 Comparison between reference and machine-translated pairs

In regard to the pair *falar francamente* ('speak frankly'), the MT outputs have also presented higher association measure results than the reference. It is interesting to notice about this pair that, even though *speak frankly* does not figure as a collocation in the *Oxford Collocations Dictionary*, this version has proven to be more fluent than the reference according to five out of six association measures. The contrary can be observed for the pair *chorar convulsivamente* ('cry convulsively'). Possible English equivalents for this pair found in the *Oxford Collocations Dictionary* are *cry hysterically* and *cry uncontrollably*, while the outputs yielded by the MT systems are *cry convulsively* for both Reverso™ and Systranet™, and *weep convulsively* for Google Translate™. In this case, the reference equivalents presented higher results than the MT outputs. In fact, the output yielded by both Systranet™ and Reverso™ for this pair, namely *cry convulsively*, does not even occur in the *Collins WordBanks Online* corpus, reason for which it does not figure in Table 6.7. With respect to the pair *weep convulsively*, the reference equivalents have presented higher results for the majority of association measures, MI being the only exception in that respect.

Overall, it could be affirmed that the pattern {V, *Adv-mente*} poses considerable difficulty to MT, since 24.1% of the MT outputs evaluated were considered to fall into class 0 – characterised by missing or erroneous translations –, as it can be seen in Table 6.2. With regard to the other two classes, outputs that fell into class 1, albeit not erroneous, have shown to be considerably less fluent than those that fell into class 2. This is noteworthy since 62.1% of cases were considered to be of class 1, which leads to the conclusion that while the MT outputs of {V, *Adv-mente*} in the direction Portuguese-English can be considered non-erroneous in most cases, the majority of MT outputs are not fluent.

These results confirm the validity of the type of research conducted in this study: verb-adverb collocations pose difficulties to high quality MT. If fluency is a goal, then, ideally, other collocational patterns should be further analysed.

Chapter 7. Conclusions and Future Work

This dissertation aimed at investigating verb and *–mente* adverb collocations in Portuguese (e.g. *mentir descaradamente*, 'lie shamelessly') in view of their extraction from corpora and their automatic translation into English.

It has been shown that the very concept of collocation is far from being clear-cut and still poses a number of problems to a precise definition and classification of this phenomenon, incurring in a wide range of formulations and approaches that can be established to address this topic. In the case of this study, the notion of collocation adopted is one that profits both from frequency of distribution (Firth, 1957) and from linguistic-based formulations (Mel'čuk, 2003; 2010), since at different stages both frequency and syntactic-semantic principles are considered for their extraction and classification.

The extraction method utilised was based on the processing of a large-sized corpus of Portuguese, including the syntactical analysis of the text. In that way, the process of retrieving verb-adverb pairs from the corpus was not based merely on the co-occurrence of the terms within a given window of words, but rather on the existence of a syntactic dependency between the terms that composed the combination. For that purpose, a number of measures had to be taken so as to guarantee that the coverage of the pairs in the corpus was as large and yet as precise as possible. Since certain *Adv-mente* classes are known to present little or no direct connection to the verb in a clause, a syntactic-semantic classification of *Adv-mente* in Portuguese (Fernandes, 2011) was substantially augmented and adopted as a criterion to filter out cases that presented no collocational potential given their lack of a straight connection with the verb. This classification was originally based on the description of French *Adv-ment* formulated by Molinier and Levrier (2000), and was later incorporated in the text processing chain used to parse the corpus.

For processing the *CETEMPúblico* corpus, to the best of our knowledge the largest corpus publicly available of Portuguese, the STRING processing chain (Mamede et al., 2012) was used. The chain includes tokenisation, morphological analysis, POS tagging, and syntactical parsing, which is performed by the XIP (Xerox Incremental Parser) finite-state rule-based parser (Aït Mokhta et al., 2002).

The syntactic behaviour of *Adv-mente* in Portuguese presents a number of peculiarities that represented a problem for the computational processing of {V, *Adv-mente*} pairs in view of the collocation extraction task. In that way, besides incorporating the linguistic classification of this adverb type in the text processing chain, the phenomenon of adverb coordination and reduction in Portuguese was also addressed. *Adv-mente* in Portuguese can be used in coordinated chains that hold a syntactic dependency with a single verb. When coordinated, all but the last adverb in the combination lose the *–mente* ('ly') suffix and take the shape of the feminine base adjective to which they are associated, posing a substantial problem to the POS disambiguation of the terms and also their dependency extraction. To address this problem, a number of disambiguating, chunking and parsing rules have been incorporated in the STRING system. Results obtained were considerably promising for the dependency extraction task, with an *F*-measure of 0.81. For POS disambiguation, an *F*-measure of 0.72 was obtained. It could be said that these results reflect the degree of difficulty posed by the problem. *Adv-mente* is an adverb class that involves a number of particularities of its own, which validates initiatives such as the one here undertaken of improving the computational processing of constructions involving this type of adverb.

The extraction of the verb-adverb pairs from the corpus yielded an output of approximately 65K word combinations, which passed through a number of filtering stages, resulting in a set of collocation candidates composed of 5,973 bigrams. A list of semantic-syntactic criteria was then devised for the classification of the 5,973 bigrams as (non-)collocations. The classification has been manually carried out.

In order to assess the intuition of native speakers of Portuguese on the collocational value of the pairs classified, a sample of 30 bigrams was randomly selected from the entire set and given to 21 subjects native speakers of Portuguese for classification. Despite the fact that an explanation of the linguistic criteria that should be used in the classification was provided, results have shown that the task of annotating word pairs as collocations is extremely challenging, which would be related to very elusive concept of collocations. Results of this experiment have shown that 62%of the subjects have agreed on the classification of cases that had previously been tagged as colocations. The overall agreement stood at 57%, with a κ value of 0.06, which could be considered to be in the range of slight agreement according to the

interpretation scale proposed by Landis and Koch (1977). That suggests that identifying a non-collocation poses a considerably higher dose of difficulty as opposed to identifying collocations proper.

It was also possible to conclude from this experiment that an annotation task of this kind aimed at the identification of collocations of the type studied requires substantial training of annotators, and would be unlikely to reach more significant agreement levels just with the provision of criteria that should be taken into account for the classification.

After classifying all candidate pairs based on the set of criteria previously established, we have checked to see how sensitive different statistical association measures were in capturing the collocational status of the pairs. The association measures used in the experiment were *t* test, Pearson's chi-square (χ^2), Mutual Information (MI), Log-likelihood Ratio (LLR), Dice Coefficient (Dice), and Unigram Subtuples (UnigSub). As to the *t* test and χ^2, it has been shown that the pre-established statistical significance threshold values of these measures are not able to satisfactorily capture the collocation pattern investigated.

The correlation of all measures with the classification has also been assessed in terms of Person's *r*. Results have shown that while the *t* test and χ^2 had were poorly correlated with the classified pairs, the other measures were more satisfactory in this respect, with UnigSub, LLR and MI presenting rather promising correlation values. It has also been shown that most measures tend to be more sensitive to highly frequent events, with LLR being the measure with the most significant correlation with cases of collocation that had low frequency in the corpus, which confirms previous studies that claim the appropriateness of this measure for the collocation extraction task.

Based on results of the association measures, we have also attempted to train an automatic collocation classifier for the linguistic pattern {V, *Adv-mente*} using Machine Learning techniques. The WEKA toolkit was adopted for that purpose where the performance of a number of different classifiers available in the toolkit has been tested. In an experiment with a balanced training set, we have noticed that RotationForest, a Meta classifier, has presented the most interesting results. However, the strategy that in fact has proven most effective was to combine different decision algorithms through the Vote Meta classifier. The combination has outperformed all the

classifiers isolated, which renders this strategy extremely promising for the task of collocation classification.

In a comparison of the performance of the combined classifier with the performance of the human subjects recruited for the annotation task previously described, the automatic classifier has achieved a considerably higher precision on the classification of cases that had been deemed to be collocations. Results were also good for a small set of unseen collocation candidates extracted from another corpus of Portuguese. Even though the evaluation carried out with unseen data was not able to lead to decisive conclusions given the small scale of the set, it can be argued that the good result is a sign that automatic classifiers represent a promising alternative to be further exploited for the task of classifying collocations.

The last stage of the project consisted of the compilation of a Portuguese-English lexicon of the collocation pattern studied and evaluation of the automatic translation of the word pairs.

For the process of compiling the lexicon of equivalent combinations in English, a collocation dictionary and three Portuguese-English parallel corpora were used as source of reference bigrams. However, the dictionary was given preference over the corpora because it would be able to provide more than just equivalent combinations, but equivalent pairs that are in fact deemed as *collocations* in the target language.

Based on the equivalent pairs in the lexicon, the evaluation of MT systems was carried out having the word pairs found in the dictionary or in the parallel corpora as reference translations. Three commercial MT engines available online were selected for the evaluation, namely Google Translate™, Systranet™ and Reverso™. The experiment was restricted to cases that were considered to pose more substantial difficulty to MT, which was based on the criteria of how morphologically/etymologically different the translations were from the original. Results have shown that while the automatic translation of {V, *Adv-mente*} collocations is accurate in the majority of cases, the outputs yielded tend not to conform to the collocation pattern in the target language. This has been demonstrated based on a comparison of association measure results for the machine-translated bigrams and their respective reference versions having a large-sized corpus of English as the source of distributional data. It follows that the MT engines evaluated were not

able to satisfactorily comply with the principle of *fluent output* – formulated by Koehn (2010) – when dealing with *Adv-mente* collocations

Results of the project as a whole demonstrate that the {V, *adv-mente*} pattern poses a considerable degree of difficulty to NLP tasks in general, from syntactical parsing, dependency extraction and POS disambiguation to MT. It is hoped that the outputs of the research undertaken have been able not only to cast light on these issues but also to contribute to their solving by resulting in a better quality of the processing of this pattern. It is also hoped that the lexicon produced is able to serve as a source of collocational information on a pattern in Portuguese that could have been considered understudied in Portuguese hitherto. Further to NLP applications, having this kind of word combinatorial knowledge is important for a number of related areas such as Linguistics and Foreign Language Learning.

As future work, the lexicon compiled could be further extended with data from other corpora of Portuguese. The collocation classifier that has been built would ideally have to be more extensively tested. Other MT engines should also be further tested as to their performance in translating *Adv-mente* collocations, ideally drawing a parallel between statistical and rule-based systems, which have shown considerable difference in the experiment carried out in this study in terms of how varied the outputs were based on the context in which the verb-adverb pairs were inserted.

Finally, it could be said that the elusive nature of collocations makes this an extremely challenging topic to deal with, especially when an equally elusive and heterogeneous grammatical class such as adverbs, and *Adv-mente* specifically, is involved. In that way, it is paramount that these issues continue to be a target of research so that we are able to better understand and make due use of them to enhance NLP applications and theoretical methodologies alike.

References

Afonso, S. (2003) *Clara e sucintamente*: um estudo em corpus sobre a coordenação de advérbios em *-mente*. In: Amália Mendes and Tiago Freitas (eds.), *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)* (Porto, Portugal, 2-4 de Outubro de 2002), Lisboa: APL, pp. 27-36.

Aït Mokhtar, S., Chanod, J. and Roux, C. (2002) Robustness Beyond Shallowness: Incremental Deep Parsing. *Natural Language Engineering*, 8, pp. 121–144. New York: Cambridge University Press.

Baptista, J. (2010) Classification of Portuguese Verbs - Guidelines. (Tech. rep.) L2F/INESC-ID Lisboa.

Baptista, J. and Català, D. (2009) Disambiguation of Focus Adverbs in Portuguese and Spanish. In: *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages*. PUFC, pp. 31-37.

Baptista, J. and Català, D. (in press) What Glues Idioms Together May Not Be Statistics After All. The Case for Compound Adverbs in Portuguese and Spanish. In: *Proceedings of Europhras*, 2010. Granada. (preprint)

Bechara, E. 1999 (2003) *Moderna gramática portuguesa*. (37a ed.) Rio de Janeiro: Editora Lucerna.

Bernth, A. and Gdaniec, C. (2000) *MTranslatability* AMTA-2000 Tutorial. Available at <http://www.isi.edu/natural-language/organizations/amta/sig-mtranslatability-tutorial.htm> [Accessed 8 May 2012].

Berry-Rogghe, G. (1973) The Computation of Collocations and their Relevance to Lexical Studies. In: Aitken, A. J., Bailey, R. W. and Hamilton-Smith, N., (eds.) *The Computer and Literary Studies*. Edinburgh University Press, Edinburgh, , pp. 103–112.

Blaheta, D. and Johnson, M. (2001) Unsupervised Learning of Multi-Word Verbs. In: *Proceedings of the 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL39)*, CNRS - Institut de Recherche en Informatique de Toulouse, and Universite´ des Sciences Sociales, Toulouse, France, July, pp. 54–60.

Cardey, S., Chan, R. and Greenfield, P. (2006). The Development of a Multilingual Collocation Dictionary. In: Proceedings of the Workshop on Multilingual Language Resources and Interoperability, Sydney, July 2006, pp. 32–39.

Casteleiro, J. (2009) *Vocabulário ortográfico da língua portuguesa*. Porto: Porto Editora.

Celso, C. and Lindley, C. (1984) *Nova gramática do português contemporâneo*. Lisboa: João Sá da Costa.

Choueka, Y. (1988) Looking for Needles in a Haytsack or Locating Interesting Collocational Expressions in a Large Corpus. In: *Proceedings of Recherche d'Information Assistée par Organiteur - RIAO '88*, pp. 609–623.

Church, K. and Gale, W. (1991) Concordances for parallel text. In: *Proceedings of the Seventh annual aonference of the UW centre for the new OED and text research,* Oxford, UK, pp. 40-62.

Church, K. and Hanks, P. (1989) Word Association, Norms, Mutual Information, and Lexicography. In: *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pp. 76–83.

Church, K. and Mercer, R. (1993) Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics*, *19*(1), pp. 1-24.

Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, pp. 37-46.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Costa-jussà, M., Daudaravicius, V. and Banchs, R. (2010) Integration of Statistical Collocation Segmentations in a Phrase-Based Statistical Machine Translation System. In: *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*.

Costa, J. (2008) *O Advérbio em português europeu*. Lisboa: Edições Colibri.

Daille, B. (1994) *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales e filtres linguistiques*. Ph.D. Université Paris 7.

Daudaravicius, V., 2009 (2010) Automatic Identification of Lexical Units. *Informatica,* 34, pp. 85-91.

Dice, L. (1945) Measures of the Amount of Ecologic Association between Species. *Ecology*, *26*(3), 297-302.

Diniz, C. (2010) RUDRICO2 – *Um Conversor Baseado em Regras de Transformação Declarativas*. M.Sc. Thesis, Instituto Superior Técnico/Universidade Técnica de Lisboa.

Diniz, C. and Mamede, N. (2011) LEXMAN - Lexical Morphological Analyser, (Tech. rep.), Lisboa L2F/INESC-ID Lisboa.

Dunning, T. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1), pp.61-74.

Evert, S. (2005) *The Statistics of Word Cooccurrences. Word Pairs and Collocations*. Ph.D. Thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Evert, S. (2008). Corpora and Collocations. In: A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, pp. 1212–1248. Mouton de Gruyter, Berlin.

Fano, R. (1961) *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA.

Fernandes, G. (2011) *Classification and Word Sense Disambiguation: The case of -mente Ending Adverbs in Brazilian Portuguese*. M.A. Thesis, Universidade do Algarve/ Universitat Autònoma de Barcelona.

Firth, J. (1935) The Technique of Semantics. *Transactions of the Philological Society*, *34*(1), pp. 36-73.

Firth, J. (1957) A Synopsis of Linguistic Theory 1930-55. In: Firth, J. R. et al. *Studies in Linguistic Analysis*. Special volume of the Philological Society. Oxford: Blackwell.

Firth, J. (1969) *Papers in Linguistics 1934-51*. Oxford: Oxford University Press.

Fisher, R. (1925) Applications of "Student's" Distribution. *Metron*, 5, pp. 90-110.

Forcada, M. (2000). Learning Machine Translation Strategies Using Commercial Systems: Discovering Wordreordering Rules. In: *Proceedings of MT2000: Machine Translation and Multilingual Applications in the New Millenium*. Exeter, UK, November 18-20, 2000, pp. 7.1-7.8.

Frankenberg-Garcia, A. and Santos, D. (2002) COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução* IX(1), pp. 61-79.

Gdaniec, C. (1994). The Logos Translatability Index. In: *Technology Partnerships for Crossing the Language Barrier. Proceedings of the First Conference of the Association for Machine Translation in the Americas* AMTA, pp. 97-105.

Gelbukh, A. and Kolesnikova, O. (2010). Supervised Learning for Semantic Classification of Spanish Collocations. In: *Proceedings of the 2nd Mexican Conference on Pattern Recognition: Advances in Pattern Recognition* (MCPR'10), Jesús Ariel Carrasco-Ochoa, José Francisco Martínez-Trinidad, and Josef Kittler (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 362-371.

Gomes, F. (2009) *Validation of Lexical-Syntactical Matrices*. M.Sc. Thesis, Instituto Superior Técnico/Universidade Técnica de Lisboa.

Gross, M. (1981) Les bases empiriques de la notion de prédicat sémantique. *Langages*, 63, pp. 7–52.

Gross, M. (1986) *Grammaire transformationelle du français. 3 - syntaxe de l'adverbe*. Paris: ASSTRIL.

Greenbaum, S. (1970) *Verb-intensifier Collocations in English. An experimental Approach*. The Hague: Mouton.

Hagège, C., Baptista, J. and Mamede, N. (2010) Caracterização e processamento de expressões temporais em português. *Linguamática* 2(1), pp. 63–76.

HarperCollins (2009) *WordBanks Online: English Corpus*. HarperCollins Publishers Ltd. Available at: <http://www.collinslanguage.com/content-solutions/wordbanks> [Accessed 10 May 2012].

Jespersen, O. (1965) *A modern English Grammar on Historical Principles*, Part VI, Morphology. London: George Allen and Unwin Ltd.

Kilgarriff, A. (1996) Which Words are Particularly Characteristic of a Text? A Survey of Statistical Approaches. In: *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition*, Sussex, UK, pp. 33-40.

Kita, K., Kato, Y., Omoto, T. and Yano, Y. (1994) A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual information vs. Cost criteria. *Journal of Natural Language Processing*, *1*(1), 21-33.

Koehn, P. (2010) *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Zens, C., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. (2007) Moses: Open Source Toolkit for Statistical Machine Translation. In: *Proceedings of the 45th Annual Meeting of the ACL, Poster and Demonstration Sessions*, pp. 177- 180.

Koehn, P., Och, F. and Marcu, D. (2003) Statistical Phrase-Based Translation. In: *Proceedings of the Human Language Technology Conference and the North American Association for Computational Linguistics*, pp. 127-133.

Krenn, B. (2000) *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*. Ph.D. Thesis, Saarland University.

Landis, J. and Koch, G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, pp. 159-174.

Levin, B. (1993) *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Liu, Z., Wang, H., Wu, H. and Li, S. (2010) Improving Statistical Machine Translation with Monolingual Collocation. In: *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, Sweden, Uppsala, pp. 825–833.

Mamede, N. (2011) STRING - A cadeia de processamento de língua natural do L2F. [PowerPoint Slides] Presented at *NILC/ICMC/USP*, São Carlos, Brasil, February 17.

Macmillan (2010) *Macmillan Collocations Dictionary Book*. Oxford: Macmillan Publishers.

Mamede, N., Baptista, J., Diniz, C. and Cabarrão, V. (2012). STRING: A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. Demo PROPOR 2012 Available at <http://www.propor2012.org/demos/DemoSTRING.pdf> [Accessed 14 May 2012].

Manning, C. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Matthews, P. (2007) *Oxford Concise Dictionary of Linguistics* (2nd ed.). Oxford: Oxford University Press.

McKeown, K. and Radev, D. (2000). Collocations. In: Dale, R., Moisl, H. and Somers, H. (eds.) *Handbook of Natural Language Processing*. New York: Marcel Dekker, pp. 507-523.

Mel'čuk, I. (1982) Lexical Functions in Lexicographic Description. In: *Proceedings of the Eighth Annual Meeting of the Berkeley Linguistics Society*, 13-15 February, 1982, Berkeley Linguistics Society - University of California, Berkeley, pp. 427-444.

Mel'čuk, I. (2003) Collocations, definition, rôle et utilité. In: Grossmann and Tutin (eds.), *Les collocations, analyse et traitement,* Amsterdam: De Werelt, pp. 23-31.

Mel'čuk, I. (2010) La phraséologie en langue, en dictionnaire et en TALN. In: *Proceedings of TALN 2010*, July 9, Montreal, Canada. Available at <http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_207.pdf> [Accessed 14 May 2012].

Molinier, C. and Levrier, F. (2000) *Grammaire des adverbes. Description des formes en* -ment. Genève: Droz.

Nobre, N. (2011) *Anaphora Resolution*. M.Sc. Thesis, Instituto Superior Técnico/Universidade Técnica de Lisboa.

Och, F. (2006) Statistical Machine Translation live. *Research Blog – The Latest News on Google Research*, [blog], 29 April. Available at <http://googleresearch.blogspot.fr/2006/04/statistical-machine-translation-live.html> [Accessed 8 May 2012].

Oliveira, D. (2010) *Extraction and Classification of Named Entities*. M.Sc. Instituto Superior Técnico/Universidade Técnica de Lisboa.

Orliac, B. and Dillinger, M. (2003) Collocation Extraction for Machine Translation. In: *Proceedings of the Machine Translation Summit IX*. New Orleans, USA, 23-27 September 2003, pp.292-298.

Orliac, B. (2006) Un outil d'extraction de collocations spécialisées basé sur les fonctions lexicales. *Terminology* 12(2), pp. 261-280.

Oxford (2009) *Oxford Collocations Dictionary for Students of English* (2nd ed.). Oxford: Oxford University Press.

Pearce, D. (2001) Synonymy in Collocation Extraction. In: *Proceesings of the NAACL 2001 Workshop: WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Carnegie Mellon University, Pittsburgh, June, pp. 41-46.

Pearce, D. (2002) A Comparative Evaluation of Collocation Extraction Techniques. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*. Spain, Las Palmas.

Pearson, K. (1896) Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society of London*, 187, pp. 253-318.

Pearson, K. (1900) On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it Can Be Reasonably Supposed to have Arisen from Random Sampling. *Philosophical Magazine, Series 5* 50(302), pp. 157–175.

Pecina, P. and Schlesinger, P. (2006) Combining Association Measures for Collocation Extraction. In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Sydney, pp. 651–658.

Pecina, P. (2010) Lexical Association Measures and Collocation Extraction. *Language Resources and Evaluation*, *44*(1), pp. 137-158.

Pinheiro, G. and Aluísio, S. M. (2003) *Córpus NILC: descrição e análise crítica com vistas ao projeto Lacio-Web*. [pdf] São Carlos: NILC-TR-03-03. Available at: <http://www.linguateca.pt/> [Accessed 28 March 2012].

Platt, J. (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Schölkopf, B., Burges, C. and Smola, A. (eds.). *Advances in Kernel Methods - Support Vector Learning*. Cambridge, MA: MIT Press, pp. 185-208.

Portela, R. (2011) *Identificação automática de nomes compostos*. M.Sc. Thesis, Instituto Superior Técnico/Universidade Técnica de Lisboa.

Quirk, R. Greenbaum, S. Leech, G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*. New York: Longman.

Ribeiro, R. (2003) *Anotação morfossintáctica desambiguada em português*, M.Sc. Thesis, Instituto Superior Tecnico/Universidade Técnica de Lisboa.

Roukos, S., Graff, D. and Melamed, D. (1995) *Hansard French/English*. Linguistic Data Consortium, Philadelphia.

Santos, D. (2010) *Extração de relações entre entidades mencionadas*. M.Sc. Thesis, Instituto Superior Técnico/Universidade Técnica de Lisboa.

Santos, D. and Rocha, P. (2001). Evaluating CETEMPúblico, a free resource for Portuguese. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics,*Toulouse, 9-11 de Julho de 2001, pp. 442-449.

Schenk, A. (1986) Idioms in the Rosetta Machine Translation System. In: *Proceedings of COLING 86*, 1986, pp. 319–324.

Seretan, V. (2011) *Syntax-Based Collocation Extraction*. *Text, Speech and Language Technology*. Dordrecht: Springer.

Shimohata, S., Sugio, T. and Nagata, J. (1997) Retrieving Collocations by Co-occurrences and Word Order Constraints. In: *Proceedings of 35th Conference of the Association for Computational Linguistics (ACL'97)*, pp. 476–481, Madrid, Spain.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Smadja, F. (1993) Retrieving Collocations from Text: Xtract. *Computational Linguistics*, *19*(1), pp. 143-177.

Smadja, F., McKeown, K. and Hatzivassiloglou, V. (1996). Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, *22*(1), pp. 1-38.

Sörensen, T. (1948) A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species and its Application to Analysis of the Vegetation of Danish Commons. *Biologiske Skrifter*, *5(4)*, pp. 1-34.

Systran (2012). *About SYSTRAN*. [Online]. Available at <http://www.systran.co.uk/systran> [Accessed 11 May 2012].

Tagnin, S. (2010) CorTrad – Um corpus de traduções inglês-português online com mil e uma possibilidades de pesquisa. [PowerPoint Slides] Presented at *Encontro de Férias da SBS 2010*. Available at: <http://www.fflch.usp.br/dlm/comet/artigos/SBS_Encontro_de_ferias_2010.pdf> [Accessed 5 May 2012].

Tsujii, J. (1986) Future Directions of Machine Translation. In: *Proceedings of the 11th International Conference on Computational Linguistics*. Bonn, pp. 655-668.

Underwood, N. and Jongejan, B. (2001). Translatability Checker: a Tool to Help Decide Whether to use MT. In *MT Summit VIII Machine Translation in the Information Age Proceedings* Santiago de Compostela, Spain 1822 September 2001. pp. 363-368.

Xu, X. (2003) *Statistical Learning in Multiple Instance Problems*. M.Sc. Thesis, The University of Waikato.

Wagner, E. (1985) Rapid Post-Editing of Systran. In: Lawson, V. (ed.) *Tools for the Trade, Translating and the Computer* 5, Alden Press, Oxford, pp. 199-213.

Way, A. and Gough, N. (2003). wEBMT: Developing and Validating an Example-Based Machine Translation System Using the World Wide Web. In: *Computational Linguistics 29*(3), pp. 421-457.

Witten, I., Frank, E. and Hall, M. (2011) *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.) Morgan Kaufmann, Burlington, MA.

Appendix A. Formulas of Statistical Association Measures

Student's *t* test

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

Where $\bar{x}$ is the sample mean, s^2 is the sample variance, N the sample size, and μ the mean of the distribution (Manning and Schütze, 1999).

Chi Square (X^2)

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where i ranges over rows of the contingency table, j ranges over columns, O_{ij} is the observed value for cell (i, j) and E_{ij} is the expected value (Manning and Schütze, 1999).

Mutual Information (MI)

$$I(x', y') = \log_2 \frac{P(x'y')}{P(x')P(y')}$$

Where x' y' would be the events between which the mutual information is calculated (Manning and Schütze, 1999).

Log-Likelihood Ratio (LLR)

$$LLR = -2\log\lambda = -2\log\frac{L(H_0)}{L(H_1)}$$

$$= 2(a\log a + b\log b + c\log c + d\log d$$
$$-(a+b)\log(a+b) - (a+c)\log(a+c)$$
$$-(b+d)\log(b+d) - (c+d)\log(c+d)$$
$$+(a+b+c+d)\ \log(a+b+c+d))$$

Where H_0 is the null hypothesis and H_1 is the alternative hypothesis; a, b, c, and d represent the cells in the contingency table (Seretan, 2011).

Dice Coefficient

$$Dice = \frac{2a}{R_1 + C_1}$$

Where R_1 corresponds to the number of segments containing an instance of word1, and C_1 corresponds to the number of segments containing an instance of word2.

Unigram subtuples

$$\log\frac{ad}{bc} - 3.29\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Where a, b, c, and d represent the cells in the contingency table (Pecina, 2010).

Contingency Table

$a = f\ (xy\ )$	$b = f\ (x\bar{y}\ )$	R_1
$c = f\ (\bar{x}y\ )$	$d = f\ (\bar{x}\bar{y}\ )$	R_{12}
C_1	C_2	N

Appendix B. Annotation Task

21 responses

Summary See complete responses

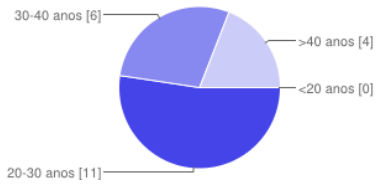
Instruções

Muito obrigado por colaborar neste questionário. Sua colaboração é muito importante para nós. Este questionário deverá levar cerca de 25 minutos a responder. O objetivo deste questionário é testar as intuições linguísticas de falantes nativos de Português relativamente ao carácter de colocação de certas combinações verbo-advérbio em frases reais. Não é preciso saber gramática, é a sua intuição quanto a estas combinações que nos interessa conhecer, portanto não há nem respostas certas nem respostas erradas. Leia atentamente as frases seguintes e, com base na sua intuição de falante nativo, indique para cada combinação verbo-advérbio (<<<marcadas desta maneira>>>) se considera que se trata ou não de uma colocação. Para perceber melhor o que se entende por colocação leia primeiro os seguintes exemplos antes de começar a responder. 1. O advérbio denota uma hipérbole (exagero), uma ironia ou um paradoxo: O Pedro adiava eternamente a resposta ao pedido do Rui. 2. O advérbio tem valor não literal na combinação: O time/a equipa venceu confortavelmente a partida. (diferente de : O time/a equipa estava confortável) Ele apaixonou-se perdidamente por ela (diferente de: Ele estava perdido) 3. A combinação faz parte de um vocabulário científico ou técnico (exemplo do domínio jurídico): Ele respondeu civilmente pelo crime que cometeu. 4. Um advérbio sinônimo/sinónimo em outros contextos deixa de o ser nesta combinatória: Ela chorava copiosamente. cp. ??*Ela chorava abundantemente. (a frase é inaceitável/incorrecta ou muito duvidosa) 5. O advérbio não se combina com o antônimo/antónimo do verbo, quando este existe:: O time/a equipa venceu confortavelmente a partida. *O time/a equipa perdeu confortavelmente a partida. (a frase é inaceitável/incorrecta) 6. O advérbio se combina com apenas um dos significados que o verbo pode ter: A secretária reproduziu fielmente os documentos. *Os coelhos reproduzem-se fielmente. (a frase é inaceitável/incorrecta) 7. Se existirem verbos de significado semelhante que se combinam com o mesmo advérbio, o status de colocação também se aplica às demais combinatórias (repare que este critério só é válido para casos em que os verbos têm significados parecidos): A professora criticou duramente o aluno. A professora reprimiu duramente o aluno. O acidente feriu gravemente os passageiros. O tiro o atingiu gravemente. O tombo lesou gravemente seu tendão direito.

Dados pessoais

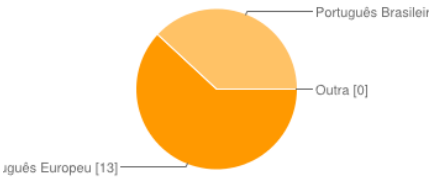
Este questionário é completamente anónimo. Os dados aqui recolhidos destinam-se exclusivamente a processamento estatístico e não serão transmitidos a terceiros.

Idade



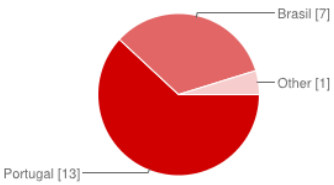
<20 anos	**0**	0%
20-30 anos	**11**	52%
30-40 anos	**6**	29%
>40 anos	**4**	19%

Língua materna



Português Europeu	**13**	62%
Português Brasileiro	**8**	38%
Outra	**0**	0%

País de residência



Portugal	**13**	62%
Brasil	**7**	33%
Other	**1**	5%

Profissão

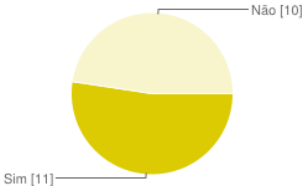
Linguista Computacional Professor de língua inglesa Professor professora estudante Estudante Estudante Professora Estudante Professor Estudante Professora professor Formador/Professor Funcionário publico Linguísta ...

Contacto

Os pares de <<<verbo-advérbio>>> nas frases seguintes são colocações?

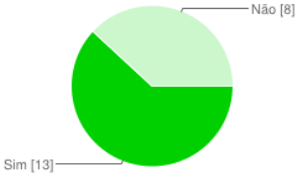
Marque "sim" ou "não" com base nos exemplos acima e na sua intuição de falante nativo.

1. O estado de espírito era o de que, custasse a habitação 40 mil ou 90 mil contos, o que importava era <<<adquirir imediatamente>>>, já que a convicção dominante levava a que se pensasse que, daí a pouco tempo, existiria uma valorização significativa do investimento .



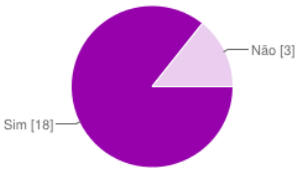
Sim	**11**	52%
Não	**10**	48%

2. Finalmente, evite problemas futuros exigindo um contrato por escrito, cujas cláusulas deverão ser <<<lidas atentamente>>> .



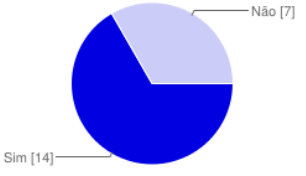
Sim	**13**	62%
Não	**8**	38%

3. Só que essa escolta não iria <<<esperar>>> por nós <<<<eternamente>>>, e nós já estávamos com dois dias de atraso (um dia perdido em Tamaransset devido ao telefone de satélite, outro desperdiçado por causa do tal funcionário fronteiriço) .



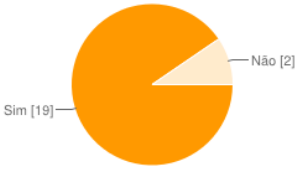
Sim	**18**	86%
Não	**3**	14%

4. Durante as filmagens de «O Corvo» -- uma história criada por James O'Barr em banda desenhada e transposta para o cinema pela mão do realizador australiano Alex Proyas --, o actor Brandon Lee foi <<<atingido acidentalmente>>> por um tiro e morreu .



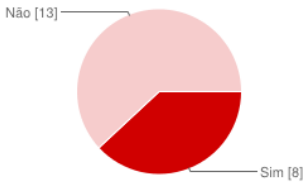
Sim	**14**	67%
Não	**7**	33%

5. Os polacos <<<contestaram fortemente>>> a arbitragem de Remy Harrel por causa da expulsão e também por terem visto um golo anulado por fora de jogo .



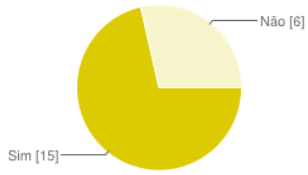
Sim	**19**	90%
Não	**2**	10%

6. O trabalho desses funcionários será aplicado no reforço de certas áreas de prestação de serviços agora deficitárias, na limpeza da Quinta da Conceição, por exemplo, e no tratamento de ecocentros e ecopontos, a <<<instalar brevemente>>> .



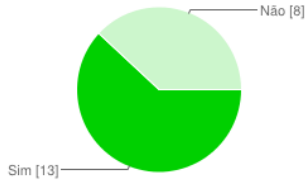
Sim	**8**	38%
Não	**13**	62%

7. No primeiro jogo, Seabra esteve sempre a dominar, tendo <<<chegado facilmente>>> ao 7-4 .



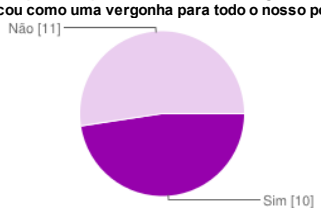
Sim	**15**	71%
Não	**6**	29%

8. É, portanto, fácil compreender que nesta genealogia houve um instante (geológico) fundamental: foi quando a nossa linhagem se <<<separou definitivamente>>> daquele reino .



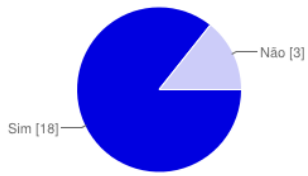
Sim	**13**	62%
Não	**8**	38%

9. Logo suspeita, a oposição islamista <<<negou imediatamente>>> qualquer responsabilidade, condenando categoricament o acto terrorista, que classificou como uma vergonha para todo o nosso povo .



Sim	**10**	48%
Não	**11**	52%

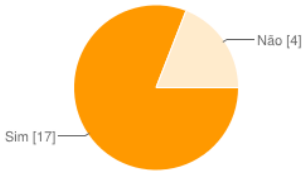
10. O que há de mais característico, em todas estas áreas, é que o respeito pelos direitos das pessoas <<<depende crucialmente>>> da condição social delas .



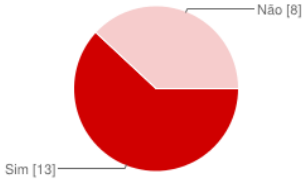
Sim	**18**	86%
Não	**3**	14%

11. O primeiro daqueles organismos <<<respondia afirmativamente>>>, enquanto o segundo se pronunciava em sentido contrário .

Sim	**17**	81%
Não	**4**	19%

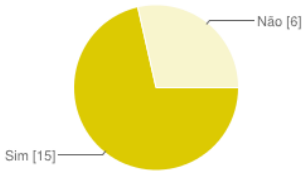


12. Há contactos em curso entre o comando das forças italianas e a Operação das Nações Unidas na Somália (Onusom) para se <<<decidir conjuntamente>>> quais as tarefas do contingente na fase de transição, bem como a duração desta -- disse o ministério .



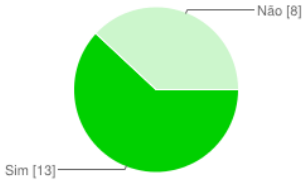
Sim	**13**	62%
Não	**8**	38%

13. O Ministério das Finanças é o cobrador-mor deste Governo e <<<desvirtuou completamente>>>, por exemplo, a recuperação das empresas .



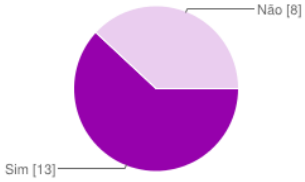
Sim	**15**	71%
Não	**6**	29%

14. Por enquanto, nenhum deles dá sinal de recuar, o que leva a crer que a tensão política <<<aumentará bruscamente>>> nos tempos mais próximos .



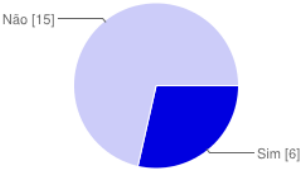
Sim	**13**	62%
Não	**8**	38%

15. Frequentemente, estes entraram irregularmente no país, não têm documentação (ou falsificaram-na) , ultrapassaram os prazos para requererem asilo, desconhecem a língua e não têm acesso a intérpretes, ou não conseguem <<<comprovar documentalmente>>> as suas razões .



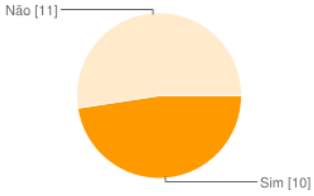
Sim	**13**	62%
Não	**8**	38%

16. A sessão de ontem da Bolsa de Zurique não produziu qualquer tipo de consequências, com os preços a caírem ligeiramente, seguindo os negócios que <<<diminuíram comparativamente>>> ao dia anterior .



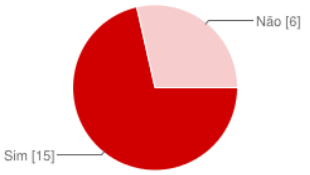
Sim	**6**	29%
Não	**15**	71%

17. A empresa nacional de navegação aérea <<<disse igualmente>>> não ter qualquer informação, recordando que em Angola não há controlo por radar .



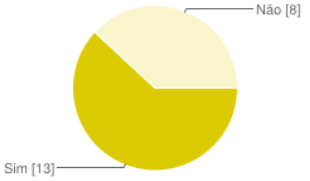
Sim	**10**	48%
Não	**11**	52%

18. Eu <<<falei francamente>>>, apesar de estar a apostar a minha vida .



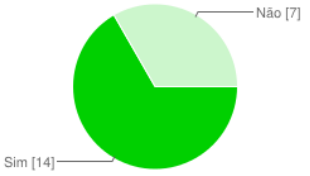
Sim	**15**	71%
Não	**6**	29%

19. Nesta entrevista <<<responde, cautelosamente>>>, às críticas e fala, empolgado, dos projectos que tem para o porto .



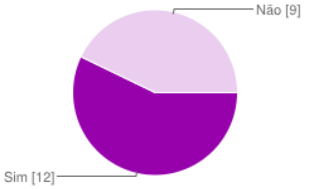
Sim	**13**	62%
Não	**8**	38%

20. Em Colónia estiveram inundados 35 hectares e a culpa foi do alter Vater Rhein -- o paizinho Reno, como os renanos <<<chamam carinhosamente>>> ao seu rio, que, em vez de se ficar pelo seu leito, resolveu visitar outras paragens .



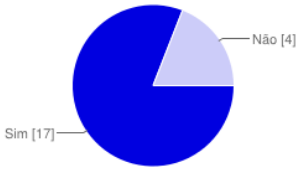
Sim	**14**	67%
Não	**7**	33%

21. Se alguma coisa correr mal <<<accionará imediatamente>>> três enormes botões vermelhos que cortam a corrente, baixam a catenária e frenam a composição no mais curto espaço de tempo possível .



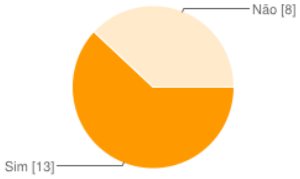
Sim	**12**	57%
Não	**9**	43%

22. Trata-se de uma posição que <<<obedece estritamente>>> a um critério económico, antipedagógico, e significa um retrocesso nítido no apoio a estes alunos», protesta José Lucas, dirigente do Sindicato dos Professores da Região Centro .



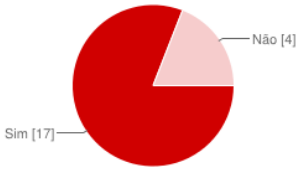
Sim	**17**	81%
Não	**4**	19%

23. Se, na verdade, Lelei não <<<manteve completamente>>> o ritmo na fase final, também não fraquejou muito .



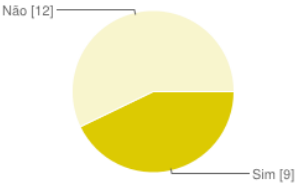
Sim	**13**	62%
Não	**8**	38%

24. Antes das eleições haverá ainda uma Ag para apriovação das contas e uma conferência de Imprensa da Direcção para <<<responder formalmente>>> aos ataques da lista opositora .



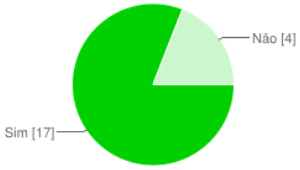
Sim	**17**	81%
Não	**4**	19%

25. Alguns dos filmes desta secção, como Justino, Assassino da 3ª Idade, ou Fresh Kill, de Shu Lea Cheang, <<<concorrem igualmente>>> na secção oficial .



Sim	**9**	43%
Não	**12**	57%

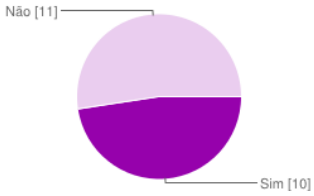
26. Quanto a Clinton, o seu ponto fraco é ser aquilo que os americanos chamam "too clever a half", qualquer coisa como «demasiado esperto», ter uma mulher demasiado inteligente e não conseguir <<<afastar completamente>>> a sua imagem de Slick Willie, «um rapaz demasiado bom para ser verdade .



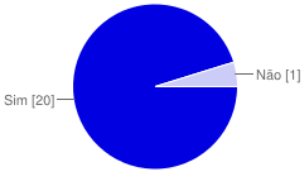
Sim	**17**	81%
Não	**4**	19%

27. Mas se são notórias as clivagens etárias, este estudo <<<detectou igualmente>>> fortes clivagens regionais (foram adoptadas as divisões das actuais Comissões de Coordenação Regional, CCR's) .

Sim	**10**	48%
Não	**11**	52%

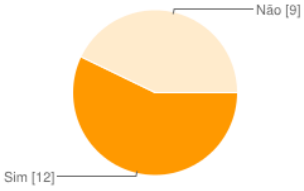


28. Estes números <<<mostram claramente>>> que a inflação continua baixa, mas este facto não impedirá o FED de manter o aperto das taxas, comentou um economista do DKB International .



Sim	**20**	95%
Não	**1**	5%

29. SAAB -- Entrando pela porta principal da Exponor e <<<virando imediatamente>>> à esquerda, o primeiro «stand» que aparece é o da Saab .

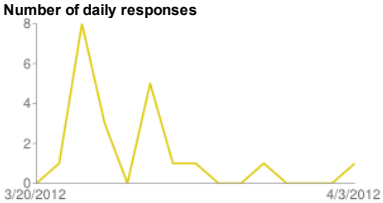


Sim	**12**	57%
Não	**9**	43%

30. Durante esta missão, os astronautas <<<testaram igualmente>>> equipamentos que vão ser utilizados em Dezembro, numa missão destinada a reparar o telescópio espacial Hubble .

Sim	**11**	52%
Não	**10**	48%

Number of daily responses



Appendix C. Sample of PT>EN Collocation Lexicon

PT	EN
abalar fortemente	*upset badly, deeply, really, terribly; shock deeply*
abandonar definitivamente	*abandon altogether, completely, entirely, totally*
abraçar efusivamente	*hug tightly, tight*
accionar criminalmente	*take/file/bring/ initiate - criminal - action*
aceitar humildemente	accept gratefully
acompanhar atentamente	follow carefully
adiar eternamente	postpone indefinitely
adoecer gravemente	fall/ get/ grow - critically, dangerously, gravelly, extremely, seriously, severely,terribly, very - ill
afectar gravemente	affect adversely, badly, seriously, severely
afirmar convictamente	state confidently, with confidence
agravar fortemente	aggravate seriously, severely
aguardar calmamente	wait patiently
analisar detalhadamente	analyse in detail, in depth
analisar exaustivamente	analyse painstakingly
aplaudir delirantemente	applaud wildly
aplaudir efusivamente	applaud enthusiastically, heartily
atacar ferozmente	attack savagely
atacar furiosamente	attack brutally, savagely, viciously, violently
atingir fortemente	hit hard
aumentar assustadoramente	increase tremendously
aumentar brutalmente	increase dramatically, drastically
bater estrondosamente	beat loudly
esperar eternamente	wait forever
explicar detalhadamente	explain in detail
falar correntemente	speak fluently
falar francamente	speak earnestly
falhar estrondosamente	fail spectacularly, completely, totally
ganhar folgadamente	win comfortably
lutar diariamente	struggle daily
obedecer estritamente	strictly comply with
mentir descaradamente	lie blatantly
olhar fixamente	look intently
pedir delicadamente	ask gently
penalizar duramente	penalise heavily, severely
penalizar fortemente	penalise heavily, severely

Appendix D. Classification of *Adv-mente*

Adv	Class	Class XIP
academicamente	MV MP	adv += [advpov=+, advmanner=+].
acaloradamente	MV	adv += [advmanner=+].
acintosamente	MS	adv += [advmansubj=+].
adversamente	MV	adv += [advmanner=+].
afetuosamente	MS	adv += [advmansubj=+].
agilmente	MS	adv += [advmansubj=+].
agressivamente	MS	adv += [advmansubj=+].
agudamente	MV	adv += [advmanner=+].
alucinadamente	MS	adv += [advmansubj=+].
amavelmente	MS	adv += [advmansubj=+].
amistosamente	MS	adv += [advmansubj=+].
amorosamente	MS	adv += [advmansubj=+].
analiticamente	MS	adv += [advmansubj=+].
analogicamente	MV	adv += [advmanner=+].
anatomicamente	MP	adv += [advpov=+].
anormalmente	PAa	adv += [adveval=+].
ardilosamente	MV	adv += [advmanner=+].
aritmeticamente	MV MP	adv += [advmanner=+,advpov=+].
arrebatadamente	MS	adv += [advmansubj=+].
arrogantemente	MS	adv += [advmansubj=+].
asperamente	MS	adv += [advmansubj=+].
assiduamente	MV	adv += [advmanner=+].
astuciosamente	MS	adv += [advmansubj=+].
atenciosamente	MS	adv += [advmansubj=+].
autonomamente	MS	adv += [advmansubj=+].
avassaladoramente	MS	adv += [advmansubj=+].
belamente	MS	adv += [advmansubj=+].
bisonhamente	MS	adv += [advmansubj=+].
brandamente	MS	adv += [advmansubj=+].
brasileiramente	MS	adv += [advmansubj=+].
burramente	MS	adv += [advmansubj=+].
calculadamente	MV	adv += [advmanner=+].
caracteristicamente	MF	adv += [advfocus=+].
celeremente	MV	adv += [advmanner=+].
centralmente	MV	adv += [advmanner=+].
chocantemente	PAa	adv += [adveval=+].
cinematograficamente	MP	adv += [advpov=+].
ciosamente	MS	adv += [advmansubj=+].
cirurgicamente	MV	adv += [advmanner=+].
civicamente	MV	adv += [advmanner=+].

Adv	Class	Class XIP
civilizadamente	MS	adv += [advmansubj=+].
coercitivamente	MV	adv += [advmanner=+].
comicamente	MS	adv += [advmansubj=+].
compassadamente	MV	adv += [advmanner=+].
competentemente	MS	adv += [advmansubj=+].
competitivamente	MV	adv += [advmanner=+].
concorrentemente	MV	adv += [advmanner=+].
confessadamente	MV	adv += [advmanner=+].
consecutivamente	MV MP	adv += [advpov=+, advmanner=+].
consensualmente	MV	adv += [advmanner=+].
continuadamente	MT	adv += [advtimeasp=+].
convictamente	MV	adv += [advmanner=+].
correspondentemente	PC	adv += [advconj=+].
corriqueiramente	MT	adv += [advhabit=+].
cortesmente	MS	adv += [advmansubj=+].
costumeiramente	PAh	adv += [advhabit=+].
crucialmente	MV	adv += [advmanner=+].
culposamente	MV	adv += [advmanner=+].
cumpridamente	MQ	adv += [advexact=+].
dedutivamente	MV	adv += [advmanner=+].
deficientemente	MV	adv += [advmanner=+].
delirantemente	MS	adv += [advmansubj=+].
depreciativamente	MV	adv += [advmanner=+].
desabridamente	MS	adv += [advmansubj=+].
desajeitadamente	MS	adv += [advmansubj=+].
desastradamente	MS	adv += [advmansubj=+].
descansadamente	MS	adv += [advmansubj=+].
descontraidamente	MS	adv += [advmansubj=+].
descontroladamente	MS	adv += [advmansubj=+].
descritivamente	MV	adv += [advmanner=+].
desdenhosamente	MS	adv += [advmansubj=+].
desmesuradamente	MQ	adv += [advsupra=+].
despreocupadamente	MS	adv += [advmansubj=+].
despretensiosamente	MS	adv += [advmansubj=+].
desproporcionalmente	MV	adv += [advmanner=+].
despudoradamente	MS	adv += [advmansubj=+].
diabolicamente	PAa	adv += [adveval=+].
diligentemente	MS	adv += [advmansubj=+].
discursivamente	MP	adv += [advpov=+].
displicentemente	MS	adv += [advmansubj=+].
dissimuladamente	MV	adv += [advmanner=+].
diuturnamente	MV	adv += [advmanner=+].
docilmente	MS	adv += [advmansubj=+].
documentalmente	MV	adv += [advmanner=+].

Adv	Class	Class XIP
dogmaticamente	MS	adv += [advmansubj=+].
dolosamente	MV	adv += [advmanner=+].
editorialmente	MV	adv += [advmanner=+].
educadamente	MS	adv += [advmansubj=+].
eleitoreiramente	MS	adv += [advmansubj=+].
emblematicamente	MV	adv += [advmanner=+].
emergencialmente	MV	adv += [advmanner=+].
engenhosamente	MS	adv += [advmansubj=+].
episodicamente	MT	adv += [advtimeasp=+].
equilibradamente	MV	adv += [advmanner=+].
escancaradamente	MV	adv += [advmanner=+].
escassamente	MQ	adv += [advinfra=+].
escrupulosamente	MS	adv += [advmansubj=+].
espacialmente	MP	adv += [advpov=+].
espertamente	MS PAs	adv += [advmansubj=+].
espetacularmente	PAa	adv += [adveval=+].
esplendidamente	PAa	adv += [adveval=+].
esquematicamente	MV	adv += [advmanner=+].
estavelmente	MS	adv += [advmansubj=+].
estilisticamente	MP	adv += [advpov=+].
estrondosamente	MS	adv += [advmansubj=+].
etimologicamente	MP	adv += [advpov=+].
etnicamente	MP	adv += [advpov=+].
explosivamente	MS	adv += [advmansubj=+].
exponencialmente	MV	adv += [advmanner=+].
exteriormente	MV	adv += [advmanner=+].
factualmente	MV	adv += [advmanner=+].
facultativamente	MV	adv += [advmanner=+].
fanaticamente	MS	adv += [advmansubj=+].
fantasmaticamente	MV	adv += [advmanner=+].
febrilmente	MV	adv += [advmanner=+].
ferrenhamente	MS	adv += [advmansubj=+].
festivamente	MV	adv += [advmanner=+].
ficcionalmente	MV	adv += [advmanner=+].
fiduciariamente	MV	adv += [advmanner=+].
figurativamente	MV	adv += [advmanner=+].
folgadamente	MV	adv += [advmanner=+].
fotograficamente	MV MP	adv += [advmanner=+,advpov=+].
fraternalmente	MS	adv += [advmansubj=+].
funcionalmente	MV MP	adv += [advpov=+, advmanner=+].
fundamentadamente	MV	adv += [advmanner=+].
fundamente	MV	adv += [advmanner=+].
galhardamente	MS	adv += [advmansubj=+].
generalizadamente	MV	adv += [advmanner=+].

Adv	Class	Class XIP
genialmente	MS	adv += [advmansubj=+].
geometricamente	MV MP	adv += [advpov=+, advmanner=+].
geopoliticamente	MP	adv += [advpov=+].
gerencialmente	MV MP	adv += [advpov=+, advmanner=+].
gostosamente	MV	adv += [advmanner=+].
gramaticalmente	MV MP	adv += [advmanner=+,advpov=+].
grotescamente	MS	adv += [advmansubj=+].
harmonicamente	MV	adv += [advmanner=+].
hereditariamente	MV	adv += [advmanner=+].
horrivelmente	MV	adv += [advmanner=+].
identicamente	MV	adv += [advmanner=+].
igualitariamente	MV	adv += [advmanner=+].
ilusoriamente	MV	adv += [advmanner=+].
imaginariamente	MV	adv += [advmanner=+].
imperfeitamente	MV	adv += [advmanner=+].
imperiosamente	MS	adv += [advmansubj=+].
impressionantemente	PAa	adv += [adveval=+].
impreterivelmente	MV	adv += [advmanner=+].
impropriamente	MV	adv += [advmanner=+].
imprudentemente	MS	adv += [advmansubj=+].
impulsivamente	MS	adv += [advmansubj=+].
incisivamente	MS	adv += [advmansubj=+].
inconstitucionalmente	MV	adv += [advmanner=+].
inconvenientemente	PAa	adv += [adveval=+].
indissociavelmente	MV	adv += [advmanner=+].
indissoluvelmente	MV	adv += [advmanner=+].
indolentemente	MS	adv += [advmansubj=+].
industrialmente	MV	adv += [advmanner=+].
inelutavelmente	MV	adv += [advmanner=+].
inextricavelmente	MV	adv += [advmanner=+].
infalivelmente	MV	adv += [advmanner=+].
infantilmente	MS	adv += [advmansubj=+].
infatigavelmente	MS	adv += [advmansubj=+].
inopinadamente	MV	adv += [advmanner=+].
inquietantemente	PAa	adv += [adveval=+].
insidiosamente	MS	adv += [advmansubj=+].
insuportavelmente	PAa	adv += [adveval=+].
intempestivamente	MV	adv += [advmanner=+].
intensivamente	MV	adv += [advmanner=+].
interativamente	MV	adv += [advmanner=+].
interminavelmente	MT	adv += [advtimeasp=+].
intermitentemente	MT	adv += [advtimeasp=+].
intrinsicamente	MV	adv += [advmanner=+].
irracionalmente	MS	adv += [advmansubj=+].

Adv	Class	Class XIP
irrecorrivelmente	MV	adv += [advmanner=+].
irrefletidamente	PAa	adv += [adveval=+].
irrestritamente	MV	adv += [advmanner=+].
irritantemente	PAa	adv += [adveval=+].
isoclinalmente	MV	adv += [advmanner=+].
isotopicamente	MV	adv += [advmanner=+].
jornalisticamente	MP	adv += [advpov=+].
justificadamente	PAa	adv += [adveval=+].
laboriosamente	MV	adv += [advmanner=+].
laconicamente	MS	adv += [advmansubj=+].
languidamente	MS	adv += [advmansubj=+].
lealmente	MS	adv += [advmansubj=+].
licitamente	MV	adv += [advmanner=+].
lindamente	MV	adv += [advmanner=+].
linearmente	MV	adv += [advmanner=+].
linguisticamente	MP	adv += [advpov=+].
longinquamente	MQ	adv += [advsupra=+].
longitudinalmente	MV	adv += [advmanner=+].
lucidamente	MS	adv += [advmansubj=+].
macroscopicamente	MV	adv += [advmanner=+].
maldosamente	MS	adv += [advmansubj=+].
massivamente	MV	adv += [advmanner=+].
melodicamente	MV	adv += [advmanner=+].
metabolicamente	MV MP	adv += [advpov=+, advmanner=+].
metodicamente	MS	adv += [advmansubj=+].
metodologicamente	MP	adv += [advpov=+].
metonimicamente	MV	adv += [advmanner=+].
miticamente	MV	adv += [advmanner=+].
miudamente	MV	adv += [advmanner=+].
molemente	MS	adv += [advmansubj=+].
monotonamente	MS	adv += [advmansubj=+].
morfologicamente	MP	adv += [advpov=+].
narcisicamente	MS	adv += [advmansubj=+].
negligentemente	MS	adv += [advmansubj=+].
nobremente	MS	adv += [advmansubj=+].
nomeadamente	MF	adv += [advfocus=+].
nostalgicamente	MS	adv += [advmansubj=+].
obscenamente	MS	adv += [advmansubj=+].
obscuramente	MV	adv += [advmanner=+].
ocultamente	MV	adv += [advmanner=+].
olimpicamente	MV	adv += [advmanner=+].
opcionalmente	MV	adv += [advmanner=+].
oportunisticamente	PAa	adv += [adveval=+].
ordenadamente	MV	adv += [advmanner=+].

Adv	Class	Class XIP
organizadamente	MS	adv += [advmansubj=+].
ortogonalmente	MV	adv += [advmanner=+].
otimamente	MV	adv += [advmanner=+].
paternalmente	MS	adv += [advmansubj=+].
patrioticamente	MS	adv += [advmansubj=+].
pausadamente	MV	adv += [advmanner=+].
peculiarmente	MS MF	adv += [advmansubj=+,advfocus=+].
pedagogicamente	MP	adv += [advpov=+].
penosamente	MV PAa	adv += [adveval=+,advmanner=+].
percentualmente	MV	adv += [advmanner=+].
perdidamente	MQ	adv += [advsupra=+].
perpendicularmente	MV	adv += [advmanner=+].
persuasivamente	MS	adv += [advmansubj=+].
pertinentemente	PAa	adv += [adveval=+].
pioneiramente	MS	adv += [advmansubj=+].
placidamente	MS	adv += [advmansubj=+].
poderosamente	MS	adv += [advmansubj=+].
polidamente	MS	adv += [advmansubj=+].
poligonalmente	MV	adv += [advmanner=+].
pomposamente	MS	adv += [advmansubj=+].
porcamente	MS	adv += [advmansubj=+].
pormenorizadamente	MV	adv += [advmanner=+].
pragmaticamente	MP MV	adv += [advpov=+, advmanner=+].
prazerosamente	MV	adv += [advmanner=+].
precedentemente	MT	adv += [advtimedate=+,t-ref-before=+,t-tempref=text].
preferivelmente	MV	adv += [advmanner=+].
premeditadamente	MV	adv += [advmanner=+].
preponderantemente	MF	adv += [advfocus=+].
presumidamente	PAm	adv += [advmodal=+].
primariamente	MF	adv += [advfocus=+].
primorosamente	MV	adv += [advmanner=+].
privadamente	MV	adv += [advmanner=+].
privativamente	MV	adv += [advmanner=+].
prodigiosamente	PAa	adv += [adveval=+].
profeticamente	MV	adv += [advmanner=+].
profusamente	MV	adv += [advmanner=+].
prosaicamente	MV	adv += [advmanner=+].
providencialmente	PAa	adv += [adveval=+].
prudentemente	MS	adv += [advmansubj=+].
psicanaliticamente	MP	adv += [advpov=+].
pudicamente	MS	adv += [advmansubj=+].
quintessencialmente	MF	adv += [advfocus=+].
racialmente	MP	adv += [advpov=+].

Adv	Class	Class XIP
radialmente	MV	adv += [advmanner=+].
realisticamente	MV	adv += [advmanner=+].
regiamente	MV	adv += [advmanner=+].
responsavelmente	MS	adv += [advmansubj=+].
restritamente	MV	adv += [advmanner=+].
restritivamente	MS	adv += [advmansubj=+].
retoricamente	MV MP	adv += [advpov=+, advmanner=+].
ritmicamente	MV	adv += [advmanner=+].
romanticamente	MS	adv += [advmansubj=+].
rudemente	MS	adv += [advmansubj=+].
sarcasticamente	MS	adv += [advmansubj=+].
sazonalmente	MT	adv += [advtimeasp=+].
secularmente	MV	adv += [advmanner=+].
selvagemente	MS	adv += [advmansubj=+].
sensatamente	MS	adv += [advmansubj=+].
sensualmente	MS	adv += [advmansubj=+].
sentimentalmente	MP MV	adv += [advpov=+, advmanner=+].
servilmente	MS	adv += [advmansubj=+].
significantemente	PAa	adv += [adveval=+].
Similarmente	PC	adv += [advconj=+].
similarmente	MV	adv += [advmanner=+].
simpaticamente	PAa	adv += [adveval=+].
sincronicamente	MP MV	adv += [advpov=+, advmanner=+].
singelamente	MS	adv += [advmansubj=+].
sinistramente	PAa	adv += [adveval=+].
sintaticamente	MP	adv += [advpov=+].
sintomaticamente	PAa	adv += [adveval=+].
sintomaticamente	PAa MV	adv += [adveval=+,advmanner=+].
sobejamente	MQ	adv += [advsupra=+].
soberanamente	MV	adv += [advmanner=+].
soberbamente	MS	adv += [advmansubj=+].
sofrivelmente	MV	adv += [advmanner=+].
sonoramente	MV	adv += [advmanner=+].
sossegadamente	MS	adv += [advmansubj=+].
subjetivamente	MV	adv += [advmanner=+].
subliminarmente	MV	adv += [advmanner=+].
subsequentemente	MT PC	adv += [advtimedate=+,t-ref-before=+,t-tempref=text,advconj=+].
subterraneamente	MV	adv += [advmanner=+].
sugestivamente	PAa MV	adv += [adveval=+,advmanner=+].
superlativamente	MQ	adv += [advsupra=+].
supletivamente	MV	adv += [advmanner=+].
taxativamente	MV	adv += [advmanner=+].
tectonicamente	MP	adv += [advpov=+].

Adv	Class	Class XIP
tediosamente	PAa	adv += [adveval=+].
tematicamente	MP	adv += [advpov=+].
temerariamente	MS	adv += [advmansubj=+].
tenazmente	MS	adv += [advmansubj=+].
tendencialmente	MV	adv += [advmanner=+].
ternamente	MS	adv += [advmansubj=+].
territorialmente	MP MV	adv += [advpov=+, advmanner=+].
tolamente	MS	adv += [advmansubj=+].
torrencialmente	MQ	adv += [advsupra=+].
transversalmente	MV	adv += [advmanner=+].
tridimensionalmente	MV	adv += [advmanner=+].
triplamente	MQ	adv += [advexact=+].
triunfalmente	MV	adv += [advmanner=+].
umbilicalmente	MV	adv += [advmanner=+].
vergonhosamente	PAa	adv += [adveval=+].
vorazmente	MS	adv += [advmansubj=+].
zelosamente	MS	adv += [advmansubj=+].
vocalmente	MV	adv += [advmanner=+].
vividamente	MV	adv += [advmanner=+].
vantajosamente	MV	adv += [advmanner=+].
valentemente	MS	adv += [advmansubj=+].
terminalmente	MV	adv += [advmanner=+].
tentativamente	MV	adv += [advmanner=+].
subconscientemente	MV	adv += [advmanner=+].
solertemente	MS	adv += [advmansubj=+].
sinuosamente	MV	adv += [advmanner=+].
serialmente	MV	adv += [advmanner=+].
sequencialmente	MV	adv += [advmanner=+].
semioticamente	MP MV	adv += [advpov=+, advmanner=+].
semelhantemente	PC	adv += [advconj=+].
saborosamente	MV	adv += [advmanner=+].
rotundamente	MQ	adv += [advsupra=+].
reverentemente	MS	adv += [advmansubj=+].
relutantemente	MS	adv += [advmansubj=+].
prolongadamente	MV	adv += [advmanner=+].
piedosamente	MS	adv += [advmansubj=+].
ousadamente	MS	adv += [advmansubj=+].
oficiosamente	MV	adv += [advmanner=+].
mutualmente	MV	adv += [advmanner=+].
monstruosamente	MQ	adv += [advsupra=+].
miraculosamente	PAa	adv += [adveval=+].
mesquinhamente	PAa	adv += [adveval=+].
maternalmente	MS	adv += [advmansubj=+].
malandramente	MS	adv += [advmansubj=+].

Adv	Class	Class XIP
localizadamente	MV	adv += [advmanner=+].
judiciosamente	MV	adv += [advmanner=+].
jocosamente	MS	adv += [advmansubj=+].
irrealisticamente	MV	adv += [advmanner=+].
interrogativamente	MV	adv += [advmanner=+].
indiscretamente	MS	adv += [advmansubj=+].
incompletamente	MV	adv += [advmanner=+].
incidentemente	MV	adv += [advmanner=+].
imprevistamente	MV	adv += [advmanner=+].
imperialmente	PAa	adv += [adveval=+].
impensadamente	MS	adv += [advmansubj=+].
imaculadamente	MS	adv += [advmansubj=+].
horrorosamente	PAa	adv += [adveval=+].
honradamente	MS	adv += [advmansubj=+].
hilariantemente	MS	adv += [advmansubj=+].
gritantemente	PAa	adv += [adveval=+].
fugazmente	MV	adv += [advmanner=+].
fraudulentamente	MV	adv += [advmanner=+].
fragorosamente	MV	adv += [advmanner=+].
fisiologicamente	MP	adv += [advpov=+].
figuradamente	MV	adv += [advmanner=+].
exuberantemente	MS	adv += [advmansubj=+].
exemplificativamente	MV	adv += [advmanner=+].
excelentemente	PAa	adv += [adveval=+].
evolutivamente	MV	adv += [advmanner=+].
eufemisticamente	MV	adv += [advmanner=+].
estudadamente	MV	adv += [advmanner=+].
estrepitosamente	MV	adv += [advmanner=+].
estatutariamente	MP	adv += [advpov=+].
esparsamente	MV	adv += [advmanner=+].
encantadoramente	MS	adv += [advmansubj=+].
empresarialmente	MP	adv += [advpov=+].
embrionariamente	MV	adv += [advmanner=+].
duradouramente	MV	adv += [advmanner=+].
divinamente	PAa	adv += [adveval=+].
divertidamente	MS	adv += [advmansubj=+].
distorcidamente	MV	adv += [advmanner=+].
disciplinarmente	MP MV	adv += [advpov=+, advmanner=+].
difusamente	MV	adv += [advmanner=+].
devastadoramente	MV	adv += [advmanner=+].
deslealmente	MS	adv += [advmansubj=+].
descuidadamente	MS	adv += [advmansubj=+].
desastrosamente	PAa	adv += [adveval=+].
desamparadamente	MS	adv += [advmansubj=+].

Adv	Class	Class XIP
desafortunadamente	PAa	adv += [adveval=+].
corretivamente	MV	adv += [advmanner=+].
correlativamente	MV	adv += [advmanner=+].
coreograficamente	MP	adv += [advpov=+].
construtivamente	MS	adv += [advmansubj=+].
constrangedoramente	PAa	adv += [adveval=+].
conseguintemente	PC	adv += [advconj=+].
condicionalmente	MV	adv += [advmanner=+].
circunstancialmente	MV	adv += [advmanner=+].
circunstanciadamente	MV	adv += [advmanner=+].
cerradamente	MV	adv += [advmanner=+].
cautelarmente	MV	adv += [advmanner=+].
causalmente	MV	adv += [advmanner=+].
caracterizadamente	MF	adv += [advfocus=+].
camaleonicamente	MS	adv += [advmansubj=+].
bovinamente	MV	adv += [advmanner=+].
biauditivamente	MV	adv += [advmanner=+].
atrevidamente	MS	adv += [advmansubj=+].
assombrosamente	PAa	adv += [adveval=+].
articuladamente	MS	adv += [advmansubj=+].
alarmantemente	PAa	adv += [adveval=+].
afoitamente	MS	adv += [advmansubj=+].
acirradamente	MV	adv += [advmanner=+].
acanhadamente	MS	adv += [advmansubj=+].
acacianamente	MS	adv += [advmansubj=+].
abreviadamente	MV	adv += [advmanner=+].
abjetamente	MS	adv += [advmansubj=+].

Appendix E. Values of Association Measures Used in the MT Evaluation

	Different class-1 bigrams (Google Translate™)	*t* test	χ^2	LLR	MI	Dice	UnigSub
1	accept humbly	1.961268717	1446514.311	10032.95747	12.57910015	0.000106898	36.41487258
2	monitor closely	18.15456051	123394749.8	465136.011	14.72584454	0.018418051	36.75472838
3	postpone forever	2.141546798	139385085.1	321554.7256	15.83178862	0.000496796	28.95093296
4	attack fiercely	3.662581763	17669079.9	76799.46695	13.47247876	0.000645206	35.2662742
5	hit heavily	3.827755389	29732828.83	349465.9786	11.90631728	0.000337323	35.33341656
6	increase alarmingly	3.95233555	1949209.944	17508.17816	12.14248659	0.000315246	38.82497599
7	talk happily	2.713507113	9938081.538	138734.1526	11.46533417	0.000142617	35.84970395
8	grow alarmingly	3.240368528	1467753.818	16797.86142	11.73482863	0.000163725	38.51547307
9	grow enormously	7.951370817	5129346.13	59362.27092	11.73482863	0.000951983	39.85457676
10	greet warmly	13.30014419	34561108.15	53111.89211	15.20910805	0.025242442	38.65838435
11	determine together	1.913328012	158572436.8	2480666.105	13.04717139	0.000270325	31.47021545
12	demonstrate fully	1.261172633	135536051	732110.5475	14.37879288	0.000151403	28.55642588
13	speak frankly	11.25060852	9119156.397	110164.7765	11.68850972	0.001811165	40.12248378
14	lie shamelessly	2.7956863	1752444.492	12080.04449	12.58758364	0.000214771	37.62892866
15	speak openly	16.15315469	10845337.48	130944.198	11.68850972	0.003664827	41.01122203
16	reject flatly	1.674136496	8681835.169	31432.90297	13.71084549	0.000170058	33.10485427
17	reply firmly	7.52010403	62942798.68	262236.0591	13.9120118	0.002846556	35.58456553
18	follow scrupulously	2.149589215	481471.4429	7872.55476	11.12209033	4.88785E-05	38.50731779
19	rise constantly	0.233074514	28881820.5	267504.5304	12.30192125	5.81249E-05	31.22771845
20	clandestinely	2.409128822	92656.8252	2264.807991	10.40450505	3.57295E-05	41.32793237
21	strongly shock	0.545507532	131093179.6	418035.6903	15.17118075	6.87238E-05	25.02778932
22	embrace warmly	7.27195902	30133894.1	52195.59151	15.04856753	0.006866174	36.80239227
23	say convincingly	1.682623782	82364.99591	11403.51173	7.277938138	8.86426E-06	42.96402708
24	calmly wait	4.836708722	7108927.079	69344.09073	12.04743977	0.000450629	37.71368027
25	attack furiously	1.635997864	10268264.98	43268.92738	13.47247876	0.000143113	32.85137143
26	beat furiously	4.034827612	4828222.954	39094.9706	12.34300424	0.000379986	37.59319299
27	weep copiously	3.604911216	5255824.645	4342.721455	16.02607244	0.003728668	37.1004218
28	grow alarmingly	3.240368528	1467753.818	16797.86142	11.73482863	0.000163725	38.51547307
29	grow enormously	7.951370817	5129346.13	59362.27092	11.73482863	0.000951983	39.85457676
30	fail disastrously	3.29492001	910104.7542	7427.867218	12.28798067	0.000240829	39.23919455
31	penalise harshly	1.727140134	90592775.66	48344.17535	17.51943376	0.001492537	29.04477672
32	fully prove	4.087189806	67535955.71	648943.8861	12.57491573	0.000490293	34.06100728
33	work jointly	6.091906354	1825951.603	45458.95583	10.40450505	0.0002423	40.57963314
34	shake strongly	3.005755629	63758253.65	372863.3179	13.36651884	0.000400976	32.7579961
35	follow closely	25.31656604	19884014.83	348389.9717	11.12209033	0.005893206	41.45198494
36	analyse thoroughly	1.912255604	91279474.52	189328.5633	15.34659956	0.000438765	29.69463562
37	applaud warmly	2.820808413	66643048.75	57323.65223	16.50993449	0.002323218	31.87609708
38	strongly affect	4.527195885	57680017.54	367232.093	13.16849275	0.000725152	34.23473764
39	weep convulsively	0.998381565	3564152.407	3029.233532	16.02607244	0.000288725	31.44948261
40	chat cheerfully	2.811666063	26871716.76	44260.65884	15.1275135	0.001110417	33.50610533
41	grow markedly	3.892687843	2477812.872	28494.92559	11.73482863	0.000237084	38.46147121
42	decide jointly	3.264187273	5156069.23	53575.62905	11.92934279	0.000201054	36.83316582
43	fail resoundingly	0.972205169	348083.1217	2861.841747	12.28798067	2.19467E-05	34.82410708
44	strictly obey	1.376527051	129868808.1	159033.0867	16.70072097	0.00040205	27.02366516

	Different class-1 bigrams (Google Translate™)	*t* test	χ^2	LLR	MI	Dice	UnigSub
45	ask politely	12.59934084	2025179.336	44116.49646	10.60333981	0.001098134	42.7712462
46	punish hard	3.258286476	171200095.7	765333.8851	15.42574742	0.000575264	29.79266867
47	search incessantly	0.968803402	4240230.692	12466.92416	14.0220582	7.17592E-05	31.19059198
48	speak out openly	2.318102422	10620498.78	130389.4575	11.68850972	0.000124466	35.18708737
49	prove absolutely	3.821227232	62938948.59	589194.3711	12.57491573	0.000446177	33.98589196
50	respond strongly	7.221648787	68385484.46	376877.958	13.50461852	0.001967061	35.32393592
51	follow strictly	64.7989773	9967838.447	115899.086	11.12209033	0.040069578	47.78317935
52	embrace enthusiastically	4.888908768	25414804.01	43402.31005	15.04856753	0.003179018	35.69306182
53	talk cheerfully	3.71802081	2382372.222	32131.89334	11.46533417	0.00018447	38.39448491
54	want passionately	3.060518877	754886.2006	26856.78989	9.777746892	4.99892E-05	39.78642239

	Reference bigrams Google Translate™	*t* test	χ^2	LLR	MI	Dice	UnigSub
1	accept gratefully	5.452252135	2578335.463	17758.01421	12.57910015	0.000797862	39.58152774
2	follow carefully	10.30222932	22069896.7	396267.2238	11.12209033	0.001105999	38.66053475
3	postpone indefinitely	12.24299409	68432965.69	86820.54026	15.83178862	0.027312454	37.04801766
4	attack savagely	2.203607962	4607248.876	18915.75568	13.47247876	0.000244744	35.21659001
5	hit hard	51.71724616	50625689.28	618132.7596	11.90631728	0.035896991	42.10150055
6	increase tremendously	5.504919042	3554493.69	32119.9739	12.14248659	0.000606487	39.13792439
7	talk animatedly	8.177791564	305537.5237	3470.015823	11.46533417	0.000831751	44.85588256
8	grow dramatically	10.67673229	13901470.73	167406.8236	11.73482863	0.001674464	39.32299087
9	grow dramatically	10.67673229	13901470.73	167406.8236	11.73482863	0.001674464	39.32299087
10	greet enthusiastically	5.559836471	28119329.94	43924.31955	15.20910805	0.004530508	35.98657794
11	decide collectively	1.985413685	4186275.216	43364.26683	11.92934279	8.41354E-05	35.35115655
12	demonstrate conclusively	5.379743008	6599333.466	15434.82083	14.37879288	0.002638522	38.16200054
13	speak earnestly	6.038432593	1489399.612	17313.84417	11.68850972	0.00053331	40.75399852
14	lie blatantly	2.587292809	2902144.148	20155.88978	12.58758364	0.000186966	36.60990168
15	state outright	3.294962685	17087267.58	40379.6307	14.46607843	0.001012332	34.86239475
16	reject outright	11.65150518	10895963.73	38272.32116	13.71084549	0.007648399	40.05719672
17	answer confidently	3.265102991	9531178.216	41867.01957	13.39523132	0.000499127	35.75381262
18	follow to the letter	11.76497632	811417.3612	12214.32868	11.12209033	0.001357349	44.1978125
19	rise steadily	17.76513285	14786771.73	125220.3256	12.30192125	0.00661212	40.84844955
20	work illegally	8.947069949	3235417.812	80907.6871	10.40450505	0.000510967	40.99893594
21	shock deeply	5.000148464	137004645.2	461710.6122	15.17118075	0.001680998	32.22152123
22	hug tightly	11.08522038	130327638.3	163319.854	16.59713847	0.023574509	35.04285022
23	state confidently	3.440619369	19155524.49	45691.33982	14.46607843	0.001093544	34.85081328
24	wait patiently	24.5762034	5195627.714	44311.90156	12.04743977	0.011036219	43.96634553
25	attack viciously	4.109107846	3716787.142	15071.751	13.47247876	0.000835586	38.00874994
26	beat badly	11.07033452	41823007.9	400535.1869	12.34300424	0.002424966	37.65865408
27	cry uncontrollably	6.474782615	4663791.37	14246.27235	13.91152317	0.002788382	39.36341165
28	rise dramatically	17.36644401	20282094.02	177256.9042	12.30192125	0.006177778	40.25388525
29	rise dramatically	17.36644401	20282094.02	177256.9042	12.30192125	0.006177778	40.25388525
30	fail completely	10.01085882	64065428.77	714444.9303	12.28798067	0.00180146	36.64284435
31	penalise heavily	3.579243427	200488718.4	480656.4224	17.51943376	0.001284395	28.66050054
32	rise steeply	13.02654223	2196618.792	16456.47696	12.30192125	0.003740497	43.08676838
33	work together	90.48378076	64469100.41	2047124.146	10.40450505	0.03642441	43.25688395

	Reference bigrams Google Translate™	*t* test	χ^2	LLR	MI	Dice	UnigSub
34	shock deeply	5.000148464	137004645.2	461710.6122	15.17118075	0.001680998	32.22152123
35	follow carefully	10.30222932	22069896.7	396267.2238	11.12209033	0.001105999	38.66053475
36	analyse in detail	6.235044387	44069261.8	68414.60483	15.34659956	0.005773074	35.60781571
37	applaud enthusiastically	3.736869027	58463251.15	47815.19783	16.50993449	0.004276115	33.2366032
38	hit hard	51.71724616	50625689.28	618132.7596	11.90631728	0.035896991	42.10150055
39	talk animatedly	8.177791564	305537.5237	3470.015823	11.46533417	0.000831751	44.85588256
40	cry uncontrollably	6.474782615	4663791.37	14246.27235	13.91152317	0.002788382	39.36341165
41	grow significantly	10.19573395	19142743.5	236345.0893	11.73482863	0.001523511	38.70480392
42	decide collectively	1.985413685	4186275.216	43364.26683	11.92934279	8.41354E-05	35.35115655
43	fail completely	10.01085882	64065428.77	714444.9303	12.28798067	0.00180146	36.64284435
44	fail completely	10.01085882	64065428.77	714444.9303	12.28798067	0.00180146	36.64284435
45	ask gently	14.11827643	8761102.219	199709.468	10.60333981	0.001407353	40.95473285
46	penalise heavily	3.579243427	200488718.4	480656.4224	17.51943376	0.001284395	28.66050054
47	search constantly	1.883650392	73725929.24	307348.6059	14.0220582	0.000247145	30.67273407
48	state outright	3.294962685	17087267.58	40379.6307	14.46607843	0.001012332	34.86239475
49	prove conclusively	9.581034604	2033700.28	13344.79993	12.57491573	0.002446743	42.04783085
50	answer confidently	3.265102991	9531178.216	41867.01957	13.39523132	0.000499127	35.75381262
51	follow to the letter	11.76497632	811417.3612	12214.32868	11.12209033	0.001357349	44.1978125
52	hug tightly	11.08522038	130327638.3	163319.854	16.59713847	0.023574509	35.04285022
53	talk animatedly	8.177791564	305537.5237	3470.015823	11.46533417	0.000831751	44.85588256
54	wish fervently	3.586042828	3395127.082	15627.03093	13.26624892	0.000554619	37.62578887

	Different class-1 bigrams Systranet™	*t* test	χ^2	LLR	MI	Dice	UnigSub
1	shake strongly	3.005755629	63758253.65	372863.3179	13.36651884	0.000400976	32.7579961
2	calmly wait	4.836708722	7108927.079	69344.09073	12.04743977	0.000450629	37.71368027
3	attack furiously	1.635997864	10268264.98	43268.92738	13.47247876	0.000143113	32.85137143
4	beat furiously	4.034827612	4828222.954	39094.9706	12.34300424	0.000379986	37.59319299
5	cry copiously	0.990006101	1184023.019	3669.64093	13.91152317	6.73764E-05	33.05299017
6	work jointly	6.091906354	1825951.603	45458.95583	10.40450505	0.0002423	40.57963314
7	use abusively	1.715461181	15388.8361	469.0947533	9.979074511	1.33068E-05	42.42075924
8	use unduly	-0.108739014	561858.098	17858.37499	9.979074511	4.42492E-06	34.1241432
9	attack ferociously	1.981524378	2373059.828	9638.801598	13.47247876	0.000197755	35.69667138
10	reach strongly	1.058721248	29173770.47	330788.0495	11.96350212	0.000107108	33.08947589
11	happily talk	2.713507113	9938081.538	138734.1526	11.46533417	0.000142617	35.84970395
12	greet effusively	3.999445602	1748361.571	2253.046124	15.20910805	0.002645284	39.30965078
13	decide jointly	3.264187273	5156069.23	53575.62905	11.92934279	0.000201054	36.83316582
14	demonstrate completely	1.767784068	142155262.5	822334.6067	14.37879288	0.000210774	29.2667271
15	speak frankly	11.25060852	9119156.397	110164.7765	11.68850972	0.001811165	40.12248378
16	strictly obey	1.376527051	129868808.1	159033.0867	16.70072097	0.00040205	27.02366516
17	lie shamelessly	2.7956863	1752444.492	12080.04449	12.58758364	0.000214771	37.62892866
18	look fixedly	5.351319756	90326.7238	2810.990894	9.898864295	0.00012164	44.63117544
19	search unceasingly	0.994590185	750245.5668	2165.576706	14.0220582	7.2881E-05	33.7164111
20	refer concretely	0.991285647	746355.6629	2663.466324	13.67528068	5.73099E-05	33.7226495

	Different class-1 bigrams Systranet™	*t* test	χ^2	LLR	MI	Dice	UnigSub
21	follow conscientiously	1.346639224	237928.324	3889.982964	11.12209033	1.95722E-05	37.3415899
22	rise constantly	0.233074514	28881820.5	267504.5304	12.30192125	5.81249E-05	31.22771845
23	analyse at great length	0.997191987	2422048.115	3123.347382	15.34659956	0.000181192	32.01378555
24	determine jointly	2.525140672	10895720.95	59500.18483	13.04717139	0.000247959	34.64191785

	Reference bigrams Systranet™	*t* test	χ^2	LLR	MI	Dice	UnigSub
1	shock deeply	5.000148464	137004645.2	461710.6122	15.17118075	0.001680998	32.22152123
2	wait patiently	24.5762034	5195627.714	44311.90156	12.04743977	0.011036219	43.96634553
3	attack viciously	4.109107846	3716787.142	15071.751	13.47247876	0.000835586	38.00874994
4	beat badly	11.07033452	41823007.9	400535.1869	12.34300424	0.002424966	37.65865408
5	cry uncontrollably	6.474782615	4663791.37	14246.27235	13.91152317	0.002788382	39.36341165
6	work together	90.48378076	64469100.41	2047124.146	10.40450505	0.03642441	43.25688395
7	use improperly	6.943718254	727852.4716	22841.61184	9.979074511	0.000225518	42.32123063
8	use improperly	6.943718254	727852.4716	22841.61184	9.979074511	0.000225518	42.32123063
9	attack savagely	2.203607962	4607248.876	18915.75568	13.47247876	0.000244744	35.21659001
10	hit hard	51.71724616	50625689.28	618132.7596	11.90631728	0.035896991	42.10150055
11	talk animatedly	8.177791564	305537.5237	3470.015823	11.46533417	0.000831751	44.85588256
12	greet enthusiastically	5.559836471	28119329.94	43924.31955	15.20910805	0.004530508	35.98657794
13	decide collectively	1.985413685	4186275.216	43364.26683	11.92934279	8.41354E-05	35.35115655
14	demonstrate conclusively	5.379743008	6599333.466	15434.82083	14.37879288	0.002638522	38.16200054
15	speak earnestly	6.038432593	1489399.612	17313.84417	11.68850972	0.00053331	40.75399852
16	strictly comply with	2.19262705	96011044.59	152205.6315	15.83475028	0.000742776	30.06694681
17	lie blatantly	2.587292809	2902144.148	20155.88978	12.58758364	0.000186966	36.60990168
18	look intently	12.72177185	759959.0018	24260.14914	9.898864295	0.000690148	44.31922073
19	search constantly	1.883650392	73725929.24	307348.6059	14.0220582	0.000247145	30.67273407
20	refer specifically	9.758910029	57203283.15	265849.8311	13.67528068	0.004183469	36.61902734
21	follow to the letter	11.76497632	811417.3612	12214.32868	11.12209033	0.001357349	44.1978125
22	rise steadily	17.76513285	14786771.73	125220.3256	12.30192125	0.00661212	40.84844955
23	analyse in detail	6.235044387	44069261.8	68414.60483	15.34659956	0.005773074	35.60781571
24	decide collectively	1.985413685	4186275.216	43364.26683	11.92934279	8.41354E-05	35.35115655

	Different class-1 bigrams Reverso™	*t* test	χ^2	LLR	MI	Dice	UnigSub
1	attack ferociously	1.981524378	2373059.828	9638.801598	13.47247876	0.000197755	35.69667138
2	strongly reach	1.058721248	29173770.47	330788.0495	11.96350212	0.000107108	33.08947589
3	talk happily	2.713507113	9938081.538	138734.1526	11.46533417	0.000142617	35.84970395
4	greet effusively	3.999445602	1748361.571	2253.046124	15.20910805	0.002645284	39.30965078
5	decide jointly	3.264187273	5156069.23	53575.62905	11.92934279	0.000201054	36.83316582
6	speak frankly	11.25060852	9119156.397	110164.7765	11.68850972	0.001811165	40.12248378
7	strictly obey	1.376527051	129868808.1	159033.0867	16.70072097	0.00040205	27.02366516
8	lie shamelessly	2.7956863	1752444.492	12080.04449	12.58758364	0.000214771	37.62892866
9	look fixedly	5.351319756	90326.7238	2810.990894	9.898864295	0.00012164	44.63117544
10	refer concretely	0.991285647	746355.6629	2663.466324	13.67528068	5.73099E-05	33.7226495
11	follow conscientiously	1.346639224	237928.324	3889.982964	11.12209033	1.95722E-05	37.3415899
12	strongly affect	4.527195885	57680017.54	367232.093	13.16849275	0.000725152	34.23473764
13	attack furiously	1.91681571	10271228.94	43271.06293	13.47247876	0.000190817	33.52137457
14	beat furiously	4.034827612	4828222.954	39094.9706	12.34300424	0.000379986	37.59319299
15	fall abruptly	4.294664397	7226086.313	85989.16707	11.71453804	0.000300354	37.37270337
16	cry copiously	0.990006101	1184023.019	3669.64093	13.91152317	6.73764E-05	33.05299017
17	punish strongly	1.144824792	140752872.3	423273.5142	15.42574742	0.000147563	26.73663924
18	prove completely	5.402698083	73501822.02	731196.0319	12.57491573	0.000726263	34.63122415
19	work jointly	6.091906354	1825951.603	45458.95583	10.40450505	0.0002423	40.57963314
20	use abusively	1.715461181	15388.8361	469.0947533	9.979074511	1.33068E-05	42.42075924
21	demonstrate completely	1.767784068	142155262.5	822334.6067	14.37879288	0.000210774	29.2667271
22	rise constantly	0.233074514	28881820.5	267504.5304	12.30192125	5.81249E-05	31.22771845

	Reference bigrams Reverso™	*t* test	χ^2	LLR	MI	Dice	UnigSub
1	attack savagely	2.203607962	4607248.876	18915.75568	13.47247876	0.000244744	35.21659001
2	hit hard	51.71724616	50625689.28	618132.7596	11.90631728	0.035896991	42.10150055
3	talk animatedly	8.177791564	305537.5237	3470.015823	11.46533417	0.000831751	44.85588256
4	greet enthusiastically	5.559836471	28119329.94	43924.31955	15.20910805	0.004530508	35.98657794
5	decide collectively	1.985413685	4186275.216	43364.26683	11.92934279	8.41354E-05	35.35115655
6	speak earnestly	6.038432593	1489399.612	17313.84417	11.68850972	0.00053331	40.75399852
7	strictly comply w	2.19262705	96011044.59	152205.6315	15.83475028	0.000742776	30.06694681
8	lie blatantly	2.587292809	2902144.148	20155.88978	12.58758364	0.000186966	36.60990168
9	look intently	12.72177185	759959.0018	24260.14914	9.898864295	0.000690148	44.31922073
10	refer specifically	9.758910029	57203283.15	265849.8311	13.67528068	0.004183469	36.61902734
11	follow to the letter	11.76497632	811417.3612	12214.32868	11.12209033	0.001357349	44.1978125
12	shock deeply	5.000148464	137004645.2	461710.6122	15.17118075	0.001680998	32.22152123
13	attack viciously	4.109107846	3716787.142	15071.751	13.47247876	0.000835586	38.00874994
14	beat badly	11.07033452	41823007.9	400535.1869	12.34300424	0.002424966	37.65865408
15	fall dramatically	15.16229338	13823423.94	167323.457	11.71453804	0.0032633	40.42961002
16	cry uncontrollably	6.474782615	4663791.37	14246.27235	13.91152317	0.002788382	39.36341165
17	penalise heavily	3.579243427	200488718.4	480656.4224	17.51943376	0.001284395	28.66050054
18	prove conclusively	9.581034604	2033700.28	13344.79993	12.57491573	0.002446743	42.04783085
19	work together	90.48378076	64469100.41	2047124.146	10.40450505	0.03642441	43.25688395
20	use improperly	6.943718254	727852.4716	22841.61184	9.979074511	0.000225518	42.32123063
21	demonstrate conclusively	5.379743008	6599333.466	15434.82083	14.37879288	0.002638522	38.16200054
22	rise steadily	17.76513285	14786771.73	125220.3256	12.30192125	0.00661212	40.84844955