Un environnement générique et ouvert pour le traitement des expressions polylexicales : de l'acquisition aux applications

Carlos Ramisch

▶ **To cite this version:**

Carlos Ramisch. Un environnement générique et ouvert pour le traitement des expressions polylexicales : de l'acquisition aux applications. Autre [cs.OH]. Université de Grenoble; Universidade Federal do Rio Grande do Sul (Porto Alegre, Brésil), 2012. Français. NNT : 2012GRENM059 . tel-00741147v2

HAL Id: tel-00741147

https://theses.hal.science/tel-00741147v2

Submitted on 9 Sep 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR ÈS SCIENCES
DÉLIVRÉ PAR L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique**

Arrêté ministériel : 7 août 2006

Présentée par

Carlos Ramisch

Thèse dirigée par **Christian Boitet**
et codirigée par **Aline Villavicencio**

préparée au sein **LIG-GETALP**
et de **EDMSTII**

A generic and open framework for multiword expressions treatment : from acquisition to applications

Thèse soutenue publiquement le ,
devant le jury composé de :

Yves Lepage
Université de Waseda (Japon), Président
Eric Wehrli
Université de Genève (Suisse), Rapporteur
Gaël Dias
Université de Caen (France), Rapporteur
Rosa Vicari
Universidade Federal do Rio Grande do Sul (Brésil), Examinatrice
Renata Vieira
Pontífice Universidade Católica do Rio Grande do Sul (Brésil), Examinatrice
Helena de Medeiros Caseli
Universidade Federal de São Carlos (Brésil), Examinatrice
Christian Boitet
Université Joseph Fourier (France), Directeur de thèse
Aline Villavicencio
Universidade Federal do Rio Grande do Sul (Brésil), Co-Directeur de thèse
Mathieu Mangeot
Université de Savoie (France), Invité

ACKNOWLEDGEMENTS

First of all, for their patience and support, I would like to thank my advisers Aline Villavicencio and Christian Boitet, who were always ready to answer my questions, discuss my research, and are true inspirations for me. Without them, I would not have been able to arrive this far. Thank you for sharing with me your knowledge and your experience on this stimulating and challenging field of computational linguistics. Your recognition has been of capital importance for keeping me motivated and focused in my work.

During the last three years, my work has been funded by a Ph.D. contract of the French Ministry of High Education and Research. I could never thank enough the people who gave me this opportunity to work on a subject I am passionate about in such an amazing research environment. I am thankful to the Brazilian agency CAPES and the French committee COFECUB, for funding a significant part of my trips between Brazil and France through the CAMELEON project (CAPES-COFECUB 707-11). Additionally, my work counted on the support of: the *Laboratoire d'Informatique de Grenoble* (LIG), the *Programa de Pós-Graduação em Computação* of the Federal University of Rio Grande do Sul (PPGC-UFRGS), the *Centre d'Initiation à l'Enseignement Supérieur (CIES) Grenoble* and the Franco-Brazilian laboratory LICIA. I would like to thank the administrative staff of these institutions, particularly Elisiane da Silveira Ribeiro (PPGC-UFRGS) and Amélie Vazquez (LIG). Both have been extremely comprehensive and always helped me find the way out in the international bureaucratic labyrinth of working simultaneously in two universities.

This thesis is not as much a personal achievement as it is the result of a collective effort. Therefore, I am thankful to all the people who directly or indirectly contributed to the achievement of this work, in particular the co-authors of papers: Magali Sanches Duran, Valia Kordoni, Evita Linardaki, Mathieu Mangeot, Cassia Trojahn dos Santos, Renata Vieira, Sandra Maria Aluísio, Maria José Finatto, Roger Granada, Marco Idiart, and Helena de Medeiros Caseli. I would also like to thank Eric Wehrli and Gaël Dias for accepting to be rapporteurs of my thesis, as well as the other members of the jury. I thank the anonymous reviewers of my papers, who provided constructive feedback which helped improve my work.

These three years as a Ph.D. student would not have been so fun if it was not for my colleagues from PLN-UFRGS, from GETALP and from the MWE community. In particular, I had many interesting discussions and obtained valuable suggestions from: Laurent Besacier, Emmanuelle Esperança-Rodier, Paulo Schreiner, Rodrigo Wilkens, Valérie Bellynck, Sara Stymne, Lucia Specia, Agata Savary, Violeta Seretan, Dimitra Anastasiou, Preslav Nakov, Paul Cook, Kyo Kageura and Francesca Bonin. I would also like to thank the *Chercheurs d'hOrizons* teammates: Antoine Bailly, Anne-Sophie Drouet, Astrid Lamberts, Marie Le Guen and Isabelle Joncour. For his help and guidance during my first steps

as a teacher, I am thankful to my teaching adviser Michel Burlet.

Many improvements in the `mwetoolkit` would not have been possible without the work of the students Sandra Castellanos, Maitê Dupont and Vitor De Araujo. I am specially indebted to Vitor for his programming skills, perspicacity, efficiency and excellent work as a developer and researcher. I would like to express my gratitude to all the anonymous users who downloaded and used the `mwetoolkit` around the world, and also to those non-anonymous ones who provided me with valuable feedback, in particular: Spence Green, Julien Corman, Agnès Tutin, Olivier Kraif, Cleci Bevilacqua, Anna Maciel and Guilherme Pilotti.

For their methodical and careful work as a proofreaders, I thank my sister Renata Ramisch and my partner Antoine Gay. Furthermore, along with the other members of my family and with my friends from Porto Alegre and Grenoble, they have been extremely patient, supportive and understanding every time I had to give priority to my work over my personal life, particularly during the last months of thesis writing. To all those who never stopped believing in my work, I say *thank you*, *merci beaucoup*, *muito obrigado*!

Carlos Ramisch

LIST OF ABBREVIATIONS AND ACRONYMS

AM	Association measure
CP	Complex predicate
DTD	Document type definition
GS	Gold standard
LM	Language model
LNRE	Large number of rare events
LSF	Lexico-semantic function
LVC	Light verb construction
MAP	Mean averaged precision
MLE	Maximum likelihood estimation
MTT	Meaning-text theory
MWE	Multiword expression
MWT	Multiword term
NLP	Natural language processing
POS	Part of speech
P	Precision
PV	Phrasal verb
R	Recall
SVC	Support verb construction
SVM	Support vector machine
TP	True positive
VPC	Verb-particle construction
XML	Extended markup language

Applications

IR	Information retrieval
MT	Machine translation
OCR	Optical character recognition
PB-SMT	Phrase-based statistical (or empirical) machine translation
SMT	Statistical (or empirical) machine translation
SRL	Semantic role labelling
WSD	Word sense disambiguation

Association measures

`dice`	Dice's coefficient
`ll`	Log-likelihood ratio
`mle`	Maximum likelihood estimator
`pmi`	Pointwise mutual information
`t-score`	Student's *t* test statistic

Conferences

ACL	Annual Meeting of the Association for Computational Linguistics
COLING	International Conference on Computational Linguistics
EACL	European Chapter of the Association for Computational Linguistics
LREC	Language Resources and Evaluation Conference
NAACL	North American Chapter of the Association for Computational Linguistics

Corpora

BNC	British National Corpus
PLN-BR	Corpus of the project Processamento de Linguagem Natural — Brasil
EP	Europarl Corpus

Language codes

`el`	Greek
`en`	English
`fr`	French
`pt`	Portuguese
`pt-BR`	Brazilian Portuguese

Parts of speech

CC	Conjunction
DT	Determiner
J	Adjective
N	Noun
P	Preposition
R	Adverb
V	Verb

Symbols

*	Ungrammatical construction
?	Unnatural construction

LIST OF FIGURES

Figure 3.1: Rank plot of the vocabulary of BNC-frg, with counts in descending order. . . . 39
Figure 3.2: Tokens in the corpus versus types in the vocabulary. . . . 40
Figure 3.3: Example of suffix tree. . . . 43
Figure 3.4: Example of suffix array. . . . 44

Figure 5.1: Framework for MWE acquisition from corpora, core modules in a prototypical acquisition chain. . . . 86
Figure 5.2: Pattern 1 matches noun phrases, pattern 2 matches sequences N_1 P N_1. 88
Figure 5.3: Example of MWT candidates extracted from the Genia corpus. . . . 95
Figure 5.4: XML fragment describing a MWT candidate extracted from the Genia corpus with `mwetoolkit`. . . . 97
Figure 5.5: Quality of candidates extracted from medium corpus, comparison across languages/MWE types. . . . 101
Figure 5.6: Time (seconds, log scale) to extract `en` nouns (bold line) and verbs (dashed line) from corpora. . . . 104

Figure 6.1: Excerpt of Greek EP from 17/12/1999. . . . 111
Figure 6.2: Tagger output containing surface form, lemma and simplified POS tag. 111
Figure 6.3: XML file containing the description of the relevant POS patterns for extraction. . . . 112
Figure 6.4: Extract of the XML output file with MWE candidates and their AM scores. . . . 112
Figure 6.5: Precision based on the EP counts. . . . 114
Figure 6.6: Quality comparison of threshold values. . . . 122
Figure 6.7: Distribution of verbs involved in CPs, pattern V + N + P. . . . 125
Figure 6.8: Distribution of verbs involved in CPs, pattern V + P + N. . . . 126
Figure 6.9: Distribution of verbs involved in CPs, pattern V + DT + N + P. . . . 126
Figure 6.10: Distribution of verbs involved in CPs, total number of CPs (all patterns). 129
Figure 6.11: Number of sentiment nouns (*y* axis) that prefer (dark bars) and accept (light bars) each pattern. . . . 131
Figure 6.12: Number of nouns (*y* axis) vs number of patterns accepted (*x* axis). . . . 131

Figure 7.1: PV semantics and quality scores per system. Scores are (1) good, (2) acceptable and (3) incorrect translation. . . . 149

LIST OF TABLES

Table 1.1: Examples of empirical MT errors due to MWEs. 5

Table 3.1: Statistics of BNC-frg — Sample of 20,000 random sentences taken from the BNC 37
Table 3.2: Counts of the 30 most frequent tokens in BNC-frg. 38
Table 3.3: Top-15 most frequent *n*-grams in BNC-frg. 45
Table 3.4: Contingency table for two random variables 47
Table 3.5: Top-15 *n*-grams (2 to 5) extracted from BNC-frg and ranked according to AMs. 49

Table 5.1: Performance of the `mwetoolkit` considering (a) no filtering threshold, (b) a threshold of $t = 1$ occurrence and (c) a threshold of $t = 5$ occurrences. 96
Table 5.2: Number of sentences and of words of each fragment of the Europarl corpus in `fr` and in `en`. 99
Table 5.3: Dimensions of the reference gold standards used and of the respective number of entries that occur at least twice in the S, M and L corpora. 100
Table 5.4: (P)recision and (R)ecall of `en` nominal candidates, comparison across corpus sizes: (S)mall, (M)edium and (L)arge. 101
Table 5.5: Intersection of the candidate lists extracted from the medium corpus. . 102
Table 5.6: Mean average precision of AMs in the large corpus. 103
Table 5.7: Summary of tools for MWE acquisition. 103

Table 6.1: Example sentences in Greek where MWEs can be at the root of translation problems. 109
Table 6.2: Inter-annotator agreement for each of the 4 categories and each evaluated AM in 2 corpora. 114
Table 6.3: Number of candidates extracted from the corpus and analysed. 123
Table 6.4: Number of candidates extracted and validated per pattern. 129
Table 6.5: Distribution of sentiment nouns according to their polarity. 130
Table 6.6: Distribution of sentiment nouns according to their source. 130

Table 7.1: Example of phrase table. 139
Table 7.2: Statistics of PVs in test corpus. 148
Table 7.3: Translation of PVs in sentences — automatic evaluation. 148
Table 7.4: Translation of PVs — human evaluation. 149
Table 7.5: Translation of idiomatic PVs — human evaluation. 151

CONTENTS

Abstract xi

1 Introduction 1
1.1 Motivations 1
1.1.1 What are multiword expressions? 1
1.1.2 Why do they matter? 3
1.1.3 What happens if we ignore them? 4
1.2 Thesis contributions 6
1.2.1 Scientific scope 6
1.2.2 Research questions 7
1.3 Thesis structure 10
1.3.1 Published work 10
1.3.2 Chapters outline 11
1.4 Summary 13

I Multiword expressions: a tough nut to crack **14**

2 Definitions and characteristics 15
2.1 Contextualisation 15
2.1.1 A brief history 16
2.1.2 MWEs in current language technology 19
2.1.3 Further reading 20
2.2 Defining MWEs 21
2.2.1 The MWE jungle 21
2.2.2 A practical definition 23
2.2.3 A note on MWEs and terms 24
2.3 Characteristics and characterisations 24
2.3.1 MWE properties 25
2.3.2 Existing MWE typologies 27
2.3.3 A simplified typology 29
2.4 Summary 31

3 State of the art 34
3.1 Elementary notions 34
3.1.1 Linguistic processing: analysis 35
3.1.2 Word frequency distributions 37
3.1.3 *N*-gram language models 40

3.1.4 Lexical association measures . . . 44
3.2 Practical context in MWE acquisition . . . 49
3.2.1 Methods for monolingual MWE acquisition . . . 50
3.2.2 Methods for bi- and multilingual MWE acquisition . . . 53
3.2.3 Existing tools . . . 54
3.3 Other MWE tasks . . . 57
3.3.1 Interpretation . . . 57
3.3.2 Disambiguation . . . 59
3.3.3 Representation . . . 61
3.3.4 Applications . . . 62
3.4 Summary . . . 64

II Automatic MWE acquisition 67

4 Evaluation of MWE acquisition . . . 68
4.1 Evaluation context . . . 68
4.1.1 Evaluation axes . . . 69
4.1.2 Evaluation measures . . . 71
4.1.3 Annotation . . . 73
4.2 Acquisition contexts . . . 75
4.2.1 Characteristics of target constructions . . . 75
4.2.2 Characteristics of corpora . . . 77
4.2.3 Existing resources . . . 79
4.3 Discussion . . . 79
4.4 Summary . . . 81

5 A framework for MWE acquisition . . . 84
5.1 Processing overview . . . 84
5.1.1 Goals and guiding principles . . . 84
5.1.2 Modules . . . 86
5.1.3 Discussion . . . 92
5.2 A case study . . . 94
5.2.1 Candidate extraction . . . 94
5.2.2 Candidate filtering . . . 96
5.2.3 Results . . . 96
5.3 Comparison with related approaches . . . 98
5.3.1 Related approaches . . . 98
5.3.2 Experimental setup . . . 99
5.3.3 Comparison results . . . 100
5.4 Summary . . . 104

III Application-oriented evaluation 107

6 Application 1: lexicography . . . 108
6.1 A dictionary of nominal MWEs in Greek . . . 108
6.1.1 Greek nominal MWEs . . . 108
6.1.2 Materials and methods . . . 111
6.1.3 Results . . . 113

6.2 Acquisition and analysis of Portuguese complex predicates 117
6.2.1 Portuguese complex predicates 117
6.2.2 Materials and methods 121
6.2.3 Results 123
6.2.4 Discussion 132
6.3 Summary 133

7 Application 2: empirical machine translation 135
7.1 A brief introduction to empirical MT 136
7.1.1 Preprocessing a parallel corpus 137
7.1.2 Learning a translation model 138
7.1.3 Decoding 139
7.1.4 Evaluating 140
7.2 MWEs and SMT 141
7.3 Integration into a PB-SMT system 143
7.3.1 Phrasal verbs in English 143
7.3.2 Experimental setup 145
7.3.3 Results 147
7.3.4 Discussion 151
7.4 Summary 152

8 Conclusions 154
8.1 Thesis achievements 154
8.2 Ongoing experiments 155
8.2.1 MWEs and MT 155
8.2.2 CAMELEON project 156
8.3 Perspectives 157

References 159

APPENDIX A Résumé étendu 182

APPENDIX B Resumo estendido 197

APPENDIX C Further reading: MWE acquisition 213

APPENDIX D Resources used in the experiments 216

APPENDIX E The `mwetoolkit`: Documentation 218

APPENDIX F The web as a corpus 227

APPENDIX G Detailed lexicon descriptions 231

ABSTRACT

The treatment of *multiword expressions* (MWEs), like *take off*, *bus stop* and *big deal*, is a challenge for NLP applications. This kind of linguistic construction is not only arbitrary but also much more frequent than one would initially guess. This thesis investigates the behaviour of MWEs across different languages, domains and construction types, proposing and evaluating an integrated methodological framework for their acquisition.

There have been many theoretical proposals to define, characterise and classify MWEs. We adopt generic definition stating that MWEs are word combinations which must be treated as a unit at some level of linguistic processing. They present a variable degree of institutionalisation, arbitrariness, heterogeneity and limited syntactic and semantic variability. There has been much research on automatic MWE acquisition in the recent decades, and the state of the art covers a large number of techniques and languages. Other tasks involving MWEs, namely disambiguation, interpretation, representation and applications, have received less emphasis in the field.

The first main contribution of this thesis is the proposal of an original methodological framework for automatic MWE acquisition from monolingual corpora. This framework is generic, language independent, integrated and contains a freely available implementation, the `mwetoolkit`. It is composed of independent modules which may themselves use multiple techniques to solve a specific sub-task in MWE acquisition. The evaluation of MWE acquisition is modelled using four independent axes. We underline that the evaluation results depend on parameters of the acquisition context, e.g., nature and size of corpora, language and type of MWE, analysis depth, and existing resources.

The second main contribution of this thesis is the application-oriented evaluation of our methodology proposal in two applications: computer-assisted lexicography and statistical machine translation. For the former, we evaluate the usefulness of automatic MWE acquisition with the `mwetoolkit` for creating three lexicons: Greek nominal expressions, Portuguese complex predicates and Portuguese sentiment expressions. For the latter, we test several integration strategies in order to improve the treatment given to English phrasal verbs when translated by a standard statistical MT system into Portuguese.

Both applications can benefit from automatic MWE acquisition, as the expressions acquired automatically from corpora can both speed up and improve the quality of the results. The promising results of previous and ongoing experiments encourage further investigation about the optimal way to integrate MWE treatment into other applications. Thus, we conclude the thesis with an overview of the past, ongoing and future work.

An extended abstract in Portuguese and in French is available in Appendices A and B.

Keywords: Natural language processing, computational linguistics, multiword expressions, lexical acquisition, machine translation, lexicography, corpus linguistics.

1 INTRODUCTION

This thesis is about multiword expressions (MWEs) and their importance for natural language processing (NLP) applications. This is a hard and open problem in NLP research, due to the complex nature of MWEs. In this chapter, we firstly motivate the importance of MWEs for NLP applications (Section 1.1). Secondly, we discuss the scientific scope, original contributions and goals of the work (Section 1.2). Thirdly, we provide an overview of the structure of this document and of previously published work (Section 1.3).

1.1 Motivations

Before we dig into detailed problems, data and their complexities, we would like to informally discuss the answers to the following three questions: what are MWEs, why do they matter and what happens if we ignore them?

1.1.1 What are multiword expressions?

The question of what counts as a multiword expression and what does not is a polemic one, and in Section 2.2 we will provide a set of rigorous formal definitions for the term. But for the moment, let us put the technical details aside and assume that all we need to know is that, put simply, MWEs are habitual recurrent word combinations of everyday language (Firth 1957). For example, when we say that someone *sets the bar high*, we use it as a metaphor to say that his/her rivals will have a hard time trying to beat him/her. There is actually no physical bar positioned in a higher position, so the meaning of the expression cannot really be guessed from the meanings of the individual words if someone is not familiar with that particular expression. This is one of the most prototypical examples of MWEs: idiomatic expressions. Analogously, a *point of view* is not a good place to take a picture, a *loan shark* is not a fish, *by the way* is not a place, *white trash* is not something that you should throw in the rubbish bin, *red wine* is actually purple, white wine is actually yellow, you can still walk when someone *stands on your feet*, you do not need a knife to *cut someone a break*, you do not need money to *buy someone some time*, there is not going to be more air available just because you *saved someone's breath*, what distinguishes a *French kiss* from other kisses is not the nationality of the kissers, an *open mind* is (fortunately) not open as a door would be open, and so on.

In addition to idiomatic expressions, many other constructions present some idiosyncrasies which would allow us to see them as MWEs. In order to recognise them, one can apply simple linguistic tests. For example, we can ask: is it possible to replace one word in the expression by a synonym? If we take the compound *full moon*, for instance, it

would be quite awkward if someone said *?entire moon*, *?total moon* or *?complete moon*. Although the meaning of the alternative forms can be easily understood and would seem natural if you are learning English, a native speaker would argue that "you do not say it like this", and then he/she would probably be incapable of telling you why. Therefore, *full moon* is a MWE with the characteristic of having arbitrarily fixed semantic variability. This makes MWEs quite hard for foreign language learners who lack experience of language use even though they master general lexical and syntactic rules. On the other hand, MWEs confer naturalness and fluency to the discourse, and are unconsciously used as markers that help spotting non-native speech in dialogue contexts.

Another test for detecting MWEs is word for word translation into another language (see Table 1.1). If the translation sounds weird, unnatural or even ungrammatical, the original expression is probably a MWE. For example, in order to express the meaning of *prince charming* in Portuguese, one says *príncipe encantado*, that is, *enchanted prince*. Alternatives like *príncipe charmoso (good-looking prince)* and *príncipe encantador (gentle prince)* seem unnatural and funny. Similarly, the *finish line* is translated as the *arrival line* in Portuguese (*linha de chegada*) and in French (*ligne d'arrivée*). A MWE in one language can be translated as a simple word in other languages. For instance, *give up* translates as *renoncer* in French and as *desistir* in Portuguese, and *thank you* translates as *merci* in French and as *obrigado* in Portuguese. Such asymmetric MWEs will be discussed in Chapter 7. As the present thesis was developed in the context of an international French-Brazilian cooperation project, some examples throughout the text will be given in French and in Portuguese. However, in order to keep the reading as accessible as possible, we will provide preferably English examples. All examples are emphasised as *italic text*, ungrammatical constructions are preceded by a star * and unnatural constructions are preceded by a question mark *?*.

Some MWEs have the singularity of breaching general language rules. For instance, time adverbs cannot, in theory, be quantified or used as interval extremities. However, it is possible to say *every now and then* and *from time to time* meaning *eventually*. Analogously, the preposition *on* (when it is not acting as a particle in a phrasal verb) requires a complement, but expressions like *from now on* and *and so on* do not respect this constraint. Also, the expression *truth be told* corresponds to a very unorthodox use of English syntax, but the equivalent "correct" expressions *truth has to be told* or *truth should be told* would not have the same meaning.

Many common names are also examples of MWEs. For instance, the tool used to suck the air and catch the dirt on the floor is a *vacuum cleaner*, a key that opens all doors is a *master key*, an automatic recorder that answers the phone when you are not there is a *voice mail* or an *answering machine*, a shoe that has a protuberance under the ankle is a *high heel shoe*, a character at the end of an interrogative sentence is a *question mark*. Sometimes, the words in the expression are collapsed and form a single word. This is the systematic behaviour in German (and other Germanic languages), but it happens sometimes in English as well (*firearm*, *honeymoon*, *sleepwalk*, *lighthouse*). MWEs in which the words are concatenated together form a single typographic word and are therefore not the main focus of our work.

Some actions require verbal MWEs in order to be expressed. Hence, there is no simple verb to express exactly the same meaning as the verbal expressions such as *make sense*, *take advantage of someone*, *have something to do with something*, *get involved*, *take for granted*, *have the last word*, *put in place*. On the other hand, some expressions do have an (almost) equivalent single-verb paraphrase, and their use might depend on the context

or simply on the speaker's intuition. Examples of such expressions are *give a wave* = *to wave*, *take a walk* = *to walk* and *take a shower* = *to shower*.

1.1.2 Why do they matter?

MWEs are very frequent in everyday language. Native speakers rarely realise it, but colloquial speech is full of formulaic expressions such as *good morning*, *my bad*, *too bad*, *what the hell* and *bye bye*. For instance, almost all the examples of the previous section were taken from a 30-minutes episode of an American TV show, and there were many more that were not included here in order to keep the text concise. Researchers in theoretical and computational linguistics evaluated the recurrence of MWEs in a more systematic way than we did. There are several publications which provide examples and figures proving how frequently MWEs occur in text collections across different languages and domains (Biber et al. 1999). It is often assumed that a native speakers' lexicon contains as many MWEs as simple words (Jackendoff 1997). Thus, any computational system dealing with human language must take MWEs into account.

In the present thesis, instead of discussing the recurrence of MWEs, we chose to present another argument, hopefully more convincing, of the importance of MWEs in natural language. That is, we analyse the impact of MWE treatment in a large although non-exhaustive list of NLP tasks and applications. The following list presents some NLP tasks and applications that will generate ungrammatical or unnatural output if they do not handle MWEs correctly.

- **Computer-aided lexicography.** Lexicographers are professionals who design and build lexical resources such as printed and machine-readable dictionaries and thesauri. Building a lexical resource is a very onerous task that demands expert knowledge and takes a lot of time. If writing a dictionary for single words is costly, dictionaries containing MWEs are even more complex and require more effort. However, as MWEs are often the source of difficulties for both human and machines to process a sentence, it is very important that lexical resources include them. One of the seminal papers in the MWE field is the work by Church and Hanks (1990). They use a lexicographic environment as their evaluation scenario, comparing manual and intuitive research with the automatic association ratio they propose. They show that tools used to support lexicographic work should also help identify MWEs and extract their meaning and syntactic behaviour from texts. We explore this application further in Chapter 7.
- **Optical character recognition (OCR).** If an OCR system recognises with equal probabilities the words *farm* and *form* in *federal farm/form credit*, it can chose the word that most likely occurs as part of this MWE (Church and Hanks 1990). Currently, this is performed using *n*-gram language models, but *n*-grams fail to model highly flexible expressions like **take** *patient risk factors and convenience* **into account**. Therefore, MWEs could help improve OCR technology, depending on the length of the *n*-gram.
- **Word sense disambiguation (WSD).** MWEs tend to be less polysemous than the composition of the senses of the individual words in it. Finlayson and Kulkarni (2011) exemplify that the word *world* has 9 senses in Wordnet 1.6, *record* has 14, but *world record* has only 1. We discuss the importance of MWEs for WSD and for other semantic tasks in Section 3.3.4.2. Additionally, we discuss in Section 6.2.1.1 the importance of MWEs for a related task, that is, the annotation of semantic role labels.

- **Part-of-speech (POS) tagging and parsing.** Recent work in parsing and POS tagging indicates that MWEs can help removing syntactic ambiguities. For instance, the French expressions *faire une marche à pied* (lit. *make a walk by foot*) and *faire un verre à pied* (lit. *make a glass by foot*) are syntactically identical. However, in the first case the adverbial complement is attached to the verb *faire* while, in the second example, the complement is attached to the noun *verre*, as it corresponds to the MWE *verre à pied (wine glass)*. The integration of MWEs into POS taggers and parsers is discussed further in Section 3.3.4.1.
- **Information retrieval (IR).** When a user queries a web search engine like Google for *rock star*, he/she is probably not looking for websites containing geological descriptions of *rocks* nor astronomy websites about *stars*. Therefore, intuitively, when MWEs like *rock star* are indexed as a unit in the system, its accuracy improves on multiword queries. This hypothesis has been validated by several related articles discussed in Section 3.3.4.3
- **Foreign language learning.** MWEs are hard for non-native speakers learning a foreign language. Dictionaries and other lexical resources containing MWE entries can be very useful to avoid common mistakes. Examples of such dictionaries for the English language include the Cambridge International Dictionary of Idioms and the COLLINS-COBUILD Dictionary of Phrasal Verbs. Therefore, MWEs play an important role in the design of computer systems for foreign language *e*-learning.
- **Machine translation (MT).** MWEs have been a concern of MT system designers from the very beginning. Often, MWEs cannot be translated word for word, and should be represented as units in the translation model. Traditional expert systems usually include dictionaries of phrases and expressions, that are looked up before performing compositional transfer. Current empirical MT systems tend to represent bilingual word sequences instead of bilingual words, thus representing a larger context that is a simplified representation of syntax and MWEs. To date, MWEs remain a challenging problem for automatic translation, independently of the MT paradigm. We discuss the relation between MT and MWEs in Section 7.2, and present some experimental results in Section 7.3.

Despite the importance of MWEs, they are often neglected in the construction of NLP applications. In 1993, Smadja pointed out that, in automatic MWE acquisition, "...the collocations [MWEs] retrieved have not been used for any specific computational task" (Smadja 1993, p. 150). Most of the recent and current academic research in the community still focuses on identification and extraction tasks instead of focusing on the integration of automatically acquired or manually compiled MWE resources into applications. That is, academic research is still slowly starting to investigate MWEs in applications, and there is a gap between industrial language technology and academic research in this field. This is one of the motivations for the work presented here.

1.1.3 What happens if we ignore them?

Taking MWEs into account is important to confer naturalness to the output of NLP systems. An MT system, for instance, needs to be aware of idiomatic expressions like *it is raining cats and dogs* to avoid literal translations. The equivalent expressions in French would be *il pleut des cordes* (lit. *it rains ropes*), in German *es regnet junge Hunde* (lit. *it rains young dogs*), in Portuguese *chove canivetes* (lit. *it rains Swiss knives*), and so on. Likewise, a parser needs to deal with verb-particle constructions like *take off from Paris* and with light verb constructions like *take a walk along the river*, in order to avoid

$\texttt{en}_{\text{SRC}}$	*I* **paid** *my poor parents* **a visit**
$\texttt{pt}_{\text{MT}}$	*Eu* **pago** *os meus pais pobres* **uma visita**
$\texttt{pt}_{\text{REF}}$	*Eu* **fiz uma visita** *aos meus pobres pais*
$\texttt{fr}_{\text{MT}}$	*J'ai* **payé** *mes pauvres parents* **une visite**
$\texttt{fr}_{\text{REF}}$	*J'ai* **rendu visite** *à mes pauvres parents*
$\texttt{en}_{\text{SRC}}$	*Students pay* **an arm and a leg** *to park on campus*
$\texttt{pt}_{\text{MT}}$	*Estudantes pagam* **um braço e uma perna** *para estacionar no campus*
$\texttt{pt}_{\text{REF}}$	*Estudantes pagam* **os olhos da cara** *para estacionar no campus*
$\texttt{fr}_{\text{MT}}$	*Les étudiants paient* **un bras et une jambe** *pour se garer sur le campus*
$\texttt{fr}_{\text{REF}}$	*Les étudiants paient* **les yeux de la tête** *pour se garer sur le campus*
$\texttt{en}_{\text{SRC}}$	*It shares the* **translation-invariance and homogeneity properties** *with the central moment*
$\texttt{pt}_{\text{MT}}$	*Ele compartilha a* **tradução invariância e propriedades de homogeneidade** *com o momento central*
$\texttt{pt}_{\text{REF}}$	*Ele compartilha as* **propriedades de invariância por translação e de homogeneidade** *com o momento central*
$\texttt{fr}_{\text{MT}}$	*Il partage la* **traduction-invariance et propriétés d'homogénéité** *avec le moment central*
$\texttt{fr}_{\text{REF}}$	*Il partage les* **propriétés d'invariance par translation et d'homogénéité** *avec le moment central*

Table 1.1: Examples of empirical MT errors due to MWEs.

PP-attachment errors.

For the empirical MT system used in the examples of Table 1.1, [1] a MWE is any sequence of words which, when not translated as a unit, generates errors. Possible problems include ungrammatical or unnatural verbal constructions (sentence 1), awkward literal translations of idioms (sentence 2) and problems of lexical choice and word order in specialised texts (sentence 3). These anecdotal translation examples clearly demonstrate how awkward the resulting sentences produced by an empirical MT system are when compared to the reference expected translations produced by humans. In an expert MT system, MWEs would typically be represented as a unit in the lexicon, otherwise the same errors may occur. More generally, in numerous other NLP applications, when the words composing a MWE are treated as separate units, this can induce the system to produce erroneous output.

We can summarise the importance of MWEs for NLP applications as a consequence of the following:

– important information can be lost if MWEs are not treated;
– MWEs confer naturalness to a system's output; and
– they are very frequent and pervasive in language, and are very likely to occur in texts to be processed.

Taking MWEs into account can be quite complicated for traditional NLP applications. The usual or conventional way of saying things, that is, the natural tendency that words have of attracting each other, is the key phenomenon behind the concept of MWE.

1. Source in English ($\texttt{en}_{\text{SRC}}$) from the web. Automatic translations (MT) in Portuguese (pt) and in French (fr) by Google Translate (http://translate.google.com/) on 2012/05/16. References (REF) by native speakers.

However, this phenomenon lies in a fuzzy zone between the lexicon and the syntax of a language, thus constituting a real challenge for NLP systems. In Section 2.3.1, we will discuss some idiosyncrasies of MWEs that make them a wild animal to tame. This is the perfect scenario for a paradox: it is at the same time difficult and necessary to deal with MWEs in applications that involve some degree of semantic interpretation of natural language.

1.2 Thesis contributions

With respect to related work presented in Chapters 2 and 3, the work presented in this thesis has several important differences. These constitute original contributions to the field of computational linguistics and, more specifically, to the academic community working on MWE treatment. In this section, we list some of our intended contributions.

1.2.1 Scientific scope

Given that the creation of language resources is an onerous task, NLP researchers have proposed techniques and tools that aid in the automatic creation and exploitation of monolingual and multilingual resources, helping linguists and domain experts to speed up lexicographic work (Preiss et al. 2007, Messiant et al. 2008). Nonetheless, when it comes to MWEs, the availability of such tools is still quite limited both in terms of effectiveness and of applicability to languages and language pairs, contrasting with the ubiquitous and pervasive nature of MWEs. Therefore, there is a need for developing, consolidating and evaluating techniques for the automatic acquisition of MWEs from corpora.

The adequate treatment of MWEs in NLP applications is an open and challenging problem. Therefore, we present our contributions as trying to answer the following questions:

- How can we *acquire* MWEs automatically in monolingual and multilingual contexts?
- How can we *estimate whether a given method for automatic MWE acquisition is useful*?
- What is the best way to *represent* MWEs in machine-readable resources?
- How can we *integrate MWEs into real applications*, specially into multilingual ones?

This thesis addresses the problem of MWE treatment in NLP applications, ranging from their automatic acquisition in raw text to their integration into two real-life applications: computer-aided lexicography and empirical MT. For the sake of simplicity, we focus on the two most frequent broad classes of MWEs: noun compounds and verbal expressions. We have developed a conceptual model for the pipeline of MWE treatment, as well as a concrete software framework that validates the proposed methodology. We have evaluated this model thoroughly and systematically. The hypotheses that guide our work can be summarised as:

- Mono- and multilingual (parallel and comparable) corpora are rich information sources for automatic lexical acquisition.
- A combination of techniques can be used as a basis for automatic MWE extraction, and will perform better than a single one of them.
- It should be possible to extract MWEs automatically for poorly resourced languages, not only from English and a few "main" languages.
- The evaluation of MWE acquisition is a research topic on its own, and designing an

evaluation scenario is as important as designing an acquisition method.
– The adequate integration of MWE treatment can improve the performance of NLP applications in terms of their linguistic quality, helping to generate more natural results and remove ambiguities in analysis.
– Different NLP applications and integration strategies will yield different performance improvements.

1.2.2 Research questions

We believe that MWE research has finally reached its maturity, and is becoming a consolidated research field. Therefore, the time has come to move forward from acquisition methods to their integration into NLP applications, for higher linguistic quality (more natural and fluent) results. In this context, the present work has three main goals:

1. We would like to develop generic and portable techniques for automatic MWE acquisition from corpora.
2. We would like to evaluate these techniques extrinsically, that is, by measuring their usefulness in real NLP applications.
3. We would like to investigate these tasks in bi- and multilingual contexts, studying how different parameters of the acquisition context, such as language, domain, type of expression and data sources, influence the quality of automatically acquired MWEs.

The motivation behind the first goal is that, currently, many punctual techniques and tools exist that focus on a small *well-defined* task. We believe that the time has come to systematise and unify these experimental approaches into a single and generic methodological framework. This framework has a proof-of-concept companion software tool, implementing all steps in the MWE acquisition pipeline (e.g., candidate generation, counting, filtering, sorting and evaluation), thus replacing other tools that only perform part of the processing. Available software and open evaluation campaigns are pointed out by Steedman (2008) as two factors that can help determine the maturity of a research field.

The second goal is motivated by our revision of the state of the art. We realise that the evaluation of automatic MWE acquisition techniques is a challenge on its own. Evaluation results depend on several factors like MWE type, corpus size, domain, existing lexical resources, among others. Therefore, our objective is twofold. On the one hand, we would like to perform a theoretical analysis of the evaluation of MWE acquisition, making explicit the axes that define an acquisition context, the evaluation measures, and the factors determining the generalisation of results. On the other hand, we would like to perform a systematic thorough evaluation of our proposed methodological framework in the context of real applications. We believe that intrinsic evaluation of acquisition results per se, as the final result of a process, can be interesting to compare several techniques and parameters. However, only extrinsic application-based evaluation can effectively prove the usefulness of MWE acquisition.

This brings us to the third goal of our research: multilingual applications. To date, there is little work on multilingual MWE acquisition and their integration into multilingual applications. Multilingualism is an important characteristic of the World Wide Web (WWW), which contains a very large amount of information expressed in natural language. Therefore, NLP systems dealing with web texts must be naturally multilingual and scalable. Thus, we would like to develop techniques for the acquisition of multi-

lingual MWEs from corpora and to evaluate their usefulness in multilingual systems, in particular for automatic translation.

The achievement of these three main goals constitutes an important and original contribution to NLP research, as we will detail later in Chapter 3. In order to achieve them, there is a number of guiding principles that we want to follow. These principles distinguish the current thesis from related work. Therefore, they constitute additional contributions that go beyond the three main aforementioned goals. In what follows, we enumerate some of these contributions that characterise and justify conceptual choices in our work.

1.2.2.1 Hybrid/mixed acquisition

To date, there is no agreement on an adequate method for acquiring MWEs. There has been much discussion about whether there is a single optimal method for acquiring any MWE type, a combination of methods, or if different methods work better for a given MWE type than for another. Therefore, one of our goals is to investigate the largest possible range of methods to automatically acquire MWEs from corpora, dissecting the influence of the different types of resources employed on the quality of the results. Our philosophy is that we do not want to elect *the* best technique for MWE acquisition, but to investigate a plethora of them, thus developing a naturally hybrid methodological framework which mixes several state-of-the-art techniques.

1.2.2.2 Integrated processing

Most of recent research on MWE treatment focused on their automatic identification and extraction from textual corpora. Some authors focus on the candidate extraction process from parsed text (Seretan 2008), others on the automatic filtering and ranking through association measures (Evert 2004, Pecina 2010). Nonetheless, few publications provide a whole picture of the MWE treatment pipeline. One of our contributions is to model the MWE acquisition as an integrated process, with modular tasks, each task having several techniques and parameters that can be optimised according to the target MWE types.

1.2.2.3 Generality

The methods developed in our work do not depend on a fixed length of MWE. Similarly, they do not depend on any adjacency assumption, as the words composing an expression might be several words away from each other in the corpus. The only constraint generally imposed is that word association does not cross sentence boundaries. This constraint could in theory be lifted, but it seems to make sense as MWEs do not split over more than one sentence. One limitation of our work is that we consider fixed word order, that is, differently from Church and Hanks (1990), Smadja (1993), we consider w_1w_2 as being different from w_2w_1, to distinguish cases like *to conform* from *conform to*. When acquiring non fixed collocations like *drastically drop*, it might be interesting to relax the order constraint. This, for the moment, falls out of the scope of our work.

1.2.2.4 Portability

Because the methods we developed are generic, they can be easily applied on virtually any language, MWE type and domain, not strictly depending on a given formalism

or tool.[2] It is possible to apply our methodology on a very large range of acquisition contexts. Intuitively, for a given language, if some preprocessing tools like POS taggers, lemmatisers and/or parsers are available and used to preprocess the input, the results will be better than those obtained by running our methods on raw corpus data. This is a hypothesis that remains to be proven in the future. Nonetheless, our software and methods were designed to be applicable even when no automatic analysis tool is available at all.

1.2.2.5 Customisation and scalability

One of the main advantages of the implementation of our methodology is that it is highly customisable. It was not designed as a push-button tool, that is, someone who is not familiar with the domain would not be able to use it without some prior training. As a counterpart, we allow for a large number of parameters to be tuned, and modules can be chained in several different ways. For instance, as opposed to similar tools, it is not necessary to work only with 2-grams, but working with arbitrarily long *n*-grams is possible. This does have some implications in terms of the association scores that can be calculated, but we leave this decision for the user: it is not taken a priori during the design of the methodology. This customisation allied with efficient methods to deal with large amounts of data constitutes an original contribution of our work.

1.2.2.6 Evaluation of MWE acquisition

Published results comparing MWE acquisition techniques usually evaluate them on small controlled data sets using objective measures such as precision, recall and mean average precision (Schone and Jurafsky 2001, Pearce 2002, Evert and Krenn 2005). On the one hand, the results of *intrinsic evaluation* are often vague or inconclusive: although they shed some light on the optimal parameters for the given scenario, they are hard to generalise and cannot be directly applied to other configurations. On the other hand, *extrinsic evaluation* consists of inserting acquired MWEs into a real NLP applications and evaluating the impact of this new data on the overall performance of the system. For instance, it is easier to ask a human annotator to evaluate the output of a MT system than to ask whether a given sequence of words constitutes a MWE. Thus, another original contribution of this thesis is application-oriented extrinsic evaluation of MWE acquisition on two study cases: computer-aided lexicography and empirical MT. Our goal is to investigate (1) how much MWEs impact on the performance and (2) what are the best ways of integrating them in the complex pipeline of the target application. In addition to evaluation results themselves, we propose a typology for MWE acquisition evaluation that classifies the evaluation context according to four orthogonal axes: according to the acquisition goals, nature of measures, available resources and type of MWE (see Section 4.1.1).

1.2.2.7 Available software tool and resources

In academic research, one often needs to re-implement techniques described in an article, most of the time because the described technique was either not implemented in a consistent software piece or because the software was not made available. Therefore, one

2. However, one of its limitations is that we do not deal with languages whose writing systems do not use spaces to separate words. Thus, when working with Chinese, Japanese, or even with German compounds, some preprocessing must be done in order to tokenise the words and insert artificial spaces between elementary linguistic words. These may not make much sense, for instance, in the case of German compounds. This languages are generally dealt with by some specialised methodology, which will probably perform better than our methods applied to artificially preprocessed data.

of our goals is to provide a usable and downloadable tool for MWE acquisition, analysis and evaluation. This tool, called `mwetoolkit`, is a practical and concrete contribution of our work. To the best of our knowledge, there are few similar tools available. These are compared to our tool in Section 5.3. The `mwetoolkit` covers a larger part of the MWE treatment pipeline, is extensible and open-source, and thus offers advantages over similar software.

1.2.2.8 Available lexical resources

As a by-product of our evaluation, we have generated some language resources that can support future research in MWE acquisition. The lexical resources resulting from the work described in Chapter 6 have been made available on the MWE community website. [3] They are, namely, a lexicon of nominal MWEs in Greek and a lexicon of complex predicates in Brazilian Portuguese. In addition, recent experiments that fall out of the scope of this thesis produced another resource: a version of the Childes corpus annotated with phrasal verbs (Villavicencio et al. 2012). This resource provides valuable data for cognitive research on the acquisition of verbal expressions by children. Since such resources are freely available, they can be extended, enriched and applied to develop NLP applications, foreign language *e*-learning systems, and can more generally be used as resources for ground research in MWE acquisition.

1.3 Thesis structure

Part of the work presented in this thesis has previously been published in conferences, workshops and journals. In most of them, the work was not carried out individually, but in collaboration with colleagues. Therefore, before describing the structure of the thesis content chapters, we first acknowledge previously published articles and their respective co-authors.

1.3.1 Published work

We have published articles on entropy-based methods for MWE acquisition (Villavicencio et al. 2007), on comparative evaluation of MWE acquisition techniques (Ramisch et al. 2008a), on the automatic acquisition of the semantics of English phrasal verbs (Ramisch et al. 2008b) and on multiword terminology extraction in the biomedical domain (Ramisch 2009). These early publications are not included in the present work, although they deal with related topics.

The `mwetoolkit` software platform, presented in Chapter 5, was developed in collaboration with Vitor De Araujo, Sandra Castellanos and Maitê Dupont. Furthermore, Christian Boitet and Aline Villavicencio gave useful advice on the methodological and conceptual design choices. Our methodology and tool was first presented to the academic community at the LREC 2010 conference (Ramisch et al. 2010c). Then, later the same year, we presented a demonstration of the tool at the COLING 2010 conference (Ramisch et al. 2010b). For the workshop MWE 2011, we published a report describing the improvements of the tool, in terms of both faster and more flexible acquisition (Araujo et al. 2011).

The comparative evaluation of the `mwetoolkit`, presented in Section 5.3, was published at the ACL 2012 student research workshop (Ramisch et al. 2012). The evaluation

3. `http://multiword.sf.net`

results in the context of computer-aided lexicography were published in three different articles. The first one was presented at the LREC 2010 workshop on the exploitation of multilingual resources and tools for Central and (South) Eastern European Languages (Linardaki et al. 2010). This paper describes the methodology employed in the creation of a lexical resource containing Greek MWEs. The second article was presented at the MWE 2011 workshop (Duran et al. 2011). It describes the creation of the CP-SRL dictionary, which contains complex predicates in Brazilian Portuguese aimed at a semantic role labelling task. The third article, presented at the 2011 conference on Corpus Linguistics, is an extension of the second (Duran and Ramisch 2011). It also describes the construction of a dictionary of complex predicates in Brazilian Portuguese. Nonetheless, the goal of this second dictionary, called CP-SENT, is distinct, since it was designed as a tool for supporting sentiment analysis and extraction of Brazilian Portuguese texts.

Experiments concerning syntax-based acquisition of MWEs were carried out on a corpus of child-directed speech transcriptions, looking at verbal expressions. This work, described in Villavicencio et al. (2012), constitutes an evaluation of the `mwetoolkit` patterns for syntactically flexible expressions. However, since it has a more cognitive bias, we decided not to include it in the thesis, as it would probably divert the user from the real focus of our work, which is the acquisition of MWEs and its evaluation.

In terms of bilingual MWE acquisition, previous publications explored the acquisition of bilingual English–Portuguese MWEs through the use of automatically aligned parallel corpora (de Medeiros Caseli et al. 2010). This method was further extended and evaluated in the context of domain-specific acquisition, considering bilingual multiword terms in Pediatrics (Ramisch et al. 2010a, Villavicencio et al. 2010). The comparison of the bilingual acquisition method with baseline monolingual acquisition methods, however, showed that coverage of the bilingual MWE lists acquired automatically is very limited. Therefore, for concision purposes, we decided not to include these results in the present thesis. Instead, we use existing bilingual resources to simulate the results of automatic acquisition in the experiments reported in Chapter 7. This has the advantage that acquisition errors are not propagated through the modules of the translation system, keeping the parameters of the experiment manageable. The insertion of automatically acquired MWEs is planned for future work, as described in Section 8.3.

Finally, we published two articles summarising and discussing some of the contributions of the present thesis. The first one, presented at the French workshop RECITAL 2012, focuses on the evaluation of the `mwetoolkit` in the context of computer-aided lexicography (Ramisch 2012b). The second paper, presented at the ACL 2012 student research workshop, is more concise and presents the main contributions of the thesis in terms lexicography and MT (Ramisch 2012a).

1.3.2 Chapters outline

This thesis is structured into eight chapters. Therefore, the reader should be able to quickly navigate through the chapters and locate parts most interesting to her/him. French and Portuguese translations of these summaries are provided in Appendix A and in Appendix B, thus allowing for speakers of these languages to grasp the main topics of the thesis.

In this first introductory chapter, we have presented the motivations of our work, its scientific scope and original contributions. Chapter 2 provides the common ground for our research. It starts with a historical review of the MWE field, exploring theoretical and computational perspectives in both academic and industrial research. Existing definitions

found in the literature for the term MWE vary from very generic ones to definitions covering a single aspect of the phenomenon. Therefore, we propose an application-oriented definition that can be instantiated according to the acquisition goal. MWEs are a recurrent and heterogeneous phenomenon, presenting varying degrees of syntactic and semantic fixedness. Thus, we present and discuss these characteristics and suggest a new typology based both on their morphosyntactic categories and on the difficulty to treat them in computational applications.

In the field of computational linguistics, MWEs have gained increasing popularity in the last decade. There is a vast body of published work witnessing this progression, and Chapter 3 is dedicated to drawing a state of the art of current MWE treatment techniques. Therefore, we start by introducing the fundamental concepts used in related work, including elementary notions of linguistic analysis, word frequency distributions, *n*-gram language models and association measures. Users familiar with such concepts might want to skip this first section. Afterwards, the remainder of the chapter is divided into two sections: firstly, we focus on MWE acquisition, and secondly, we briefly describe other tasks in MWE treatment like interpretation, disambiguation, representation and applications.

Chapter 4 is more abstract, and its goal is to persuade the reader about the difficult and challenging nature of the evaluation of MWE acquisition. We introduce a new classification that describes the evaluation context through four orthogonal axes. The measures such as precision and recall, as well as the annotation guidelines provided to the human judges, are described in detail in this chapter. Evaluation results depend on several factors such as corpus size, corpus nature, level of preprocessing, type of MWE, language, domain and existing resources. Thus, the evaluation is often difficult to interpret and generalise, making this one of the open problems in the field. This fact motivates the use of extrinsic rather than intrinsic evaluation, as described later in Chapters 6 and 7.

We present a methodological framework and a corresponding software tool called `mwetoolkit`, developed for the automatic acquisition of MWEs from corpora. Its modules are presented and discussed in Chapter 5, which can be complemented by the software documentation found on the website [4] and reproduced in Appendix E. In addition to a detailed description of the modules and of how they can be combined, we present a pedagogical experiment in which we go step by step, extracting MWEs from a toy corpus. There are some similar freely available tools, and we compare the `mwetoolkit` with them in terms of linguistic quality, use of computational resources and flexibility.

Thorough extrinsic evaluations of the proposed methodology are performed in Chapter 6 and 7. In the former, we evaluate it in the context of computer-aided lexicography, showing how it helps to create several lexical resources. These lexical resources are: a dictionary of nominal MWEs in Greek and two dictionaries of complex predicates in Brazilian Portuguese, one aimed at semantic role labelling (CP-SRL) and the other aimed at automatic sentiment analysis (CP-SENT).

We start Chapter 7 with a brief review of empirical methods for automatically learning a translation model from a parallel corpus. Again, users familiar with empirical MT can skip this first section. The next section provides a discussion of current approaches to MWEs in existing expert and empirical MT systems. Then, we show and analyse the preliminary results of our experiments on the integration of verbal expressions into an English–Portuguese empirical MT system. We chose this language pair in order to study the asymmetries that arise when MWEs in one language (phrasal verbs in English) are translated as simple words (single verbs in Portuguese). At the end of this more practical

4. `http://mwetoolkit.sf.net`

chapter, we discuss further directions for our ongoing experiments.

At the end of the thesis, the reader will find in Chapter 8 the conclusions which summarise the achievements of this work and future experiments that we intend to perform. The promising results found in the present thesis allow us to discuss the future perspectives and the long-term goal of this research.

1.4 Summary

Multiword expressions are a hard and open problem in natural language processing, due to their complex nature. The question of what counts as a MWE is a polemic one. Put simply, MWEs are habitual recurrent word combinations of everyday language (Firth 1957). Probably the most prototypical types of MWEs are idiomatic expressions like *loan shark*, *stand on someone's feet*, *cut someone a break*, *buy someone some time*, *save someone's breath*, *French kiss*, and *open mind*. In addition to idiomatic expressions, other constructions can be seen as MWEs. Further examples of MWEs include common names (e.g., *vacuum cleaner*, *voice mail*, *high heel shoe*) and verbal expressions (e.g., *make sense*, *take advantage*, *take a shower*, *take for granted*).

Native speakers rarely realise it, but colloquial speech is full of formulaic expressions such as *good morning*, *my bad*, *too bad* and *bye bye*. It is often assumed that a native speakers' lexicon contains as many MWEs as simple words (Jackendoff 1997). Thus, any computational system dealing with human language must take MWEs into account. In numerous NLP applications, when the words composing a MWE are treated as separate units, this can induce the system to produce erroneous output. An MT system, for instance, needs to be aware of MWEs to avoid literal translations.

Taking MWEs into account can be complicated for traditional NLP applications, as MWEs lie in a fuzzy zone between the lexicon and the syntax of a language. Therefore, the availability of tools and resources containing MWEs is still limited both in terms of effectiveness and of applicability to languages, contrasting with the ubiquitous and pervasive nature of MWEs. As a consequence, there is a need for developing, consolidating and evaluating techniques for the automatic acquisition of MWEs from corpora.

This thesis addresses the problem of MWE treatment in NLP applications, ranging from their automatic acquisition in raw text to their integration into two real-life applications: computer-aided lexicography and empirical MT. We have developed a conceptual model for the pipeline of MWE treatment, as well as a concrete software framework that validates the proposed methodology. We have evaluated this model thoroughly and systematically. We can summarise the goals of the present thesis as follows:

1. To develop generic and portable techniques for automatic MWE acquisition from corpora.
2. To evaluate these techniques extrinsically, that is, by measuring their usefulness in real NLP applications.
3. To investigate these tasks in bi- and multilingual contexts, studying how different parameters of the acquisition context influence the quality of automatically acquired MWEs.

Part I

Multiword expressions: a tough nut to crack

2 DEFINITIONS AND CHARACTERISTICS

In this chapter, we provide a broad yet not exhaustive discussion of the foundations, definitions, properties and current research trends in MWE treatment. Although we provide some pointers toward linguistic and psycholinguistic studies, most of the related work cited in this chapter presents a strong computational bias, as the present thesis is inserted in a scientific context of computational linguistics.

As motivated in Chapter 1, the computational treatment of MWEs is a tough problem. However, it does not constitute a new problem neither in linguistics nor in computational linguistics. Therefore, Section 2.1 starts with a brief overview of the history of the field, discussing some seminal papers, recent results and current trends in academic and industrial research.

Afterwards, Section 2.2 provides a set of more or less standard definitions for the term "multiword expression", which will engender the definition adopted in this thesis. We close this section on a note on the similarities and differences between MWEs and terms.

In Section 2.3, we characterise MWE properties, presenting arguments and examples based on linguistic intuition and coming from corpus evidence. We complement this discussion with a presentation of existing taxonomies for MWE types and a suggestion of a new rough classification which groups similar constructions in terms of their morphosyntactic category and difficulty degree.

2.1 Contextualisation

MWEs are a hot topic and an exciting research field of computational linguistics. Research has made significant progress in recent years, and this is reflected on the large number of papers that focus on data-driven (semi-)automatic acquisition of multiword units from corpora. A considerable body of techniques, resources and tools to perform automatic extraction of expressions from text are now available. This is an evidence of the growing importance of the automatic MWE acquisition field within the NLP community (see Section 3.2.1). A clear evidence of this "change of status" is that the annual workshop on MWEs attracts the attention of an ever-growing number of participants.

Researchers from several fields view MWEs as a key problem in current NLP technology. And yet, there are still important and urgent open matters to be solved. This section provides an overview of the MWE field, starting with a discussion of the seminal papers in the domain in theoretical and computational linguistics (Section 2.1.1). Then, we present the current industrial and academic research contexts, detailing the main research trends (Section 2.1.2). At the end of this section, we provide a set of pointers for obtaining further up-to-date information about the organisation of the research community, which are

potentially useful for any researcher wanting to know more about MWEs (Section 2.1.3).

2.1.1 A brief history

The study of MWEs is almost as old as linguistics itself. Traditional generative linguistic studies present an idealised point of view in which grammatical phenomena can be formally classified into *lexical* or *syntactic* levels. The *lexical level* considers words as separate units, independently of their neighbour words. It deals with questions such as morphology, inflection (e.g., number, gender, verb tense), word formation (prefixes, suffixes) and lexical semantics (the meanings of a word). The main object of lexical studies is the lexicon, that is, the set of words used in a language, and its description, which constitutes a dictionary. The *syntactic level* deals with word order in natural language utterances. Grammars are used to formalise the rules that govern the position of words and phrases, and how they can be combined. Syntax studies investigate, for instance, the place of epithet adjectives with respect to their corresponding noun, and the order of verb, subject and object in languages.

However, when trying to classify linguistic phenomena into either lexical or syntactic, one realises that some of them, and in particular MWEs, lie in between these two levels. Therefore, linguistic and computational approaches to grammar need to include MWE representations in their models. In what follows, we present the linguistic and computational work that gave origin to the current research field of MWEs.

2.1.1.1 Theoretical linguistics

In the traditional generative grammatical framework, the representation of idioms poses a challenge. For example, the English idiom *first off* is an adverbial locution synonym to *firstly*, that is, before anything else. The information about the morphosyntactic category and meaning of each of these two words taken individually, *first* and *off*, is contained in the lexicon. However, by combining them, it is impossible to guess the syntactic category and the meaning of the idiom as a whole. Moreover, general syntax rules of English formally forbid to combine an adjective with a preposition in order to form an adverbial phrase. This would make us to consider such idioms as a single lexical unit containing a space. However, other idioms such as *spill the beans* also have idiomatic meaning but they conform to general syntactic rules (for example, the verb can be inflected). Should one consider all possible inflections as separate lexical units, thus filling the lexicon with redundancy? Should one represent it in the grammar as separate entries, thus supposing that the idiom allows free modification according to general syntactic rules?

Such grammar engineering questions show that there are limitations in the structural approach to language à la Chomsky and Tesnière. One of the seminal papers of the *construction grammar* trend is the work of Fillmore et al. (1988). They illustrate and discuss in detail the weaknesses of this idealised atomistic approach to grammar, arguing that:

> As useful and powerful as the atomistic schema is for the description of linguistic competence, it doesn't allow the grammarian to account for absolutely everything in its terms. [...] the descriptive linguist needs to append to this maximally general machinery certain kinds of special knowledge—knowledge that will account for speakers' ability to construct and understand phrases and expressions in their language which are not covered by the grammar, the lexicon and the principles of compositional semantics, as these are

> familiarly conceived. Such a list of exceptional phenomena contains things which are larger than words, which are like words in that they have to be learned separately as individual whole facts about pieces of the language, [...] (Fillmore et al. 1988, p. 504)

Construction grammar suggests that there must be an appendix to the set of lexical units and syntactic rules of a language model. This appendix is a repository containing a large amount of *idiomatic* entries and their specific syntactic, semantic and pragmatic characteristics. Idioms become thus part of the core of the grammar: a language can be fully described by its idioms and their properties. These idioms correspond to what we call MWEs in this thesis.

Another linguistic theory that gives much importance to MWEs is the *meaning-text theory* (MTT). This theory proposes a rigorous description of the lexicon in the form of an explanatory combinatorial dictionary (Mel'čuk et al. 1984, Mel'čuk et al. 1988; 1992; 1999). Mel'čuk et al. (1995, p. 17) state that "Exaggerating a little, we could even say that the set of lexies [the lexicon] *is* the language." [1]

According to Mel'čuk and Polguère (1987), a dictionary entry contains three zones: (i) the semantic zone, (ii) the syntactic zone, and (iii) the lexical combinatorics zone. MWEs are present at two points of the computational MTT model: as *phrasemes* and as lexico-semantic functions (LSF) in the *lexical combinatorics zone*. The head of an entry in the explanatory combinatorial dictionary is a "lexie", that is, a lexeme or a phraseme used with a specific meaning. This second type of entry, the phraseme, represents a fixed expression that needs to be described as a separate lexical unit in spite of the fact that it is composed of more than one word. Phrasemes are more rigid MWEs that only allow very low or no morphosyntactic flexibility. The second type of MWE present in the explanatory combinatorial dictionary is in the lexical combinatorics zone. The latter contains a set of LSFs describing the interactions of the described lexical unit with other lexical units. A lexico-semantic function can be, for instance, the diminution of a word: in order to say that the *rain* is not intense, one uses the adjective *light*, thus **AntiMagn**(*rain*)=*{light, thin}*. The content of the lexical combinatorics zone in the explanatory combinatorial dictionary is what linguists usually describe as collocations, that is, habitual or conventional words that are used together with the target lexical unit. [2] Mel'čuk et al. (1995, p. 46) explain the difference between phrasemes and collocations as follows:

> The ECD [Explanatory Combinatorial Dictionary] does not describe all phrasemes in the same way. The *complete phrasemes* [...] and *quasi-phrasemes* [...], that is, the phrasemes that cannot be completely described based on at least one of their constituents, form independent entries — like the lexemes. The *semi-phrasemes* (= *collocations* [...]) are described under the entry of one of their constituents — through what we call lexical functions. [3]

Recently, psycholinguistic and cognitive linguistics have shown interest in MWEs. Researches in language acquisition propose cognitively plausible models for the acquisi-

1. "En exagérant quelque peu, on pourrait même dire que l'ensemble des lexies [le lexique] *est* la langue."

2. See Section 2.2 for a clarification on the difference between MWE and collocation.

3. "Le DEC [Dictionnaire Explicatif Combinatoire] ne décrit pas tous les phrasèmes de la même façon. Les *phrasèmes complets* [...] et les *quasi-phrasèmes* [...], c'est-à-dire les phrasèmes qui ne peuvent pas être complètement décrits en fonction d'au moins un de leurs constituants, forment des entrées indépendantes — tout comme les lexèmes. Les *semi-phrasèmes* (= les *collocations* [...]) sont décrits sous l'entrée d'un de leurs constituants — par ce qu'on appelle les fonctions lexicales."

tion of MWE knowledge from exposure to language. Thus, there has been work on learning verb-particle constructions (Villavicencio et al. 2012), noun compounds (Devereux and Costello 2007), light verb constructions (Nematzadeh et al. 2012) and multiword terms (Lavagnino and Park 2010) based on corpora evidence and sophisticated cognitive models. In particular, these models try to validate computational models for MWE acquisition by checking their correlation with experiments that use similar models for human language acquisition (Joyce and Srdanović 2008, Rapp 2008).

An extensive account of MWEs in different linguistic theories falls out of the scope of this work, but we recommend the reading of Seretan (2008, p. 20–27), where a quite complete discussion about theoretical linguistic aspects of MWEs can be found.

2.1.1.2 Computational linguistics

In computational linguistics, the study of MWEs arose from the availability of very large corpora and of computers capable of analysing them by the end of the 80's and beginning of the 90's. One of the main goals of these first attempts to process MWEs using machines was to build systems for computer-assisted lexicography and terminography of multiword units. Among the seminal papers of the field, one of the most often cited ones is Choueka (1988), who proposed a method for collocation extraction based on *n*-gram statistics.

Another ground-breaking work is that of Smadja (1993). He proposed Xtract, a tool for collocation extraction based on some simple POS filters and on mean and standard deviation of word distance. His approach has the advantage of handling non-contiguous constructions. His work is strongly based on the notion of collocation as outstanding co-occurrence. The reported precision on specialised texts was of around 80%.

Church and Hanks (1990) suggested the use of a more sophisticated association measure based on mutual information. They provided theoretical justification for it and then tested it on relatively large corpora for the extraction of terminological and collocational units. Later, Dagan and Church (1994) proposed a terminographic environment called Termight, which uses this association score. Termight performs bilingual extraction and provides tools to easily classify candidate terms, find bilingual correspondences, define nested terms and investigate occurrences through a concordancer.

Also in the context of terminographic extraction, Justeson and Katz (1995) proposed a simple approach based on a small set of POS patterns and frequency thresholds. Using minimal linguistic information combined with an intuitive quantitative technique, they obtained surprisingly good results given the simplicity of the technique. [4]

The indiscriminate use of association measures was first criticised by Dunning (1993). He argued that the assumption underlying most measures is that words are distributed normally, but corpus evidence dos not support this hypothesis. Therefore, he proposed a 2-gram measure called *likelihood ratio*. It estimates directly how more likely a 2-gram is than expected by chance. In addition to being theoretically sound, Dunning's score is also easily interpretable. Nowadays, measures based on likelihood ratio (e.g., the log-likelihood score) are still largely employed in several MWE extraction contexts.

At the beginning of the 2000's, the Stanford MWE project [5] has revived interest of the NLP community in this topic. One of the most cited publications of the MWE project is the famous "pain-in-the-neck" paper by Sag et al. (2002). It provided an overview of

4. A pedagogical example of the application of this technique on a corpus is given by Manning and Schütze (1999, p. 156).

5. `http://mwe.stanford.edu/`

MWE characteristics and types and then presented some methods for dealing with them in the context of grammar engineering. The Stanford MWE project is also at the origin of the MWE workshop series.

2.1.2 MWEs in current language technology

One of the actors interested in MWEs is the industry of language technologies. The vertiginous growth in the size of the lexicons of commercial systems is an evidence that they incorporate automatic techniques for MWE acquisition in order to build their lexicons (Section 2.1.2.1). In terms of academic research, we describe the two main trends in current MWE research (Section 2.1.2.2). One of the goals of this comparison between industrial and academic research is to provide some evidence of the current gap between these two contexts.

2.1.2.1 The industrial scenario

In parallel to the academic research community, MWEs are present in many commercial tools, specially in terminology extraction and translation technology. However, it is very hard to obtain concrete information in this context. In the name of industrial secret, descriptions are rarely published in academic conferences. Therefore, here we report the official figures from company websites and present our hypotheses about how they do it.

Automatic MWE acquisition technology is included in many computer-aided translation tools like Similis, Trados and DejaVu. More details on proprietary tools that perform some kind of MWE acquisition are provided in Section 3.2.3.2 and in Section 7.2.

In industrial MT, a notable case is the growth of Fujitsu's package technical dictionaries. This product is a complement to the state-of-the-art MT system ATLAS-II. [6] Their dictionary of English–Japanese technical terms went from 70,000 entries in 1983 to more than 5.5 million entries in the last version (Boitet et al. 2006). [7, 8] Given the huge amount of work that these figures represent, it is possible that the Fujitsu team performed some kind of (semi-)automatic MWE extraction from specialised documents. Otherwise, it would be barely impossible to gather so many entries in 28 different specialised domains.

Another way of inserting MWEs into a MT system is to allow direct input by the users. Systran provides this functionality through what they call the *user dictionary*. According to their documentation, one can perform automatic or manual terminology extraction and then import it into the MT system in order to personalise it. It is even possible to build such dictionaries using phrase tables built empirically from parallel corpora, in the fashion of empirical MT systems. [9]

2.1.2.2 The research scenario

We identify two trends in the current academic research community, the *grammar engineering trend* and the *computational semantics trend*. The former is a minor trend in the NLP field, as MWEs are (too) often ignored in the design of applications. Related work

6. ATLAS-II is a transfer MT system for translating from and to Japanese and English, French, German and Spanish. It is based on a semantic pivot. See also `http://www.fujitsu.com/global/services/software/translation/atlas/`

7. Assuming that field dictionaries do not overlap, there are 2,837,000 entries in English–Japanese and 2,761,000 in Japanese–English in version 14.

8. A detailed description is provided in Appendix G.1.

9. `http://www.systran.fr/support/informations-importantes/gestionnaire-de-dictionnaires/guide-utilisateur-de-codage-dictionnaire`

tends to describe the automatic or manual construction of lexicons containing multiword entries (Izumi et al. 2010). This is often carried out in a larger context of some lexicography or terminography project (Laporte and Voyatzi 2008). Another concern of this trend is to optimise the internal representation of the relations between the words composing a MWE (Schuler and Joshi 2011, Graliński et al. 2010).

Most of the time, techniques in this trend use corpora and intuition as a source of information. The use of morphosyntactic patterns, frequency information and heuristic filters is standard. Often, an expert or a team of experts goes through the automatically extracted lists in order to manually validate the data. The resulting lexical resources may be used for various purposes: from printing a dictionary to building a syntactic parser.

In the *computational semantics trend*, the lexical and compositional semantics of MWEs are explored. Thus, the problem of identifying a MWE is analogous to the problem of deciding whether the semantics of a sequence of words is compositional. Work in this trend often use some kind of semantic resource like Wordnet (Pearce 2001, Ramisch et al. 2008b). The use of automatically generated thesauri can replace manually constructed resources (McCarthy et al. 2003). Such semantic representations based on word co-occurrences in a corpus are often referred to as *distributional methods*.

Generally, large corpus statistics are used in this trend. As in traditional corpus linguistics, studies tend to concentrate on a specific type of construction or phenomenon. Results are punctual contributions and generally do not take into account practical considerations such as portability and scalability. A review of the state of the art in current academic research is the theme of Chapter 3.

2.1.3 Further reading

The MWE research community is organised and shares some common resources. The first and most important place to exchange ideas on MWE research is the annual workshop on MWEs. It is a series of workshops that have been held since 2001 in conjunction with major computational linguistics conferences (Bond et al. 2003, Tanaka et al. 2004, Rayson et al. 2006, Moirón et al. 2006, Grégoire et al. 2007; 2008, Anastasiou et al. 2009, Laporte et al. 2010, Kordoni et al. 2011b). The recent editions of the workshop show that there is a shift from research on identification and extraction methods work toward more application-oriented research. The evaluation of MWE processing techniques and multilingual aspects are also current issues in the field. For example, there has been some published work on the automatic translation of several types of MWEs. However, there is a gap between research methods for bi- and multilingual MWE acquisition and commercial translation tools (see Section 2.1.2.1).

From 2012 on, the MWE workshop will be absorbed as an area of a new conference called *SEM, which gathers several sub-fields of NLP around a common theme: natural language semantics. [10] In addition to the specialised workshops, main computational linguistics conferences such as COLING, ACL and LREC regularly feature papers on MWEs. For example, the best paper award of COLING 2010 went to a paper about compositionality measures for multiword units (Bu et al. 2010).

Most of the information concerning past editions of the MWE workshop series can be found at the MWE community website. [11] The site also hosts a repository with several annotated data sets and a list of software capable of dealing with MWEs. The community also uses as communication tool a mailing list to which anyone can subscribe.

10. http://ixa2.si.ehu.es/starsem/
11. http://multiword.sourceforge.net/

Finally, as a complement to workshops and conferences, special issues on MWEs have been published by leading journals in computational linguistics: the journal of Computer Speech and Language (Villavicencio et al. 2005a) and the journal of Language Resources and Evaluation (Rayson et al. 2010a). At the time of writing this thesis, there is an open call for papers for a future special issue of the ACM Transactions on Speech and Language Processing (Kordoni et al. 2013). The special issues generally gather publications describing consolidated research projects around MWEs. Thus, they provide a broad overview and present the most relevant research results coming from different authors and research groups working on the subject.

2.2 Defining MWEs

MWEs are hard to define. Yet, it is important to define them because evaluation depends on the definition. Annotators need to know what they are looking for, otherwise they cannot perform a binary choice of telling whether a word combination constitutes a MWE or not. For example, the English expression *take a shower* seems fairly compositional, that is, the meaning of the whole is similar to the meaning of the noun. Therefore, using compositionality as a criterion, this expression would not be a MWE. However, when translated into Italian, a word-for-word translation is impossible as the correct corresponding Italian expression would be *fare la doccia* (lit. *make a shower*) instead of the literal *?prendere la doccia*. Therefore, for a MT system, it would be important to treat this expression as a MWE, either during analysis or transfer. In this section, we present and discuss the coverage of some classical definitions, concluding on the definition adopted in this work.

2.2.1 The MWE jungle

Before diving into the multiword expression jungle, we must define what we consider to be a *word*. However, behind this apparently simple question, there are deep theoretical questions and little agreement on the answer. For instance, Mel'čuk et al. (1995, p. 15) say that "we know the restive character of the word *word*, which, to date, has escaped the attempts to circumscribe it with precision although much has been written about this subject throughout the decades." [12] As pointed out by Manning and Schütze (1999, p. 125), "the question of what counts as a word is a vexed one in linguistics, and often linguists end up suggesting that there are words at various levels". They suggest to simply yet operationally define graphic words as "contiguous alphanumeric characters with spaces on either side". Hovever, this definition poses several problems in English for tokens involving hyphens (*language-independent approach*), punctuation (*google.com*, *US$ 1,299.99*), contractions (*do not* as *don't*). In languages other than English, this definition is not suitable: for example, the writing system of many Asian languages (e.g., Japanese, Chinese) do not use whitespace between words at all, Germanic noun compounds are concatenated together as a single word, [13] and other morphologically rich languages like Turkish and Finnish tend to form new words by appending lexical units rather than using spaces. In this work, we adopt the definition by Evert (2004) who considers a "word as an entirely generic term which may refer to any kind of lexical item, depending on the underlying

12. "on connaît tout aussi bien le caractère rétif du mot *mot*, qui, jusqu'à présent, a échappé aux tentatives de le circonscrire avec précision et a fait couler beaucoup d'encre pendant des décennies.

13. Intervening material may be added, as it is the case for letter *s* in some German compounds like *Prüfungstermin* = *Prüfung* + *Termin*.

theory or intended application."

At this point, a clarification on the nomenclature adopted is required. In the following definitions, we use the terms *collocation*, *idiom* and *multiword expression*. In the literature, the three terms are commonly and sometimes interchangeably employed, with no unique definition as both theoretical and computational linguists did not reach a consensus to date. Actually, there are slight differences between them, from our point of view.

The term *idiom* is generally employed in the construction grammar tradition to denote a combination of words with non-compositional semantics. In other words, an idiom is a combination whose meaning or syntax cannot be modelled by applying general grammar rules to combine the meaning/syntax of the individual words (Fillmore et al. 1988). The degree of idiomaticity or non-compositionality of an idiom may vary in a continuum from almost transparent idioms to completely opaque ones.

The notion of *collocation*, however, does not depend as much on compositionality as it does on co-occurrence. Combinations such as *heavy rain*, *strong coffee* and *drop drastically* are prototypical examples of collocations. In many grammatical theories and, in particular, in MTT, the term collocation expresses a combination of words usually appearing together in a given (sub-)language (Mel'čuk et al. 1995). Collocations correspond to the usual way of expressing something in a language. Formally, for a target *base* word, there is a set of usual *collocates* that modify and disambiguate it (Yarowsky 2001).

It is worth noting that, while the term collocation has been used for a long time in linguistics and also in the beginning of the 90's in computational linguistics, the term multiword expression has gained popularity in the beginning of the 2000's after the seminal "pain in the neck" article by Sag et al. (2002) and the Stanford MWE project [14]. Thus, the term *multiword expression* (or *multiword unit*) is more popular in computational linguistics and represents a generalisation of the two former denominations in the sense that it covers linguistic phenomena that cross the borders between words without being freely syntactic combinations (Sag et al. 2002). As in Sag et al. (2002), we will assume that "the term collocation [refers] to any statistically significant co-occurrence, including all forms of MWE". That is, the notion of collocation is corpus-dependent and encompasses the notion of MWE. However, there are collocations (e.g., *doctor—nurse*) that are not MWEs because they breach the property that a MWE "has to be listed in a lexicon" (Evert 2004, p. 17). The term MWE seems to be the most generic one and matches the goals of the present work, therefore this will be the term employed from now on. However, in the quotations below, we keep the original denominations for the sake of coherency with the source.

The notion of MWE has its origin in Firth's famous quotation "you shall know a word by the company it keeps". He affirms that "collocations of a given word are statements of the habitual and customary places of that word" (Firth 1957, p. 181). Analogously, Smadja (1993) considers collocations as "arbitrary and recurrent word combinations". The definition adopted by Choueka (1988) focuses on non-compositionality; for him, a collocation is "a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components". For Fillmore et al. (1988, p. 504), non-compositionality is also the main property of MWEs, as "an idiomatic expression or construction is something a language user could fail to know while knowing everything else in the language", that is, that cannot be modelled using general lexical knowledge, grammatical rules and compositional semantics. Sag et al. (2002) generalise this same property to vaguely define MWEs as "idiosyncratic

14. `http://mwe.stanford.edu/`

interpretations that cross word boundaries (or spaces)".

In some publications, the authors do not define MWEs, but instead enumerate examples. For instance, the latest special issue on MWEs of the Language Resources and Evaluation journal (Rayson et al. 2010a) starts as follows:

> MWEs range over linguistic constructions such as idioms (*a frog in the throat, kill some time*), fixed phrases (*per se, by and large, rock'n roll*), noun compounds (*telephone booth, cable car*), compound verbs (*give a presentation, go by [a name]*), etc.

This enumeration may be well suited, specially because this is not a closed class and new interpretations of linguistic phenomena may create "new" types of MWEs, like the discourse relation markers suggested by Joshi (2010).

For a broad inventory of MWE definitions, one may refer to appendix B of Seretan (2008, p. 182-184).

2.2.2 A practical definition

All the definitions cited above are valid in a given experimental setup. Although, the definition of MWE adopted will influence the evaluation results, as it will be used to write annotation guidelines and/or to select reference lists. Therefore, in this thesis, we adopt the definition by Calzolari et al. (2002), who define MWEs as:

Definition 2.1 (Multiword expression) *[...] different but related phenomena [...]. At the level of greatest generality, all of these phenomena can be described as a sequence of words* [15] *that acts as a single unit at some level of linguistic analysis.*

This generic and intentionally vague definition can be narrowed down according to the application needs. For example, for the empirical MT system used in the examples shown in Table 1.1, a MWE is any sequence of words which, when not translated as a unit, generates errors or unnatural output.

Another concrete example: when translating technical documentation of the technology domain, should filenames and paths containing spaces be considered as MWEs? In theory, no, as they are not true "words" in general language. However, before running a MT system, a preprocessing step must identify such cases in order to keep them untranslated and/or post-process them after translation. They are sort of "easy" MWEs because simple patterns based on regular expressions can detect them (e.g., most email clients detect URLs in the messages you write). From the point of view of this work, at some level of processing these items need to be treated as a unit. Therefore, they can be considered as MWEs. In practice, however, our work aims at more challenging and flexible constructions that cannot be identified using simple regular expressions.

The level at which a combination of words needs to be treated as a unit varies. In a complete analysis system, for instance, fixed expressions such as *ad hoc* and *by and large* will probably constitute lexical entries on their own, while more flexible constructions like *take off* and *bus stop* will be dealt with as a unit during syntactic parsing and more non-compositional constructions like *kick the bucket* are likely to be treated during semantic processing.

15. Although this definition refers to word *sequences*, thus assuming contiguous MWEs, we assume word *combinations* for greater generality.

2.2.3 A note on MWEs and terms

A definition related to MWEs in the context of domain-specific texts is that of *term* (Cabré 1992):

Definition 2.2 (Term) *A terminology is a specialised lexicon corresponding to the set of words that characterise a specialised language of a domain. A term is the basic lexical unit of a terminology.*

MWEs and terms have some similar aspects: both have non-conventional semantics and both are a challenge for NLP systems. Manning and Schütze (1999, p. 152) point out that:

> There is considerable overlap between the concept of collocation and notions like *term*, *technical term* and *terminological phrase*. As these names suggest, the latter three are commonly used when collocations are extracted from a technical domain (in a process called terminology extraction). The reader should be warned, though, that the word *term* has a different meaning in information retrieval. There, it refers to both words and phrases. So it subsumes the more narrow meaning that we will use.

From the point of view of the present work, there are several differences between MWEs and terms, not only epistemologically but also pragmatically. First, terms may be either simple (single-word) or multiword units like nominal and verbal locutions, whereas MWEs are inherently composed of two or more words. Second, MWEs occur in both technical/scientific language and in general-purpose everyday language while terms occur only in the former. Even though the sharp distinction between general and specialised communication has been questioned, the difference is important here because the available computational methods to deal with MWEs in general-purpose texts are potentially different from methods to handle specialised corpora and terminology. *Multiword terms* (MWT) lie in the intersection between terms and MWEs (SanJuan et al. 2005, Frantzi et al. 2000, Ramisch 2009):

Definition 2.3 (Multiword term) *A multiword term is a specialised lexical unit composed of two or more typographic words, and whose meaning cannot be directly inferred by a non-expert from its parts because it depends on the specific area and on the concept it describes.*

Notice that this definition is essentially different from what terminologists consider to be a *phraseological expression* like in *to initiate a remote digital loopback test*. Phraseological expressions are much more related to specialised collocations than to our conception of MWT. Specialised phraseology deals with more complex constructions that often involve more than one domain-specific concept, and are often seen as intermediary entities between terms and institutionalised sentences. We, on the other hand, consider a MWT as a multiword lexical representation of an abstract term, but sharing with the latter the same properties of monosemy and limited variability. In other words, a MWT, as well as a single-word term, is a lexical manifestation of a specialised concept.

2.3 Characteristics and characterisations

As for almost everything else — and as a consequence of the fact that there is no unique definition for the concept of MWE — there is also no unique taxonomy to organise

them into classes presenting similar characteristics. However, the literature reports several attempts to typify MWEs based mostly on their syntactic and semantic idiosyncrasies. In this section, we first present some characteristics of MWEs (Section 2.3.1). Then, we summarise several proposals of hierarchical classification for MWE types (Section 2.3.2). Finally, we present our own rough classes used in this work (Section 2.3.3). Our typology is based on the morphosyntactic role of the whole expression in a sentence and on the degree of difficulty to treat them in NLP systems. Wherever possible, we explicit the link between "our" classes and the cited classification propositions.

2.3.1 MWE properties

Based on intuition and on corpus observation, researchers describe some common properties of MWEs in general, which we summarise below. It is important to keep in mind that these are not binary yes/no flags, but values in a continuum going from completely flexible, ordinary word combinations to totally prototypical and/or fixed expressions. Thus, any particular expression will present the properties described below to a variable extent and will probably not manifest a high degree for all of them simultaneously.

1. **Arbitrariness**. This is probably the most challenging property of MWEs. Their arbitrary character is well illustrated by Smadja (1993, p. 143–144), who listed eight different and valid ways of referring to the Dow Jones index, from which only four are acceptable. [16] This happens because sometimes a perfectly valid construction both syntactically and semantically is not acceptable simply because people do not use to talk that way. That is also why MWEs are hard for second language learners, who know the lexicon and grammar of the language but lack of knowledge about language use.
2. **Institutionalisation**. MWEs are recurrent, as they correspond to conventional ways of saying things. Fillmore et al. (1988) argue that the inventory of constructions in a language is too large to be considered as an appendix or as a list of exceptions. Jackendoff (1997) estimates that they correspond to half of the entries of a speaker's lexicon. Sag et al. (2002) point out that this may be an underestimate if we take domain-specific MWEs into account. Indeed, some researchers assume that the proportion of multiword terms in a specialised lexicon is around 70% (Krieger and Finatto 2004). Empirical measurements showed that this ratio is between 50% and 80% in a corpus of scientific biomedical abstracts (Ramisch 2009). This property is directly related to collocational behaviour, and motivates using frequency information (and all the related statistical tools) in order to automatically identify MWEs in corpora.
3. **Limited semantic variability**. MWEs do not undergo the same semantic compositionality rules as ordinary word combinations. This is often expressed in terms of the following sub-properties:
   (a) **Non-compositionality**. The meaning of the whole expression often cannot be directly inferred from the meaning of the parts composing it. Therefore, there is a lack of compositionality that ranges in a continuum from completely compositional MWEs (*bus stop*) to completely opaque ones (*kick the bucket* as *to*

16. One can say *The Dow Jones average of 30 industrials*, *The Dow average*, *The Dow industrials* or *The Dow Jones industrial*, but never ?*The Jones industrials*, ?*The industrial Dow*, ?*The Dow of 30 industrials* nor ?*The Dow industrial*.

die). The MTT models a MWE as being composed of two parts: a base which carries the core meaning (e.g., *rain* in *heavy rain*) and a collocate that modifies the sense of the base (*heavy*). While this model captures semi-fixed expressions, it fails to generalise when the meaning of the MWE is closer to the edges of the compositionality spectrum. For instance, it is hard to designate a base and a collocate for completely idiomatic expressions (e.g., *big deal*) and for expressions where both elements seem to be equally relevant to the meaning of the expression (e.g., *cable car*)..

(b) **Non-substitutability**. Because MWEs are non-compositional, it is not possible to replace part of a MWE by a related (synonym/equivalent) word or construction. This motivates the notion of *anti-collocations* (Pearce 2001), which are awkward or unusual word combinations (e.g., *strong coffee* vs *?powerful coffee*). Syntactic and semantic variations are used in several techniques aimed at the automatic identification and classification of MWEs (Pearce 2001, Fazly and Stevenson 2007, Ramisch et al. 2008b).

(c) **No word-for-word translation**. This is a consequence of the above properties. However, it constitutes a useful test to decide whether a construction should be considered as a MWE or not, as we exemplified in Section 1.1.3. This motivates the application-oriented evaluation of Chapter 7. Ideally, the knowledge about MWEs should be available in both, source and target languages, within a MT system (Smadja 1993). However, knowing MWEs at the source side already can improve the quality of MT (Carpuat and Diab 2010), and sometimes it is better to transliterate them instead of translating them (Pal et al. 2010). Translational (non-)equivalences can also be used to detect MWEs when parallel data is available (de Medeiros Caseli et al. 2010, Attia et al. 2010), as discussed in Section 3.2.2.

(d) **Domain-specificity/idiomaticity**. Smadja (1993) emphasises that MWEs are related to a specific sublanguage. Thus, for the layman not familiar with it, it is hard to identify them. A sublanguage may be a specialised scientific or technical domain (e.g., epistemology, chemistry, cars, fashion), a regional or dialectal variation (e.g., Brazilian vs European Portuguese), or a text genre (e.g., poetry vs textbooks).

4. **Limited syntactic variability (or non-modifiability)**. Standard grammatical rules do not apply to MWEs, and this can be demonstrated by the following sub-properties:

(a) **Extragrammaticality**. Fillmore et al. (1988) introduce this property, arguing that such expressions are unpredictable and seem "weird" for somebody (e.g., a second language learner) who only knows general lexical and morphosyntactic rules. Examples of extragrammatical MWEs include, in English, *kingdom come*, *by and large*; in Portuguese, *dar para trás*, *um Deus nos acuda*, *prós e contras*; and in French, *faire avec*, *sens dessus dessous*, *de par le monde*.

(b) **Lexicalisation**. Somehow, the knowledge that a set of words "belongs together" must be available to NLP applications. Because "MWEs can be regarded as lying at the interface between grammar and lexicon" (Calzolari et al. 2002), parsing engineers often need to chose where each MWE belongs. It is not enough to list them all in the lexicon, because this would result in undergeneration. Conversely, listing them all in the grammar as free combinations

would make a parser overgenerate. In other words, they have a variable degree of lexicalisation, and identifying this degree of lexicalisation for each MWE (class) is important for NLP analysis and generation tasks. This property relates to what Smadja (1993) calls "cohesive lexical clusters" and to the assumption of Evert (2004), who argues that MWEs need to be represented as a lexical unit.

5. **Heterogeneity**. It is not a coincidence that MWEs are hard to define, as the term encompasses a large amount of distinct phenomena. This complexity makes them hard to deal with by NLP applications, which cannot use a unified approach and usually need to rely on some type-based approach using one of the multiple MWE classifications available (see Section 2.3.2).

2.3.2 Existing MWE typologies

2.3.2.1 Constructionist typology

Fillmore et al. (1988) suggest a typology based on the predictability of a construction with respect to standard syntactic rules (and somehow related to semantic compositionality). They define three classes among the four possibilities of unfamiliar/familiar pieces unfamiliarly/familiarly combined.

- **Unfamiliar pieces unfamiliarly combined**. this class contains idiomatic constructions that are extremely unpredictable. This may concern, for instance, words that only appear in a specific idiom (*ad hoc*, *with might and main*) or very specialised syntactic configurations that do not occur anywhere else in language (*the more, the merrier*, and more generally expressions of the type *the X-er, the Y-er*).
- **Familiar pieces unfamiliarly combined**: these constructions require special syntactic and semantic rules for their interpretation. Examples are *all of a sudden*, *stay at home* and constructions of the type *first cousin two times removed*.
- **Familiar pieces familiarly combined**: constructions in this class do not present any particular syntactic idiosyncrasy and their members are combined using standard grammatical rules. However, they have an idiomatic interpretation like in *pull someone's leg* and *tickle the ivories*.

2.3.2.2 MTT typology

Another classification is suggested by Mel'čuk et al. (1995), who use as a criterion the relevance of a given expression as an entry in a dictionary. According to them, complete phrasemes and quasi-phrasemes must be full entries in the dictionary while semi-phrasemes are represented as LSFs in the lexical co-occurrence zone of the entries corresponding to the base words. Their goal is to optimise the access to information based on the most probable circumstance in which a speaker would require it. Their classification can be easily expressed in terms of semantic compositionality, and contains the following classes: complete phrasemes, semi-phrasemes and quasi-phrasemes.

- **Complete phraseme**: fully idiomatic expression, that is, the meaning of the expression has no intersection with the meaning of any of its components, for instance, *kick the bucket* and *Achilles' hill*. In other words, the expression is fully non-compositional.
- **Semi-phraseme**: expressions in which the meaning of at least one of the components is present in the meaning of the whole expression, but the global meaning still does not correspond to the systematic composition of all individual meanings.

This is the case of most collocations that can be expressed in terms of a base word (the one which contributes its meaning to the expression) and the collocate word (which modifies or adds some new interpretation to the meaning of the base word). Examples include *take a nap* and *break up*. Such expressions can be expressed in terms of LSFs in the explanatory combinatorial dictionary. They correspond to the set of familiar pieces familiarly arranged in Fillmore's classification.

– **Quasi-phrasemes**: expressions in which all the words keep their original meanings but an extra element of meaning is added by the fact that they occur in an expression. So for instance a *bus stop* is actually a place where the bus stops, but not any place. If the bus stops at the traffic light, nobody can get on and off the bus, so this will not be considered as a bus stop. Analogously, not every light that helps controlling the traffic is a *traffic light*.

2.3.2.3 "Pain-in-the-neck" typology

A slightly more sophisticated classification scheme is proposed by Sag et al. (2002). They separate institutionalised from lexicalised expressions and further classify the latter into three sub-types according to their degree of syntactic flexibility: institutionalised and lexicalised phrases.

– **Institutionalised phrases**: are sets of words which co-occur often but have no syntactic idiosyncrasy, and whose semantics are fairly compositional. As they undergo full syntactic variability, they cannot be represented using a words-with-spaces approach. This class corresponds to the notion of collocation denominated in the previous schemes as semi-phrasemes and familiar pieces familiarly combined.
– **Lexicalised phrases**: present some idiosyncratic syntactic or semantic characteristics. This class can be further divided into three sub-classes according to their degree of flexibility: fixed, semi-fixed and syntactically flexible expressions.
  – **Fixed expressions**: they are expressions that do not allow any morphosyntactic modification. This includes expressions containing words that do not occur in isolation (*ad hoc*, *vice versa*) and extragrammatical expressions such as *kindom come* and *in short*. These expressions are immutable, not allowing any morphological inflection (**in shorter*), syntactic modification (**coming kingdom*) and internal insertion (**in very short*). This corresponds to the set of unfamiliar pieces unfamiliarly combined and would be included in the class of complete phrasemes.
  – **Semi-fixed expressions**: they have strict syntactic and semantic interpretation but undergo morphological inflection. This class contains notably noun compounds (*part of speech*) and proper names (*San Francisco*), as well as non-decomposable idioms (*kick the bucket*, *shoot the breeze*). The latter are idiomatic expressions with completely opaque semantic interpretation. That is, it is impossible to decompose the semantics of the whole into pieces assigned to parts of the expression.
  – **Syntactically flexible** [17] **expressions**: they allow a much larger range of syntactic variation than the former, still presenting idiosyncratic semantics. Examples include phrasal verbs (*take off*, *give up*), light verb constructions (*take a picture*, *make a mistake*) and decomposable idioms, that is, idioms in which it is possible to assign parts of a possibly non-standard meaning to parts of the expression

17. We would have preferred to use the term *syntactically variable expressions*, but in order to be coherent with the source we keep the original term.

(*spill the beans* can be analysed as *spill* = *reveal*, *the beans* = *the secret*).

2.3.2.4 Xtract typology

Probably, the typology most similar to the one we propose in this work in terms of selection criteria is that of Smadja (1993). His classification is oriented toward automatic terminology recognition. His classes are inspired by the types of filters which he applies during MWE acquisition. [18]

- **Predicative relations**: they correspond to a word modifying or adding some new meaning to the semantics of a base word, like *make a presentation* and *drop drastically*. According to the author, this class corresponds to the LSFs of Mel'čuk et al. (1995). Therefore, it presents a large overlap with what, in the previous classifications, is denominated as semi-phrasemes, institutionalised phrases and familiar pieces familiarly combined.
- **Rigid noun phrases**: they are nominal expressions such as noun compounds (*stock market*) and proper nouns (*the Dow Jones index*). Smadja characterises these expressions as allowing little or no flexibility (fixed and semi-fixed expressions), and often being used to describe a concept in a specialised domain (see multiword term in Section 2.2.3). These constructions are full lexical entries, as semantically equivalent formulations are not valid or do not designate the same concept.
- **Phrasal templates**: they are whole phrases prototypical in specialised texts. This class covers what terminologists often call phraseological expression, that is, some usual way of saying something in a domain. Phraseological expressions can be as long as a whole sentence and often look like a template containing gaps for variable parts, for example, *The average finished the week with a net loss of* NUMBER.

2.3.3 A simplified typology

In this work, we use a simplistic typology based firstly on the morphosyntactic role of the whole expression in a sentence (Section 2.3.3.1), and secondly on its difficulty to be dealt with using computational methods (Section 2.3.3.2).

2.3.3.1 Morphosyntactic classes

In this typology, expressions that act as or are heads of noun phrases are broadly classified as nominal expressions (Section 2.3.3.1.1); expressions containing a verb and other lexical items attached to it (adverbs, objects, complements) are considered verbal expressions (Section 2.3.3.1.2); likewise for adverbial and adjectival expressions (Section 2.3.3.1.3).

By no means do we argue that the classification proposed here is superior or more general than the existing ones. Actually, it would be quite easy to show the opposite, that is, that other schemes are more powerful and take MWE properties better into account than ours. However, the existing schemes are quite complex and/or intentionally vague. Thus, for the sake of practicality, we suggest and adopt our own taxonomy. Even though it is far from being exhaustive, it is quite simple and yet rigorous and precise enough to describe the MWEs dealt with in our experiments and in particular in NLP applications.

18. This classification is largely incomplete. It does not cover the whole set of MWEs defined in our work.

2.3.3.1.1 Nominal expressions

Nominal expressions are word combinations acting as nouns, that is, as noun phrases or heads of noun phrases, in a sentence. This class covers the following sub-types: noun compounds, proper names and multiword terms.

- **Noun compounds**: they are sequences formed by head nouns and other elements appended to it, like other nouns (*traffic light*), adjectives (*Russian roulette*) and adjectival locutions introduced by prepositions (*part of speech*). Noun compounds generally denote a specific concept for which there is no equivalent single-word formulation.
- **Proper names**: they are very similar to noun compounds except that they usually denote a very specific named entity of the world such as a name of a city (e.g., *Porto Alegre*), institution (e.g., *United Nations*) or person (e.g., *Alan Turing*). In some languages, they are distinguished from regular nouns using capitalised initials. According to Manning and Schütze (1999, p. 186), "[proper names] are usually included in the category of collocations in computational work although they are quite different from lexical collocations". The question of whether proper names should be considered as MWEs is an open one because (a) not all proper names are MWEs (e.g., *Paris*, *Google*) and (b) computational methods used for the automatic identification of proper names are different from methods to identify regular noun compounds. We will assume that identification of proper names is the concern of another NLP area called named entity recognition, which has its own methods and goals, thus falling out of the scope of this work.
- **Multiword terms**: they are also noun compounds with the specificity of being used in a specialised domain to denote a specific concept (see Section 2.2.3). Similarly to proper names, multiword terms are the main object of study in automatic term recognition, which is a research area on its own in computational linguistics, not being covered by our work.

2.3.3.1.2 Verbal expressions

Verbal expressions are those in which a verb is the head of the expression and other elements are appended to it. According to the class of the appended elements, they can be further sub-classified as phrasal verbs and light verb constructions.

- **Phrasal verbs**: they are constructions in which a preposition or adverb plays the role of a particle adding some meaning to the meaning of the base verb. This includes basically two types of constructions: (a) transitive prepositional verbs that are compositional but require a specific preposition introducing the object, like *rely on* and *agree with*, and (b) more opaque verb-particle constructions where the particle is actually attached to the verb, forming a cohesive lexical-semantic unit, like *give up* and *take off*. They are frequent in Germanic languages like English and German but rare in Latin languages like Portuguese and French. A detailed characterisation of phrasal verbs is provided in Section 7.3.1.
- **Light verb constructions**: they are also sometimes called support verb constructions, and correspond to a semantically "light" verb used with a noun that conveys most of the meaning of the expression. Thus, when one *takes a shower*, most of the semantics comes from the noun *shower* and not from the highly polysemous verb *take*. The complement of the verb is often a deverbal nominalisation derived from

a simple verb [19] which is synonym of the light verb construction, like *to shower = take a shower* and *to present = make a presentation.* [20] As the choice of the light verb is rigid and unpredictable in this kind of construction, a simple test to verify whether it is a genuine light verb construction consists of trying to replace the verb, yielding unnatural expressions (e.g., ?*make a shower*, ?*get a shower*). A detailed characterisation of complex predicates including light verb constructions is provided in Section 6.2.1.

2.3.3.1.3 Adverbial and adjectival expressions

Expressions acting as adverbs or adjectives in sentences belong to a separate class. Examples in English are *upside down*, *second hand*, *on fire*, *at stake*, and *in the buff*. These expressions are also quite frequent in other languages such as in French (*à poil*, *à la bourre*, *dans l'absolu*, *la tête en bas*) and Portuguese (*sem mais nem menos*, *de braços abertos*, *mais ou menos*, *de cabeça para baixo*). However, due to time and space limitations, they are out of the scope of this work.

2.3.3.2 Difficulty classes

In addition to these three main types, we also define three orthogonal types that are more related to the computational methods used to treat MWEs. Fixed expressions can be dealt with using relatively simple techniques while idiomatic expressions are very hard to recognise and require the use of external semantic resources. The last class contains what we call "true" collocations because they correspond to the notion of words that co-occur more often than expected by chance.

- **Fixed expressions**: they correspond to the fixed expressions of Sag et al. (2002), that is, it is possible to deal with them using the words-with-spaces approach. Such expressions often play the role of functional words (*in short*, *with respect to*), contain foreign words (*ad infinitum*, *déjà vu*) or breach standard grammatical rules (*by and large*, *kingdom come*).
- **Idiomatic expressions**: idiomatic MWEs have completely non-compositional semantics, that is, the literal interpretation of its members is completely unrelated to the meaning of the expression. Therefore, they are very hard to automatically identify in texts without the help of semantic resources. Examples include expressions from the three morphosyntactic classes above like nominal expressions (*dead end*, *dry run*), verbal expressions (*put in place*, *shoot the breeze*) and adjectival expressions (*on the same wavelength*, *all ears*).
- **"True" collocations**: MWEs corresponding to the notion of institutionalised phrases are fully compositional expressions both syntactically and semantically, but co-occurring more often than expected by chance. This class corresponds to Firth's definition of MWEs as " habitual and customary places of that word" and can be modelled using Mel'čuk's notion of LSF.

2.4 Summary

19. Sometimes, the opposite may occur, that is, the simple denominal verb may come from the corresponding noun in the construction, for instance, *give an example = exemplify*.

20. *To present* may also mean *make a gift*, and the use of the analytic expression using the same support verb helps disambiguating.

The study of MWEs is almost as old as linguistics itself. When trying to classify linguistic phenomena into either lexical or syntactic, one quickly realises that some of them, and in particular MWEs, lie in between these two levels. Therefore, there are limitations in the structural approach to language à la Chomsky and Tesnière. One of the seminal papers of the *construction grammar* trend is the work of Fillmore et al. (1988). They illustrate and discuss in detail the weaknesses of this idealised atomistic approach to grammar. In construction grammar, idioms are part of the core of the grammar: a language can be fully described by its idioms and their properties. These idioms correspond to what we call MWEs. Another linguistic theory that gives much importance to MWEs is the *meaning-text theory* (MTT). MWEs are present at two points of the computational MTT model: as *phrasemes* and as lexico-semantic functions in the *lexical combinatorics zone*. An account of MWEs in different linguistic theories is provided in Seretan (2008, p. 20–27).

MWEs are hard to define, as there is little agreement on the definition of the word *word* itself. The notion of MWE has its origin in Firth's famous quotation "you shall know a word by the company it keeps". He affirms that "collocations of a given word are statements of the habitual and customary places of that word" (Firth 1957, p. 181). Smadja (1993) considers collocations as "arbitrary and recurrent word combinations". For Choueka (1988), a collocation is "a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components". For Fillmore et al. (1988, p. 504), "an idiomatic expression or construction is something a language user could fail to know while knowing everything else in the language". Sag et al. (2002) generalise this same property to vaguely define MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)".

All these definitions are valid in a given experimental setup. Although, the definition of MWE adopted will influence the evaluation results, as it will be used to write annotation guidelines and/or to select reference lists. Therefore, in this thesis, we adopt the definition by Calzolari et al. (2002). For us, MWEs are "[...] different but related phenomena [...]. At the level of greatest generality, each of these phenomena can be described as a [combination] of words that acts as a single unit at some level of linguistic analysis." This generic and intentionally vague definition can be narrowed down according to the application needs. For example, for an MT system, a MWE is any combination of words which, when not translated as a unit, generates unnatural output. The level at which a combination of words needs to be treated as a unit varies according to the type of system and expression.

The literature describes some common properties of MWEs in general: arbitrariness, institutionalisation, limited semantic variability (non-compositionality, non-substitutability, no word-for-word translation, domain-specificity/idiomaticity), limited syntactic variability (extragrammaticality, lexicalisation), and heterogeneity. These are not binary yes/no flags, but values in a continuum going from completely flexible, ordinary word combinations to totally prototypical and/or fixed expressions.

There exist several typologies to classify MWEs, based on different views on grammatical theories: constructionism, meaning-text theory, grammar engineering, and automatic MWE acquisition. In this work, we propose a typology based firstly on the morphosyntactic role of the whole expression in a sentence, and secondly on its difficulty to be dealt with using computational methods. The first typology classifies MWEs into nominal, verbal and adverbial/adjectival expressions. Nominal expressions cover noun compounds (*Russian roulette*), proper names (*Porto Alegre*), and multiword terms (*DNA-*

binding domain). Verbal expressions include phrasal verbs (*give up*) and light verb constructions (*take a shower*). Adverbial and adjectival expressions include examples such as *upside down* in English, *à poil* in French, and *sem mais nem menos* in Portuguese. In addition to these types, we define three orthogonal types that are more related to the computational methods used to treat MWEs: (i) fixed expressions like *in short*, (ii) idiomatic expressions like *dry run*, *put in place* and *on the same wavelength*, and (iii) "true" collocations, corresponding to fully compositional expressions co-occurring more often than expected by chance. This typology is quite simple and yet rigorous enough to describe the MWEs dealt with in our experiments.

3 STATE OF THE ART

In the previous chapter, we provided the historical and theoretical foundations for the study of multiword expressions. The set of definitions, characteristics and types described give an idea of the difficulty of the computational tasks involving MWEs. The goal of the present chapter is to draw an overview of the state of the art in computational methods for MWE treatment, focusing on acquisition. State-of-the-art techniques to deal with MWEs are the starting point of the methodology proposed in Chapter 5. Information contained in the present chapter allow a better comparison and contextualisation of the present work with respect to the computational linguistics community.

In Section 3.1, we start with a brief review of some practical *elementary notions*, defining concepts like *n*-grams, frequencies and association measures. These are the tools used by the techniques for automatic *MWE acquisition* described in Section 3.2. Although the largest part of research effort in the community has been devoted to acquisition, other tasks such as interpretation, disambiguation and representation are also relevant. Mainly in the last decade, work on these tasks started to emerge, and this is presented in Section 3.3.

3.1 Elementary notions

In this section, we review standard NLP concepts useful in the present work. We focus on general notions that appear recurrently throughout the thesis, while more detailed explanations of concepts specific to a certain experiment are provided later, whenever they are required. [1]

We define a *corpus* as a body of texts used in empirical language studies (Manning and Schütze 1999, p. 6). One usually wants for corpora to be representative of the target language, where the meaning of *representative* depends on the context (e.g., application, domain, genre). In our experiments, we use only written corpora like the collection of speech transcripts from the European parliament (Koehn 2005), the 100-million words British National Corpus (Burnard 2007) and the collection of news from the Brazilian *Folha de São Paulo* newspaper. [2] Intuitively, half a dozen of sentences in French are not a big enough corpus of general French language, as well as a million sentences of computer

1. The goal of this section is not to provide a substantial introduction to empirical methods in computational linguistics. Instead, we remind and try to disambiguate as much as possible the definitions of concepts that are already familiar to the reader to some extent. If this is not the case, we recommend Jurafsky and Martin (2008) as a consolidated and wide introduction to NLP and Manning and Schütze (1999) for a more specific introduction to empirical methods. Our text is inspired on these two standard reference textbooks.

2. More detailed descriptions of the corpora used in our experiments can be found in Appendix D

science articles in English are not representative if the target application will deal with botany texts or with texts in Portuguese. The WWW can be used as a huge corpus, please refer to Appendix F for more details.

Usually, following the standard methodology of empirical NLP, part of the corpus is used as *training set* while a small part is held out as *test set*. A corpus may contain data in one language (monolingual) or in several languages (multilingual); when the sentences in one language are translations of sentences in another language, we consider it as a sentence-aligned *parallel corpus*. We will also use the term *general-purpose* corpus to refer to corpora that contain a wide variety of texts corresponding to most common language use over a given time span, while a *specialised* corpus contains texts of a specific knowledge domain or sub-language, like botany, computer science or sailing. We consider a *word token* to be an occurrence of a word in a corpus while a *word type* is a unique occurrence of a word as a lexeme in a dictionary, thesaurus or other lexical resource. The set of unique word types in a corpus constitutes its *vocabulary*.

In Section 3.1.1, we introduce linguistic notions such as part of speech and dependency syntax. We provide an overview of the statistical distribution of words in a corpus in Section 3.1.2. Section 3.1.3 deals with *n*-gram language models, presenting the basics of probability estimation through the chain rule, the Markov assumption and the data structures used to represent *n*-grams. In Section 3.1.4, we present statistical tools used in the automatic acquisition of MWEs, namely, lexical association measures.

3.1.1 Linguistic processing: analysis

Linguistic analysis is the process of creating more abstract representations from raw text. It is generally seen as a set of steps, each of which must solve ambiguities inherent to language. However, more sophisticated systems should not try to *solve* ambiguities, but represent multiple solutions in the form of weighted lattices. However, for concision purposes, we present here an over-simplified example of analysis steps which can be applied on corpora for MWE acquisition.

A corpus may be structured as a set of documents, each document being composed of several paragraphs, which in turn are sequences of sentences. While the higher level divisions are optional and task-dependent, most of the current NLP systems require the text to be split into sentences prior to processing. Splitting the sentences in running text can be accomplished through language-dependent regular expressions on anchor punctuation signs (such as periods and question marks) and lists of common exceptions like abbreviations (*Ph.D.*), acronyms (*Y.M.C.A.*) numbers (*1,399.99*), proper names (*Yahoo!*), filenames and web addresses (*www.google.com*). Although apparently simple, sentence splitting is challenging in highly structured texts like scientific articles containing many tables, itemised lists and mathematical formulas. Already at the sentence splitting level, ambiguities about possible splitting points must be dealt with by the system.

Further decomposition takes us from sentences to words. The definition of *word* is discussed in Section 2.2.1. In practice, for languages whose writing system uses spaces to separate words, one also needs to split from adjacent words punctuation such as commas, periods, apostrophes, dashes and contractions (e.g., the English possessive marker *'s*). It may also be necessary to split contractions such as *du* = *de* + *le* in French and *no* = *em* + *o* in Portuguese. [3] Other morphological phenomena like prefixes, suffixes and agglutination can also be dealt with at this point. The process of word splitting is called *tokenisation*,

3. Contraction identification usually requires context-aware analysis. For instance, in French, the contraction *des* = *de* + *les* is homonym to the partitive article *des*.

and is generally accomplished using simple regular expressions. For example, consider the sentence:

Example 3.1 *"Tomorrow, I'll be paying a visit to Mary's parents."*

After tokenisation, it becomes: [4]

Example 3.2 *"␣Tomorrow␣,␣I␣'ll␣be␣paying␣a␣visit␣to␣Mary␣'s␣parents␣.␣"*

A word in the corpus occurs in its inflected form, also called the *surface form*. A surface form like *parents* is the plural of the base form *parent*, the verb *paying* is the gerundive of *pay*, and so on. The morphology of languages is responsible for word formation and inflection, and the latter often encodes information such as gender, number, tense, mode, person and case of words. Moreover, distinctive capitalisation marking the beginning of a sentence, for example, needs to be normalised so that *Tomorrow* and *tomorrow* are considered as being the same word. [5] The base form from which an inflected word is derived is called *lemma*. The process of assigning lemmas to words is called *lemmatisation*.

Generally, lemmatisation is performed simultaneously with another process called *part-of-speech tagging*. The latter is the assignment of part-of-speech (POS) tags to each word. POS tags are useful, for example, to distinguish closed-class words or *function words* like prepositions, pronouns and determiners, from open-class words or *content words* like verbs, nouns, adverbs and adjectives. The set of all POS tags that can be assigned to words in a language is the *tagset*, and the software performing POS tagging and, usually, lemmatisation, is the *POS tagger*. In some of our experiments, we use a POS tagger called TreeTagger, described in Appendix D (Schmid 1994). When the English module is applied to the sentence of Example 3.1, the system performs sentence splitting, tokenisation, lemmatisation and POS tagging, resulting in:

Example 3.3 *" $tomorrow_{[NN]}$, $I_{[PP]}$ $will_{[MD]}$ $be_{[VB]}$ $pay_{[VBG]}$ $a_{[DT]}$ $visit_{[NN]}$ $to_{[TO]}$ $Mary_{[NP]}$ $'s_{[POS]}$ $parent_{[NNS]}$. "*

The process of going from surface forms to more abstract representations like POS and lemmas is called *analysis*. In order to perform a deeper analysis, one can group POS-tagged words into chunks like noun phrases, and represent chunks as part of a *syntax tree*. There are countless formalisms to represent syntactic structures in theory and in practice. The one adopted in this thesis is *dependency syntax*. In dependency syntax, the nodes of the tree are the words themselves and the arrows are the dependency relations, tagged with the corresponding relation type (e.g., direct object, subject, determination). A software capable of generating such trees from sentences is a *dependency parser*. For English, we use the RASP parser, described in Appendix D. RASP generates the following surface dependency tree [6] for the example sentence:

4. We use the character ␣ only to emphasise the spaces between words.

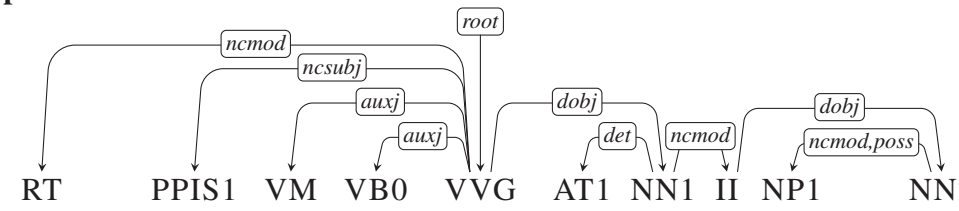
5. However, it is not enough to lowercase the whole text as case information may be important, for instance, in domain-specific texts (chemical element *NaCl*), acronyms (*NASA*, *CIA*) and to distinguish named entities (*Bill Gates*, *March*) from common words (*pay the bill*, *open the gates*, *the soldiers march*).

6. Actually, RASP does not generate dependency relations directly, but it infers *grammatical relations* using equivalence rules applied to a traditional constituent parsing tree. They claim that the relations are mostly acyclic and exceptions can be dealt with on a case by case basis.

# of sentences	20,000
# of tokens	414,602
# of types	37,649

Table 3.1: Statistics of BNC-frg — Sample of 20,000 random sentences taken from the BNC

Example 3.4



Notice that, for the dependency parser, some nodes like punctuation are ignored, as they are considered irrelevant for syntax. [7] Also, the tagset and the lemmas used by the RASP parser [8] for morphosyntactic analysis are more fine-grained than those of the TreeTagger. Finding equivalences and adapting the granularity of POS tags is a practical problem in NLP and demands some manual trial-and-error work.

Many other parsers and formalisms exist for English and for other languages, and related work on MWE acquisition explores some of them, as we will discuss in Section 3.2.1. Nonetheless, we will limit our discussion to dependency parsing because it is the formalism used in our experiments. The advantage of the dependency formalism is that the resulting tree can be represented on a word basis, that is, for every word we assign two labels: the other word of the sentence on which it depends and the type of the relation. This has practical implications in the data structures used to represent parsed corpora, as we will discuss in Chapter 5. Moreover, more meaningful relations such as subject and object tend to appear closer to the root while auxiliaries and determiners appear as leafs, as shown in Example 3.4.

3.1.2 Word frequency distributions

In order to design statistical methods for dealing with corpora, one needs to understand how words and word counts behave in text. We will use as a toy example a fragment of 20,000 English sentences randomly chosen from the British National Corpus, henceforth BNC-frg. Table 3.1 summarises the number of tokens and types in the toy corpus. It can be considered as a sample of English language, and therefore the size N of the sample is the number of tokens, that is, around 414K tokens (surface forms). The vocabulary V is a set containing around 37.6K distinct word types. [9]

7. This is a simplification, as described by Briscoe et al. (2006).

8. Documentation about RASP's tagset and grammatical relations is available at `http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-662.pdf` and in Appendix C of Jurafsky and Martin (2008).

9. The *type/token ration*, that is, the number of types with respect to the number of tokens in a text, has been used as a measure of the richness of the vocabulary. However, this measure is not a good one because it depends on the corpus size (Baayen 2001). In BNC-frg, the type/token ratio is of 0.091.

r	$c(w)$	w	r	$c(w)$	w	r	$c(w)$	w
1	20,765	the	11	3,248	for	21	2,173	you
2	19,031	,	12	3,064	it	22	2,146	’
3	16,022	.	13	2,996	was	23	2,029	’
4	11,923	of	14	2,899	’s	24	1,899	by
5	9,830	to	15	2,816	I	25	1,800	are
6	9,771	and	16	2,550	on	26	1,782	at
7	7,346	a	17	2,535	be	27	1,727	have
8	6,758	in	18	2,405	with	28	1,668	not
9	4,351	that	19	2,356	as	29	1,532	from
10	4,029	is	20	2,255	The	30	1,496	he

Table 3.2: Counts of the 30 most frequent tokens in BNC-frg.

Thus, let us define a function $c(w) : V \to \mathbb{N}$ which associates to each word type w in a vocabulary its number of occurrences in a corpus. In order to avoid ambiguity, when more than one corpus is considered simultaneously, the function is subscripted with the name of the corpus in which the token was counted, for instance, $c_{\text{BNC-frg}}(\textit{Mary}) = 18$.

The values of $c(\cdot)$ for the 30 most frequent tokens of BNC-frg are listed in Table 3.2. The most frequent words in any corpus are generally function words like prepositions, determiners, pronouns and punctuation signs. Notice that, as our corpus was not analysed, the words *the* and *The* are considered as two distinct tokens. Also, because of an encoding problem, there are two different apostrophe characters.

At this point, we need to clarify a point on the nomenclature: the definition of the word *frequency*. In the French statistical literature, the term *nombre d’occurrences* is used to denote how many times an event occurs, that is, the value of counting function $c(\cdot)$. The term *fréquence* is defined as the proportion obtained when one divides the count of the event by the total number of events N. For instance, using the French nomenclature, the frequency of the event *Mary* in the sample BNC-frg is $\frac{c(\textit{Mary})}{N} = \frac{18}{414{,}602} = 0.000043415$. In the Anglo-Saxon statistical tradition, however, the word *frequency* is used as a synonym of number of occurrences. According to this nomenclature, when one divides the number of occurrences by the total number of events, one obtains the *relative frequency* of that event. As a consequence, in many textbooks, function c is referred to as f (Jurafsky and Martin 2008, Manning and Schütze 1999, Baayen 2001). Thus, according to the Anglo-Saxon nomenclature, the frequency of *Mary* in the corpus is not 0.000043415, but 18. In the present thesis, in order to avoid ambiguity, we will adopt the following nomenclature conventions:

- $c(\cdot)$ denotes the *count* or *number of occurrences* of a token;
- the ratio $\frac{c(\cdot)}{N}$ denotes the *relative frequency* of a token;
- the term *frequency* will be avoided because of its promiscuous use in the Anglo-Saxon literature;
- $p(\cdot)$ denotes the *probability* of a token or n-gram. It will not be used for the statistical description of empirical data. Instead, it will be reserved for a value provided by some underlying theoretical model, that is, a probability density function depending on a certain number of parameters;

In Section 3.1.3, our goal is to estimate the probability of an arbitrary token or se-

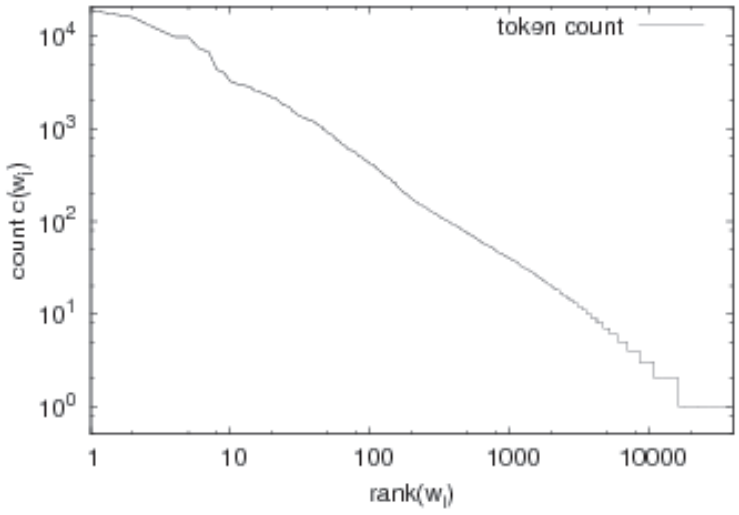


Figure 3.1: Rank plot of the vocabulary of BNC-frg, with counts in descending order.

quence of tokens. Therefore, it is useful to study the empirical distribution of the values of function $c(\cdot)$. Unlike the heights of humans or the grades of students in a class, the word counts in a corpus are not normally distributed around the mean. Instead, they are distributed according to a *power law* distribution, also known as *Zipfian distribution*. Many other events in the world are distributed according to power laws (Newman 2005).

In order to illustrate the Zipfian distribution, we will use the rank plot of Figure 3.1. A rank plot is a graphic where the word counts are sorted in descending order and assigned to their rank positions r, like those in the first column of Table 3.2. Formally, the rank r of a given word w can be defined as the value of a bijective function $rank(w) : V \rightarrow [1..|V|]$ which assigns a distinct integer to each word respecting the constraint $(\forall w_1, w_2 \in V)[rank(w_1) \leq rank(w_2) \iff c(w_1) \geq c(w_2)]$. Notice that, in the presence of ties, the *rank* function is not uniquely defined. Any valid function respecting the aforementioned constraint could be used. Therefore, we assume that lexicographic order is used to assign the ranks of words with identical numbers of occurrences, thus uniquely defining the *rank* function.

The rank plot of Figure 3.1 is in logarithmic scale, otherwise it would be impossible to visualise the counts. The main characteristic of power laws is that there is a very large number of rare events. In BNC-frg, for example, there are 21,423 word types occurring only once in the corpus, [10] that is, almost 57% of the vocabulary. On the opposite side of the rank, frequent words correspond to a tiny portion of the vocabulary. The graphic shows that the number of words decreases exponentially in the ranked set. In other words, Zipf's law states that the number of occurrences of a type in the corpus is inversely proportional to its position in the rank.

A derived property is that the size of the vocabulary increases logarithmically with the size of the corpus. This is exemplified by plotting the number of tokens and of types in a corpus (and its corresponding vocabulary) as its size, that is, the number of sentences, increases. Figure 3.2 shows these two measures for increasing corpus sizes, from 50 to 2,000 sentences. While the number of tokens increases linearly with the number of sentences, the number of types increases fast at the beginning and much slower afterwards.

10. A word occurring once in the corpus is often called a *hapax*, from the Greek *hapax legomena*.

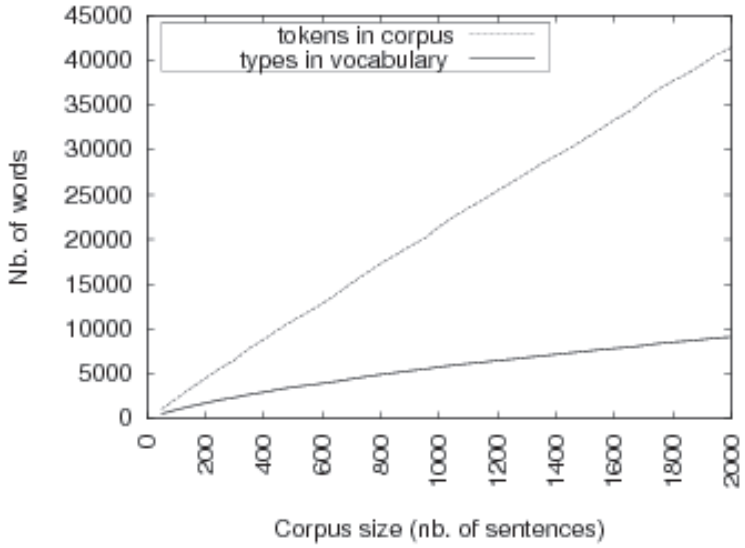


Figure 3.2: Tokens in the corpus versus types in the vocabulary.

This happens because, as the sample increases, many common words tend to be repeated while new words become harder to find. One could assume that, if an infinitely large corpus (the sample) was available, the size of the vocabulary would stop growing at some point, converging to the total number of words used in that language (the population).

The statistics of a corpus depend on its size, language, genre and domain, but the logarithmic relation between the number of word tokens and the number of different word types holds as well as Zipf's law. This means, for instance, that in general more than half of the words in a corpus occur only once. Distributions like these are called *large number of rare events* (LNRE). When the underlying model is a LNRE distribution, specific statistical tools able to deal with sparse data must be employed. Besides, one needs to be careful because standard assumptions for a sample drawn from a population normally distributed do not apply to corpora. Operations like parameter estimation, hypothesis testing and the like need to be adapted when working with LNRE distributions. For further details, one may refer to Baayen (2001).

3.1.3 *N*-gram language models

When we consider word sequences, each token in the corpus is represented as w_i, where the subscript i stands for its position with respect to other tokens. For instance, a sequence of n consecutive tokens in the corpus can be represented as $w_1 w_2 \dots w_{n-1} w_n$. Such contiguous sequences are called n-grams. [11, 12] We use the abbreviated notation w_i^j to represent an n-gram formed by $j-i+1$ words w_i through w_j. By extension, the function $c(\cdot)$ can be applied to n-grams and returns the number of times they occur in a corpus. For example, $c_{\text{BNC-frg}}(I\ will\ be) = 5$ because this 3-gram occurs 5 times in the corpus BNC-frg.

A *language model* (LM) is a tool that determines to what extent a sequence of words belongs to a certain language. An n-gram LM is a set of probability density functions

11. Discontiguous sequences are sometimes referred to as *flexigrams*, that is, n-grams with gaps.

12. In the field of statistical MT, the term *phrase* is employed to denote a sequence of contiguous words. In this thesis, we refer to n-gram to denote a sequence of words in the context of MWE acquisition, and to phrase to denote a sequence of words in the phrase table of a statistical MT system.

that estimate the probabilities of any *n*-gram in a language. For instance, an *n*-gram like *I will be paying* is more plausible in English than *will paying I be*, thus the model will assign it a higher probability. LMs are widely employed not only in MWE acquisition but in many other NLP applications like speech recognition and MT. They are often used to choose among several possible outputs because they simulate grammatical and semantic preferences in sentences.

3.1.3.1 Probability estimation

One way to estimate *n*-gram probabilities is to learn them from a sample of language: the training corpus. A very simple language model can count all the *n*-grams in the training corpus and then return as a probability estimates the relative frequencies of the *n*-grams. [13] For instance, according to Table 3.1, the BNC-frg corpus contains 414,602 tokens or 1-grams. If we use BNC-frg as training corpus, the probability estimate p of the 1-gram *Mary* is $p(Mary) \approx \frac{18}{414,602}$ and the probability estimate of the 3-gram *I will be* is $p(I\ will\ be) \approx \frac{c(I\ will\ be)}{N} = \frac{5}{414,602}$.

Although a good idea in theory, it is not feasible to store all the counts for each distinct *n*-gram of arbitrary length (1 to N) in a large corpus, as the number of *n*-grams grows quickly. [14] In order to solve this practical problem, we first apply the probability chain rule, that is, for an arbitrary *n*-gram:

$$p(w_1^n) = p(w_1) \times p(w_2|w_1) \times p(w_3|w_1^2) \ldots p(w_n|w_1^{n-1}) = p(w_1) \times \prod_{k=2}^{n} p(w_k|w_1^{k-1}) \quad (3.1)$$

We further simplify calculations by applying the Markov assumption in order to approximate the conditional probability of a token given a short history instead of using the whole preceding sequence. That is, given $m > 1$ as the fixed maximum size of *n*-gram that we can store, [15] we ignore all words preceding w_{k-m+1}. This simplification assumes that the presence of a word only depends on a short number of words to the left of it, completely ignoring the right context. [16] The advantage of applying the Markov assumption is that we only need to store the probability estimates for *n*-grams of fixed length m, that is, w_k and the preceding $m-1$ words (for $m > 1$ and $k \geq m$): [17]

$$p(w_k|w_1^{k-1}) \approx p(w_k|w_{k-m+1}^{k-1}) \quad (3.2)$$

Thus, by replacing Equation 3.2 in Equation 3.1, we obtain that the probability estimate of an arbitrary-length *n*-gram depends only on the probability estimates of smaller *m*-grams, that is:

$$p(w_1^n) = p(w_1) \times \prod_{k=2}^{n} p(w_k|w_{k-m+1}^{k-1}) = p(w_1) \times \prod_{k=2}^{n} \frac{p(w_{k-m+1}^{k})}{p(w_{k-m+1}^{k-1})} \quad (3.3)$$

13. In theory, a corpus with N tokens contains $N-n+1$ *n*-grams. However, as $n \ll N$, we can safely assume that $N-n+1 \approx N$.

14. For instance, if we consider sentence start (<*s*>) and sentence end (</*s*>) markers as tokens, BNC-frg contains 37,651 1-grams, 210,183 2-grams, 346,450 3-grams and so on.

15. The order m of the model typically ranges from 2 to 5 according to the target application.

16. Linguistic studies demonstrates that this assumption makes sense, as in speech the right context (what is going to be said after the current word) is always unknown.

17. Limit cases for $k < m$ would, according to the formula, yield undefined probability estimates like $p(w_2|w_{-1}^1)$. In such cases, we simply assume that the words with indices lower than zero should be removed from the formula, for instance, if $k = 2$ and $m = 4$, $p(w_2|w_{-1}^1) = p(w_2|w_1)$.

Now, we can estimate the conditional probabilities through the relative n-gram frequencies $p(w_j^k) = \frac{c(w_j^k)}{N}$ and thus:

$$p(w_1^n) = \frac{c(w_1)}{N} \times \prod_{k=2}^{n} \frac{\frac{c(w_{k-m+1}^k)}{N}}{\frac{c(w_{k-m+1}^{k-1})}{N}} = \frac{c(w_1)}{N} \times \prod_{k=2}^{n} \frac{c(w_{k-m+1}^k)}{c(w_{k-m+1}^{k-1})} \tag{3.4}$$

For instance, let us consider a model of order $m = 2$, built using the BNC-frg corpus as training data. Given this model, we want to estimate the probability of the 4-gram *I will be visiting*. Thus, $p(I\ will\ be\ visiting) = p(I) \times p(will|I) \times p(be|will) \times p(visiting|be) = \frac{c(I)}{N} \times \frac{c(I\ will)}{c(I)} \times \frac{c(will\ be)}{c(will)} \times \frac{c(be\ visiting)}{c(be)} = \frac{2{,}816}{414{,}602} \times \frac{34}{2{,}816} \times \frac{312}{1{,}093} \times \frac{1}{2{,}535} = 0.000000009$. [18]

This model uses the principle of *maximum likelihood estimation* (MLE), that is, it assumes that the sample *is* the population. In other words, the chosen model parameters are those that maximize the likelihood of the observed sample. The probability estimates p are given by relative frequencies $\frac{c(\cdot)}{N}$ on the training corpus, so that for each size of n, $\sum_{w_1^n} \frac{c(w_1^n)}{N-n+1} = \frac{N-n+1}{N-n+1} = 1$. This means that, even though each n-gram length has a probability space, the probability estimates summed over all possible n-gram lengths sum up to a value larger than one. Therefore, the probabilities returned by the LM as a whole do not constitute a probability space.

The problem with MLE is that it does not take into account n-grams that were not observed in the corpus as a side effect of sampling a very large event space. In other words, no matter how large a training corpus is, a large number of perfectly valid n-grams will surely be missing from the model, thus yielding zero probability for the whole product.

In order to solve this problem, current LM tools implement sophisticated *smoothing* techniques. The idea of smoothing is to assign some probability mass to unseen events, discounting it from the probabilities of seen n-grams (Chen and Goodman 1999, Good 1953, Kneser and Ney 1995). Furthermore, it is also possible to use *backoff* in order to estimate the probabilities of larger unseen n-grams by combining the probabilities of smaller n-grams contained in them. Because such techniques are rarely employed in empirical MWE acquisition from corpora, we will not discuss their details here. One of the rare works concerning smoothing for MWE acquisition is that of Lapata (2002).

3.1.3.2 Data structures

When dealing with very large corpora, it is crucial to have efficient access to n-gram counts in order to estimate their probabilities. The intuition behind quick access to n-gram counts in a corpus is to organise the n-grams in a data structure that allows fast search (that is, direct access or binary search). For example, for a 1-gram LM, we could store it as a hash table that associates each word (the key w_i) with a number of occurrences (the value $c(w_i)$). This structure is fairly simple, allows constant-time access and fits into memory. Unfortunately, it is not scalable for larger values of n.

N-gram models with a fixed order m can be represented using structures based on suffix trees. A *suffix tree* is a representation in which each edge is labelled with a word

18. In practice, in order to avoid numerical underflow and to speed up computations, one usually sums the logarithm of the probability estimates instead of directly calculating this product.

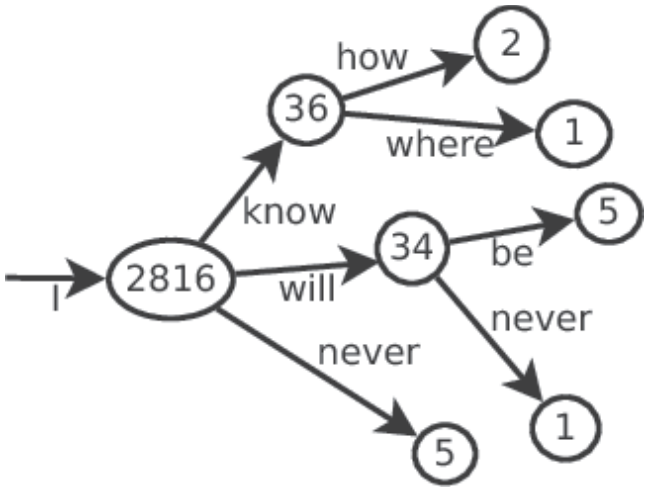


Figure 3.3: Example of suffix tree.

and each node contains a count. Concatenating the words on the edges of a path from the root to a node n_i generates an n-gram whose count is stored in n_i. For example, in Figure 3.3, the path *I will be* leads to a node containing the value of $c(\text{I will be}) = 5$. In order to optimise the access to the child nodes, it is possible to build hash tables for constant access or ordered lists for binary search.

Not only speed, but also memory consumption needs to be minimised. Therefore, we can use a compact representation in which each word is assigned to a 4-byte integer that uniquely identifies it. When a corpus or n-gram is read from a file, the vocabulary hash table assigns an integer identifier to each word and the remainder of the processing only considers integer identifiers instead of strings. To make comparisons easy, the identifiers can be assigned to words in such a way that lexicographic order is preserved. Thus, for each pair of words, if a word lexicographically precedes another, it will also have a lower integer identifier.

While suffix trees are appropriate for LMs with fixed order, counting arbitrarily long n-grams requires another kind of data structure. A *suffix array* is an efficient structure to represent n-grams of arbitrary size (Manber and Myers 1990, Yamamoto and Church 2001). The corpus is viewed as an array of N words w_1 to w_N. Each word w_i is the beginning of a corpus suffix of size $N-i+1$, for instance, $w_{N-2}w_{N-1}w_N$ is a suffix of size 3. The trick is that the list containing all the N suffixes is sorted in lexicographic order. Therefore, one can perform binary search in order to locate the first and the last positions starting with the searched n-gram. For example, in Figure 3.4 we represent part of a suffix array of BNC-frg. If we want to know how many times the n-gram *I will be* occurs in the corpus, we will perform two binary searches in $O(\log N)$ time to find the first index F and last index L in the array containing a suffix which starts with the searched n-gram. The number of occurrences of the n-gram is then simply $L-F+1 = 108-104+1 = 5$. If now we need to obtain the count for *I will*, we repeat the procedure and find $133-100+1 = 34$.

In the actual implementation (see Section 5.1.2) each suffix is represented with an integer index pointing to the position in the corpus where it starts, thus optimizing memory use. Thus, a suffix array uses a constant amount of memory with respect to N: if every word and every word position in the corpus is encoded as a 4-byte integer, a suffix array uses precisely $4 \times 2 \times N$ bytes, plus the size of the vocabulary, which is generally very small if compared to N.

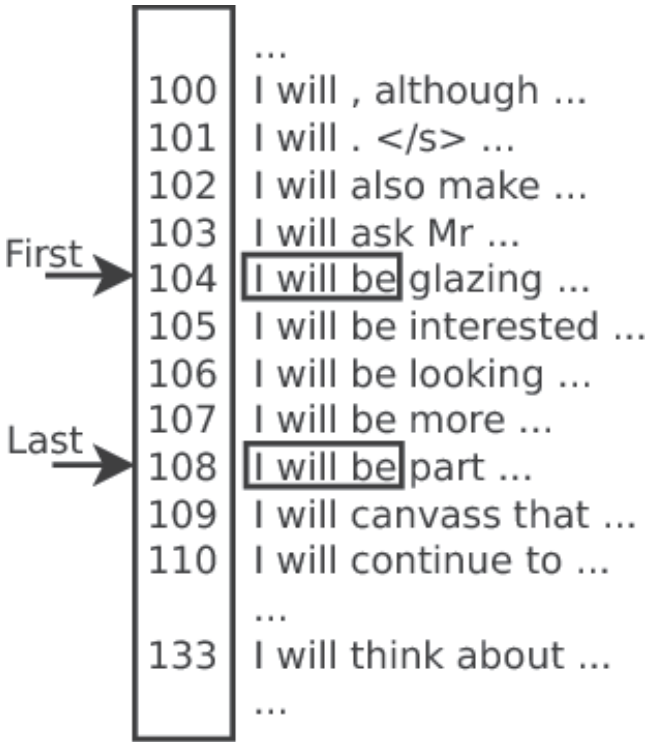


Figure 3.4: Example of suffix array.

3.1.4 Lexical association measures

The principle of corpus-based MWE acquisition is that words that form an expression will co-occur more often than if they were randomly combined by a coincidence of syntactic rules and semantic preferences. In this context, lexical association measures are applied to *n*-gram counts in order to estimate how much the occurrences of two or more words depend on each other. In this section, we will explain how this is possible and illustrate it with examples.

A simple and intuitive method to acquire MWEs from corpora is to use ranked *n*-gram lists. For example, Table 3.3 lists the 15 most frequent *n*-grams of BNC-frg. Unfortunately, all of the returned items are uninteresting combinations of function words like determiners *the* and *a*, prepositions and auxiliary verbs *be* and *have*. Moreover, the list only contains 2-grams and no 3-grams and larger *n*-grams. This is a consequence of the fact that the count of a larger *n*-gram will always be less than or equal to the count of the *n*-grams that it contains, thus biasing the acquisition towards short *n*-grams.

We could solve these problems by separately acquiring *n*-grams of different lengths, using regular expression patterns to filter out sequences of function words contained in stopword lists or matching unwanted POS tags. This is actually performed in many real-life systems, specially for automatic terminology acquisition, with surprisingly good results (Justeson and Katz 1995, Ramisch 2009). However, if we are to acquire general MWEs (and not only multiword terms), we need a more sophisticated way to tell whether an *n*-gram is just a random co-occurrence of frequent words or whether there is some statistical idiosyncrasy about it, that deserves further analysis.

A common preprocessing step when dealing with *n*-gram counts is to eliminate all combinations that occur less than a fixed threshold. This is very important because statistics tend not to be reliable in low frequency ranges. As the counts decrease, it is impossi-

r	$c(w_1w_2)$	w_1w_2	$c(w_1)$	$c(w_2)$	$E(w_1w_2)$	t-score
1	3060	of the	11923	20765	597.2	44.5
2	1788	in the	6758	20765	338.5	34.3
3	1139	to the	9830	20765	492.3	19.2
4	772	on the	2550	20765	127.7	23.2
5	738	and the	9771	20765	489.4	9.2
6	733	to be	9830	2535	60.1	24.9
7	687	for the	3248	20765	162.7	20.0
8	526	at the	1782	20765	89.3	19.0
9	525	by the	1899	20765	95.1	18.8
10	500	that the	4351	20765	217.9	12.6
11	473	of a	11923	7346	211.3	12.0
12	457	from the	1532	20765	76.7	17.8
13	456	with the	2405	20765	120.5	15.7
14	369	it is	3064	4029	29.8	17.7
15	362	in a	6758	7346	119.7	12.7

Table 3.3: Top-15 most frequent n-grams in BNC-frg.

ble to distinguish statistically significant events from coincidences due to sampling error. Unfortunately, there is no rule or algorithm for determining the value of such threshold except common sense and trial and error. For example, statistics calculated over hapax are surely unreliable while setting the threshold at 100 occurrences will probably result in too little data (if any).

Now, in order to investigate whether an n-gram is a MWE, let us assume that words are combined randomly. That is, the occurrence of words at a given position are independent events. This hypothesis does not hold, otherwise languages would not have grammar. Nonetheless, it provides a powerful way to test the association strength between words. By the definition of statistical independence, if the occurrence of a word w_2 does not depend on the occurrence of the preceding word w_1, then we expect that the joint probability of the 2-gram is the product of the probabilities of the individual events, that is:

$$p(w_1^2) = p(w_1) \times p(w_2) \tag{3.5}$$

For the sake of simplicity, let us use MLE estimators for the probabilities of the individual words through relative frequencies, that is $p(w_i) = \frac{c(w_i)}{N}$. Then, for an arbitrary n-gram w_1^n, the expected relative frequency would be the probability:

$$p(w_1^n) = \frac{c(w_1)}{N} \times \frac{c(w_2)}{N} \times \ldots \times \frac{c(w_n)}{N} = \frac{c(w_1) \times c(w_2) \times \ldots \times c(w_n)}{N^n} \tag{3.6}$$

We can scale this probability estimate by the approximate number of n-grams in the corpus ($N - n + 1 \approx N$) to obtain the expected count $E(w_1^n)$:

$$E(w_1^n) = N \times \frac{c(w_1) \times c(w_2) \times \ldots \times c(w_n)}{N^n} = \frac{c(w_1) \times c(w_2) \times \ldots \times c(w_n)}{N^{n-1}} \tag{3.7}$$

Column 6 of Table 3.3 shows the values of $E(w_1^n)$ for the top-30 most frequent 2-grams in BNC-frg. Combinations of frequent words are expected to occur frequently

while combinations involving rarer words are expected to occur less. One way to test whether the difference between the expected count $E(\cdot)$ and the observed count $c(\cdot)$ is statistically significant is to use a hypothesis test. In theory, we should perform an exact binomial test that models the discrete distribution of n-gram counts (Evert 2004). In practice, however, this test is computationally costly and it is possible to approximate it using a z test.

A very common test employed in MWE acquisition is *Student's t test*, a heuristic variation of the z test in which the standard deviation of the sample is estimated through its observed count $c(w_1^n)$ rather than from the expected count. This approximation holds if we consider the corpus as a sequence of randomly generated n-grams and a Bernoulli trial that assigns 1 to the occurrence of w_1^n and 0 otherwise. Then, the probability p of generating 1 is the mean of the sample, $\bar{x} = p = \frac{c(w_1^n)}{N}$. For small values of p, $s^2 = p \times (1-p) \approx p$, thus the standard deviation of the sample s^2 is equivalent to the mean $\bar{x}$. Finally, the estimated theoretical mean μ is the normalised estimated count $\frac{E(w_1^n)}{N}$, thus yielding the following formulation for the t statistic: [19]

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} = \frac{\frac{c(w_1^n)}{N} - \frac{E(w_1^n)}{N}}{\sqrt{\frac{c(w_1^n)}{N^2}}} = \frac{c(w_1^n) - E(w_1^n)}{\sqrt{c(w_1^n)}} \tag{3.8}$$

As we have seen in Section 3.1.2, word counts do not follow a normal distribution, but they can be modelled using a power law distribution, and the same applies to n-grams. As a consequence, from a theoretical perspective, the application of Student's t test here does not make sense as it assumes that the $c(w_1^n)$ follows a normal distribution. Nonetheless, most of the time in MWE acquisition, our goal is to rank candidate n-grams according to their association strength. Thus, the value of the t test statistic is not used to calculate the p-value, but is used directly as a ranking criterion. This ranking measure is called the `t-score`, and it is interpreted as follows: a large value means strong word association and thus a potential MWE, a small value means that the combination is more likely to be a random word combination, thus uninteresting for MWE acquisition. Notice that, for the examples in Table 3.3, the statistic is larger when the combination is composed of rarer words.

The `t-score` is an example of lexical *association measure* (AM), that is, a numerical score that measures the degree of independence or association strength between the number of occurrences of the n-gram and the number of occurrences of the individual words that compose it. Similarly to n-gram counts, when more than one corpus is involved, we will subscribe the name of the association measure with the name of the corpus from which the counts used to calculate it were obtained, like in $\texttt{t-score}_{\text{BNC-frg}}$. In addition to the `t-score`, there are many other proposed measures in the literature. Church and Hanks (1990), for instance, suggest to use *pointwise mutual information* (`pmi`), a notion coming from information theory which measures the predictability of a word given the preceding words:

$$\texttt{pmi} = \log_2 \frac{c(w_1^n)}{E(w_1^n)} \tag{3.9}$$

Another commonly employed AM is Dice's coefficient, a classical measure used in

19. The *test statistic* is a variable with a known distribution from which we can obtain the p-value. In French, the test statistic is called *variable de décision*.

	w_2	$\neg w_2$	
w_1	$c(w_1 w_2)$	$c(w_1 \neg w_2)$ $= c(w_1) - c(w_1 w_2)$	$c(w_1)$
$\neg w_1$	$c(\neg w_1 w_2)$ $= c(w_2) - c(w_1 w_2)$	$c(\neg w_1 \neg w_2)$ $= N - c(w_1) - c(w_2) + c(w_1 w_2)$	$c(\neg w_1)$ $= N - c(w_1)$
	$c(w_2)$	$c(\neg w_2)$ $= N - c(w_2)$	N

Table 3.4: Contingency table for two random variables: the occurrence of the first word w_1 and the occurrence of the second word w_2. The notation $\neg w_i$ expresses the occurrence of any word except w_i.

information retrieval to calculate the similarity between two sets:

$$\texttt{dice} = \frac{n \times c(w_1^n)}{\sum_{i=1}^{n} c(w_i)} \tag{3.10}$$

All of the measures above are applicable to arbitrary-length n-grams, but they are mostly heuristics motivated by practical applications. However, more robust and theoretically sound AMs exist for the special case of 2-grams. These measures are based on *contingency tables*, that is, a representation like the one showed in Table 3.4, in which we consider the occurrence of two words as two random variables. We denote as $\neg w_i$ the occurrence of any word different from w_i. Notice that all the cell values are derived from the count of the 2-gram $c(w_1 w_2)$, the individual word counts $c(w_1)$, $c(w_2)$ and the total number of tokens in the corpus N. The values in the last row represent the sum of the values of the inner cells, and analogously for the last column. These are often called marginal counts because they belong to the margins of the contingency table. The value of the cell in the last row and column corresponds the number of elements in the sample N, and is equivalent to the sum of the marginal counts in both directions.

For every cell in the contingency table, it is possible to calculate the equivalent expected value if the occurrences of the two words were independent events, as follows:

$$\forall w_i \in \{w_1, \neg w_1\}, \forall w_j \in \{w_2, \neg w_2\}, E(w_i w_j) = \frac{c(w_i) \times c(w_j)}{N} \tag{3.11}$$

We can employ the χ^2 test in order to estimate whether the difference between observed and expected contingency tables is statistically significant, that is, if the word pair co-occurs more often than would be expected by chance. The X^2 test statistic is a scaled mean squared error measure between observed and expected cell values, that is, for all values of $w_i \in \{w_1, \neg w_1\}$ and $w_j \in \{w_2, \neg w_2\}$,

$$X^2 = \sum_{w_i, w_j} \frac{[c(w_i w_j) - E(w_i w_j)]^2}{E(w_i w_j)} \tag{3.12}$$

The X^2 test statistic for two random variables has an asymptotic χ^2 distribution with one degree of freedom and thus it is possible to obtain the p-value which, if sufficiently small, indicates a significant difference between the tables. However, as for the t test, usually the test statistic is considered by itself as a ranking criterion.

A very popular AM based on contingency tables is the *log-likelihood ratio* (`ll`), proposed for the first time for MWE acquisition by Dunning (1993). This measure is preferable over X^2 because, for small samples with LNRE distributions, it provides more accurate association estimators, as demonstrated through numerical simulation by Dunning (1993). The simplified version of the `ll` AM is:

$$\texttt{ll} = 2 \times \sum_{w_i, w_j} c(w_i w_j) \times \log \frac{c(w_i w_j)}{E(w_i w_j)} \tag{3.13}$$

This measure has the advantage that, in addition to being theoretically sound, numerically simple and robust to low frequencies, it has a simple interpretation. Its value equals the number of times the 2-gram is more likely under the hypothesis that the words are not independent than the individual counts would suggest. While on the one hand `ll` is robust and theoretically sound, on the other hand it is only applicable to the case where $n = 2$. Extensions to larger n-grams, although possible, are far from being intuitive (see the documentation of the NSP package, described in Section 3.2.3.1, for an example).

There are numerous AMs available for MWE acquisition. Pecina (2008b) presents a table containing 84 measures among which some are rank-equivalent to each other. The adaptation of traditional AMs for word pairs in which one word is very frequent and the other is rather rare, like it is the case for English phrasal verbs formed by rare verbs (e.g., *nail*) with frequent prepositions and adverbs (e.g., *down*), has been evaluated by Hoang et al. (2009).

Table 3.5 shows the top-15 n-grams acquired from BNC-frg as ranked by some of the AMs presented here. A threshold of at least 3 occurrences was set to reduce noise. The first measure, the `t-score`, seems to retrieve rather long specialised MWEs like proper names (*Unix System Laboratories Inc*) and terminological phraseology (*reported first quarter net profit*). The list illustrates one of the problems with n-gram based methods: the extraction of nested expressions, that is, a shorter expression like *first quarter* contained in a larger one like *first quarter net profit*. Delimiting the borders of a MWE is a current challenge in acquisition tools and methods.

The `dice` measure, on the other hand, retrieves shorter n-grams among which we find many MWE types like proper names (*Sri Lanka*, *Winston Churchill*), noun compounds (*Greenhouse Effect*, *molecular biology*), formulaic sequences (*Yours sincerely*) and fixed expressions (*per cent*, *inter alia*). Both `t-score` and `dice` tend to retrieve rarer sequences, which only occur 3 to 4 times in the corpus.

The other two measures seem to fail in extracting any interesting MWE, as they give much weight to frequent combinations of function words. The `ll` measure, however, retrieves some cases of rare double commas or double *the* determiners. Most of the applications of `pmi` and `ll` in the literature are targeted, as these AMs are used to classify possible collocates for a given fixed word, and not to blindly acquire unknown MWEs from a corpus (Dunning 1993, Church and Hanks 1990). The unfortunate reality in AMs for MWE acquisition is that sometimes the most theoretically sound measures perform worse than intuitive heuristics.

This example is an illustration of how AMs work and shows that their results are complementary, suggesting that their combination should be envisaged for broad coverage acquisition. Although there is some published work on fair comparisons among AMs (Pearce 2002, Evert and Krenn 2005, Wermter and Hahn 2006, Schone and Jurafsky 2001), this falls out of the scope of our work and is not the goal of our example. Moreover, the measures have different weaknesses: some overestimate the importance of

t-score	pmi	dice	ll
Net earnings per share amounted	of the	CHANCERY DIVISION	of the
reported first quarter net profit	in the	homoclinic orbits	in the
Microsoft Corp 's Windows NT	, and	Los Angeles	, but
(7) mm Hg	to be	Yours sincerely	to be
or fume or other impurity	, but	Greenhouse Effect	I 'm
earnings per share amounted to	on the	Hong Kong	have been
7) mm Hg in	for the	gon na	do n't
dust or fume or other	. '	inter alia	, and
has reported first quarter net	to the	Khmer Rouge	will be
[CHANCERY DIVISION]	at the	Inland Revenue	the the
Inc has reported first quarter	by the	Sri Lanka	per cent
; [1991] 2	from the	Cruz Operation	, ,
N C V O	it is	per cent	has been
Unix System Laboratories Inc	will be	molecular biology	on the
you 're gon na get	it 's	Winston Churchill	the .

Table 3.5: Top-15 *n*-grams (2 to 5) extracted from BNC-frg and ranked according to AMs.

rare *n*-grams while others are not capable of dealing with frequent items. Thus, different count thresholds should be applied for each AM, specially for such a small corpus as the BNC-frg. In addition, further cleaning of function words and punctuation is an easy step that should be performed in any case.

Besides association measures, there are also other types of statistical measures that can be used as evidence for MWE discovery in corpora. Pecina (2005), for example, discovered that context measures that consider the adjacent words of the *n*-grams are more adequate to acquire idiomatic expressions. In terminology acquisition, contrastive measures like C-NC and csMWE are employed as a way of verifying the pertinence of the *n*-gram to the target domain (Frantzi et al. 2000, Bonin et al. 2010a).

A complete survey on statistical measures for the automatic acquisition of MWEs is out of the scope of the present work. For a deeper understanding of AMs, please refer to Evert (2004), Seretan (2008), Pecina (2008b). A summary of common association measures can also be found on Stefan Evert's website `http://www.collocations.de/`.

3.2 Practical context in MWE acquisition

The tasks involved in the computational treatment of MWEs have been structured by the organisers of the 2009 MWE workshop (Anastasiou et al. 2009) as follows.

- **Identification (or acquisition).** Given a text as input, try to locate the interesting multiword units in it.
- **Interpretation.** Given a multiword unit out of context, try to discover its internal structure both in terms of syntactic and semantic relations.
- **Disambiguation.** Given a multiword unit in its context, try to classify it with respect to a closed set of categories. Typically, one tries to distinguish literal from idiomatic uses, but other disambiguation tasks are possible, for instance, distinguishing general-purpose from specialised uses and performing multiword sense

disambiguation.

– **Application.** Given a lexicon of MWEs, try to integrate it in another application such as parsing, information retrieval or MT.

Interpretation and disambiguation are similar as both can be modelled as classification tasks. However, they are distinct as the former concerns MWE types whereas the latter deals with MWE tokens as they occur in text. In addition to the four topics above, we consider an additional task which lies between disambiguation and application, representation:

– **Representation.** Given a lexicon containing MWEs (automatically or manually acquired), try to optimise their representation in a given formalism considering their properties and the target application.

As pointed out in the call for papers of the MWE 2009 workshop: [20]

> The above topics largely overlap. For example, identification can require disambiguating between literal and idiomatic uses since MWEs are typically required to be non-compositional by definition. Similarly, interpreting three-word noun compounds like *morning flight ticket* and *plastic water bottle* requires disambiguation between a left and a right syntactic structure, while interpreting two-word compounds like *English teacher* requires disambiguating between (a) 'teacher who teaches English' and (b) 'teacher coming from England (who could teach any subject, e.g., math)'.

As a large part of the research developed and presented in this thesis focuses on the first task, the present section is entirely dedicated to MWE acquisition. We start with a summary of related work on monolingual acquisition in Section 3.2.1, and on multilingual acquisition in Section 3.2.2. Then, we present a more practical description of tools that perform automatic acquisition, distinguishing between those freely available developed by academic researchers and those which were developed and commercialised in a proprietary context.

3.2.1 Methods for monolingual MWE acquisition

In this section, we discuss the more relevant papers, and Appendix C gives a more comprehensive listing of monolingual acquisition methods per language. The references discussed here and in Appendix C are complemented by the work that has been developed for other MWE tasks (Section 3.3).

One of the goals of monolingual acquisition techniques is to help and speed up the creation of lexical resources such as printed or machine-readable dictionaries and thesauri containing multiword entries. We distinguish two types of acquisition:

– In *MWE identification*, the input is a text and the expected output is a mark-up indicating the places where MWEs occur. This may include the use of an existing dictionary or the discovery of new MWEs. What makes MWE identification more difficult than simple regular-expression matching is non-adjacency, morphological inflection and ambiguity of some MWEs that can be used both as compositional and idiomatic sequences (e.g., *look up* as consult a dictionary or as staring towards a higher position). In MWE identification, a token-based evaluation is required, taking into account the context in which the expression occurs.

– In *MWE extraction*, the input is a text and the expected output is a list of MWE candidates found in the text. The evaluation can be done on a type basis, as if

20. `http://multiword.sourceforge.net/mwe2009`

each expression was an entry of a lexicon, independently of the input corpus. In extraction, it is usual to consider two separate steps: (a) *candidate extraction* and (b) *candidate filtering and/or ranking*. We consider that an *MWE candidate* is a sequence of words which has some of the characteristics described in Section 2.3 as measured by some objective measure, but that was not yet validated by a manual or automatic evaluation process.

Candidate extraction methods are based on some kind of pattern matching, where the patterns range from simple *n*-grams to structured sequences of part-of-speech tags and syntactic relations. The level of linguistic information employed in candidate extraction depends on various factors such as the language, the type and syntactic variability of the target MWEs and the available analysis tools.

The use of surface forms alone is rare, as generally at least minimal patterns based on stopwords or POS are employed (Gurrutxaga and Alegria 2011). However, there might be cases where flat *n*-gram extraction is required, for instance, when the target MWEs are generic keyphrases for document description and indexation (Silva and Lopes 2010). The sliding window method consists of considering as MWE candidates pairs that co-occur in a window of at most *w* words, thus retrieving discontiguous *n*-grams (Smadja 1993). The extraction of candidates using sliding windows can pose a challenge in terms of computational performance. Indeed, optimised data structures and algorithms must be used because the number of possible combinations, even for relatively small sizes of *n*, explodes with the size of the corpus (Gil and Dias 2003).

Part of speech sequences are one of the major approaches in candidate extraction because (i) many languages have available push-button POS taggers and (ii) this approach provides good results when the target constructions are relatively rigid in terms of word order, like fixed phrases and nominal MWEs. POS sequences have been used originally in multiword terminology acquisition (Justeson and Katz 1995, Daille 2003), but have also been applied to the extraction of other MWE types, specially noun compounds (Vincze et al. 2011). Even when dealing with more variable constructions such as verbal expressions, POS tag patterns can be used in the absence of syntactic information (Baldwin 2005b, Duran et al. 2011). POS patterns can be defined based on various criteria, from linguistic intuition and expert knowledge (Bonin et al. 2010b) to systematic empirical observation of a sample (Duran et al. 2011). Sequences of POS can also be automatically learnt from the annotated corpus, using the same methodology as for words, that is, by maximizing some AM on the extracted POS *n*-grams (Dias 2003).

When a parser is available, patterns based on syntactical relations can be used for candidate extraction. For example, one may retrieve all candidates that are formed by a noun which is the direct object of a verb (*take/V* $\leftarrow_{\text{DOBJ}}$ *time/N*). According to the accuracy of the parser, simple syntactic patterns can be much more precise than POS sequences, specially in the extraction of non-fixed MWEs like "true" collocations (Seretan and Wehrli 2009, Seretan 2008). Tree substitution grammars can also be used in order to learn syntactic MWE models from annotated corpora, as it is performed for the French version of the Stanford parser (Green et al. 2011). Regardless of the syntactic information and labels, structural regularities in parsing trees can also be used to retrieve MWE candidates using a minimal description length algorithm (Martens and Vandeghinste 2010)

In addition to analysed corpora, other monolingual and multilingual resources can be used for MWE acquisition. For instance, by comparing the titles of Wikipedia pages using cross-language links, it is possible to detect multiword titles whose translation in one of the other languages is a single word (Attia et al. 2010). Another way to use the

web as a source of information for MWE acquisition is to generate candidates according to generic combination rules and further validate them using web search engine hit counts (Villavicencio et al. 2005b). This is explored in our experiments in Section 6.2.3.2. The current trend is the integration of several complementary information sources (including linguistic analysis, statistics, the web) in order to maximise the recall of the extraction (de Medeiros Caseli et al. 2009, Attia et al. 2010).

More sophisticated candidate extraction methods, not based on pattern matching, have also been proposed. The LocalMaxs algorithm, for instance, performs extraction based on the maximisation of an AM applied to adjacent word pairs. Thus, it naturally handles nested expressions, extracting maximal sequences that recursively include adjacent words while the overall AM score increases (Silva and Lopes 1999). Similarly, a tightness measure is used in a Chinese IR system for the automatic identification, concatenation and optimised querying of strongly associated word sequences (Xu et al. 2010). A string matching algorithm inspired by computational biology has been proposed to extract sequences that occur recurrently throughout the corpus. Sentences are viewed as DNA sequences and a dynamic programming algorithm matches corresponding parts for each sentence pair in the corpus, taking into account gaps that represent variable parts of the expression (Duan et al. 2006). These techniques generally do not distinguish candidate extraction from filtering, performing both simultaneously.

As for candidate filtering, some straightforward procedures are the use of stopword lists and of count thresholds to remove candidates for which statistical information is insufficient. Lexical association measures like those described in Section 3.1.4 are also widely employed to rank the candidates and keep only those whose association score is above a certain threshold (Evert and Krenn 2005, Pecina 2005). When several AMs are available and must be combined, possibly considering additional information coming from auxiliary resources, one can use machine learning. Thus, it is necessary to annotate part of the data or to obtain an annotated dataset. Then, a supervised learning method can be used to build a classifier modelling the optimal weights of all the AMs and extra features (Ramisch et al. 2008b, Pecina 2008b).

There is a strong predominance of methods based on 2-grams (or more generally on word pairs) in current techniques for monolingual MWE acquisition. This is justified because (i) the majority of the interesting and challenging MWEs are formed by two words and (ii) "experiments with longer expressions would require processing of much larger amount of data and [there is a] limited scalability of some methods to [handle] high order *n*-grams" (Pecina 2005). While this seems like a reasonable justification to keep the methodology simple, it does not correspond to the reality of NLP applications, where the many MWEs longer than 2 words also require proper treatment.

Monolingual methods have been developed in several languages and are sometimes language independent. The advantage of language-independent methods is that they do not depend on the availability of a specific resource (POS tagger, parser) and can thus be applied to virtually any language, including poorly resourced ones. On the other hands, the use of linguistic information generally improves the precision and the coverage of the acquisition. Finding an adequate trade-off between language independence and quality when designing a method for monolingual acquisition is a challenging problem. However, as MWEs seem to be a universal phenomenon, being present in all human languages, it is important to build methods and evaluate them in multilingual contexts (Seretan and Wehrli 2006).

3.2.2 Methods for bi- and multilingual MWE acquisition

Even though many of the methods described in the previous section can be applied to arbitrary corpora, independently of the language, they are still considered as monolingual methods because the result is a list of MWEs with no cross-language correspondences. The extraction of bilingual MWEs is a task in which the resulting list of expressions is bilingual, that is, if a candidate is returned in one language, it contains translation links which relate it to its correspondent candidate in the other language. Hence, bi- and multilingual MWE acquisition is different from language-independent MWE acquisition. Existing techniques for bilingual MWE acquisition are frequently based on parallel corpora. To the best of our knowledge, there is no account in the MWE literature of truly multilingual techniques for MWE extraction, dealing simultaneously with more than two languages.

Automatic word alignments can provide lists of MWE candidates by themselves, as described in de Medeiros Caseli et al. (2010). They aligned a Portuguese–English corpus in both directions using GIZA++, and then joined the alignments using the grow-diag-final heuristic. Word sequences of two or more words on the source side aligned to sequences of one or more words on the target side were filtered using several stopword patterns and the resulting candidates were considered as MWEs. The comparison with a simple monolingual *n*-gram method showed that alignment-based extraction is much more precise, but has very limited recall.

Bai et al. (2009) present an algorithm capable of mining translations for a given MWE in a parallel aligned corpus. Then, the different translations are ranked according to standard association measures in order to chose the appropriate one. They integrated this extraction method into the empirical MT system Moses for the English–Chinese language pair, obtaining improved translations when compared to baseline translations.

The automatic discovery of non-compositional compounds from parallel data has been explored by Melamed (1997). Considering a statistical translation model, he introduced a feature based on mutual information and proposed an iterative algorithm that retrieves an increasing number of compounds. These can in turn be used to improve the quality of the statistical translation system itself.

Conversely, it has been shown that MWEs can improve the quality of automatic word alignment. The English-Hindi language pair presents large word order variation, and it has been shown that MWE-based features that model compositionality can help reducing alignment error rate (Venkatapathy and Joshi 2006). When compared with baseline GIZA++, a system enriched with MWE features obtains significantly lower error rates, from 68.92% to 50.45%.

The acquisition of bilingual verbal expressions requires not only the availability of parallel corpora, but also of syntactic analysis of both languages. Zarrieß and Kuhn (2009) used syntactically analysed corpora and GIZA++ alignments to extract verb-object pairs from a German–English parallel corpus. They considered a candidate as a true MWE if (i) a verb on the source side was aligned to a verb on the target side, (ii) the noun heading the object of the verb on the source side was tagged as a noun on the target side and (iii) there was a syntactic object relation on the target side between the target verb and the target noun. Their method retrieves 82.1% of correct translations, and almost 60% of translations which can be considered as MWEs.

Instead of relying on large parallel word-aligned corpora, which are not always available for a given language pair, it is possible to use comparable corpora as a source for acquisition. Daille et al. (2004) performed multiword term extraction independently in

French and in English using comparable corpora in the environmental domain. Then, using the distances between the context vectors of the acquired terms, they obtained cross-lingual equivalences that were evaluated against a bilingual terminological dictionary. The dictionary reference translation occurred among the top-20 retrieved translations in 47% to 89% of the translations, depending on the translation relation type (single word vs multiword).

The acquisition of bilingual MWEs has been explored more often in the context of machine translation. In Section 3.3.4.4, we provide an overview of attempts to integrate MWEs into different MT applications. This is further developed in the experiments described in Section 7.3.

3.2.3 Existing tools

The maturity of a research field depends not only on theoretical models and experimental results, but also on concrete tools and available software on the basis of which it is possible to reproduce results, build extensions and perform systematic evaluations. Thus, tools for the automatic acquisition of MWEs are very important for the evolution of this research field. Here, we distinguish two types of tools: those which are freely available for the community (Section 3.2.3.1) and those that are either commercialised or available in restricted contexts (Section 3.2.3.2).

3.2.3.1 Freely available tools

To date, the existing research tools follow the main trends in the area, using linguistic analysis and statistical information as clues for finding MWEs in texts. Here, we present a non-exhaustive list of freely available tools that can be used for mostly monolingual MWE acquisition.

1. **LocalMaxs**: `http://hlt.di.fct.unl.pt/luis/multiwords/`
   The "Multiwords" scripts are the reference implementation [21] of the LocalMaxs algorithm. It extracts MWEs by generating all possible *n*-grams from a sentence and then further filtering them based on the local maxima of a customisable AM's distribution (Silva and Lopes 1999). On the one hand this approach is based purely on word counts and is completely language independent. On the other hand, it is not possible to directly integrate linguistic information in order to target a specific type of construction or to remove noisy ungrammatical candidates. [22] The tool includes a strict version, which prioritises high precision, and a relaxed version, which focuses on high recall. A separate tool is provided to deal with big corpora. A variation of the original algorithm, SENTA, has been proposed to deal with non-contiguous expressions (da Silva et al. 1999). However, it is computationally costly because it is based on the calculation of all possible *n*-grams in a sentence, which explodes when going from contiguous to non-contiguous *n*-grams. Furthermore, there is no freely available implementation.
2. **Text::NSP**: `http://search.cpan.org/dist/Text-NSP`
   The *N*-gram Statistics Package (NSP) is a standard tool for the statistical analysis of *n*-grams in text files developed and maintained since 2003 (Pedersen et al. 2011, Banerjee and Pedersen 2003). It provides Perl scripts for counting *n*-grams in a

21. Recommended by the author of the algorithm in personal communication.
22. Although this can be simulated by concatenating words and POS tags together in order to form a token.

text file and calculating AMs, where an n-gram is either a sequence of n contiguous words or n words occurring in a window of $w \geq n$ words in a sentence. While most of the measures are only applicable to 2-grams, some of them are also extended to 3-grams and 4-grams, notably the log-likelihood measure. The set of available AMs includes robust and theoretically sound measures such as Fischer's exact test. The input to the NSP tool is a corpus and a parameter value fixing the size of the target n-grams. The output is a list of types extracted from the corpus along with the counts, which can further be used to calculate the AMs. Although there is no direct support to linguistic information such as POS, it is possible to simulate them to some extent using the same workaround as for LocalMaxs. [22] The tool allows complex expressions in order to express what counts should be calculated in terms of the sub-n-grams contained in a given n-gram.

3. **UCS**: `http://www.collocations.de/software.html`

   The `UCS` toolkit provides a large set of sophisticated AMs, in addition to other mathematical procedures like dispersion test, frequency distribution models and evaluation methods. It was developed in Perl and uses the R statistics package. UCS focuses on high accuracy calculations for 2-gram AMs, but, unlike the other approaches, it does not properly perform MWE acquisition. Instead of a corpus, it receives a list of candidates and their respective counts, relying on external tools for corpus preprocessing and candidate extraction. Then, it calculates the measures and ranks the candidates. Therefore, the question about contiguous n-grams or support of linguistic filters is not relevant for UCS.

4. **jMWE**: `projects.csail.mit.edu/jmwe`

   The jMWE tool (Kulkarni and Finlayson 2011) is aimed at dictionary-based in-context MWE token *identification* in running text, which makes it quite different from *extraction* tools. It is available in the form of a Java library, and expects a corpus as input, possibly annotated with lemmas and parts of speech. In addition, it requires an initial dictionary of valid known MWEs. The system then looks for instances (occurrences) in the corpus of the MWEs included in its internal dictionary. It does not perform any automatic discovery of new expressions, thus the quality of the output heavily depends on the availability of MWE dictionaries. While jMWE is not language independent, it can be configured and straightforwardly adapted to other languages for which a suitable dictionary is available. The system allows quite powerful instance search, similar to multilevel regular expressions. It is possible to deal with non-contiguous expressions and to apply successive filters on the output. jMWE also provides heuristics for disambiguating nested compounds. On the other hand, it is not possible to express constraints based on syntax, nor to apply AMs in order to remove words that co-occur by chance.

5. **Varro**: `http://sourceforge.net/projects/varro/`

   This tool is not specifically aimed at MWE acquisition, but rather at finding regularities in treebanks (Martens 2010). It implements an optimised version of the *Apriori* algorithm with many adaptations that allow for the efficient and compact representation of tree structures. Statistical scores based on description length gain have been proposed to rank regular subtrees returned by the tool, thus helping in the acquisition of MWEs (Martens and Vandeghinste 2010). In contrast with the preceding tools, the Varro toolkit is not based on word sequences but it requires syntactically analysed corpora as input. It is thus well suited for the extraction

of flexible expressions such as idioms, formulaic phrases, “true” collocations and verbal expressions.

There are also numerous freely available web services and downloadable tools for automatic term extraction. These tools are generally language dependent, having versions for major European languages like English, Spanish, French and Italian. Although multiword terms are included in our definition of MWE, these tools are not appropriate for general-purpose extraction of expressions in everyday language. Examples of such tools are TermoStat [23], AntConc [24] and TerMine. [25] The Wikipedia page on terminology extraction [26] lists many other freely available tools.

The methodological framework introduced in the present work has also been implemented in a freely available tool, the MWE toolkit. [27] This tool is described in detail in Chapter 5.

3.2.3.2 Proprietary commercial tools

There are numerous commercialised systems for automatic terminology extraction from specialised texts. As a great part of terminology is multiword, this kind of software performs MWE acquisition at some point. At Xerox and, in particular, at their research centres, such techniques and tools for term extraction have been developed for a long time. Déjean et al. (2002), for example, describe a method that uses morphosyntactic patterns for monolingual term recognition. Afterwards, they perform automatic alignment and extract English–German terminology, reaching an F-measure of around 80%. This kind of technique has certainly been integrated into their Xerox Terminology Suite (XTS). This software is not commercialised any more, since it has been acquired by the text mining company Temis. [28] Nowadays, it has become part of the Luxid® information extraction package. [29]

Another large company which developed a tool for terminology extraction is Yahoo!. Their term extraction service is freely available for research and personal purposes, limited to 5,000 queries per day per IP address [30]. However, this service is limited to short English texts and is probably based on term dictionaries and gazetteers.

The Fips parser, developed at the University of Geneva, has been used for collocation extraction in several languages (Seretan and Wehrli 2009). Even though it is academic research, the collocation extraction tool FipsCo, based on Fips, is not freely available. The tool is able to extract collocations from monolingual corpora in English, French, Spanish and Italian, and there is a version for Greek (Michou and Seretan 2009). The tool has been used in MT experiments, suggesting that it is able to extract bilingual collocations from word-aligned parallel corpora. Although the system itself is not free, its visualisation tool, FipsCoView [31], is freely available as a web interface (Seretan and Wehrli 2011).

Translation memory software may use MWEs as basic segments to retrieve. Indeed, MWEs are somehow in-between sentences and words. On the one hand, the retrieval of simple words in a hypothetical translation memory would be of little usefulness. The

23. http://olst.ling.umontreal.ca/~drouinp/termostat_web/
24. http://www.antlab.sci.waseda.ac.jp/software.html
25. http://www.nactem.ac.uk/software/termine/
26. http://en.wikipedia.org/wiki/Terminology_extraction
27. http://mwetoolkit.sourceforge.net
28. http://www.temis.com/
29. http://www.temis.com/index.php?id=201&selt=1
30. http://developer.yahoo.com/search/content/V1/termExtraction.html
31. http://129.194.38.128:81/FipsCoView

number of possible translations for a word out of its context is potentially large and additional information is required to chose among the options. Therefore, it would lack of precision. On the other hand, the retrieval of whole sentences would be highly precise, but an extremely large translation memory would be required in order to obtain reasonable recall. If the memory of previously translated segments is small, only from time to time (and with some luck) a sentence will be retrieved. Many sentences containing part of the translation would be useful, but will be ignored by a sentence-based exact match system.

One example of system performing bilingual MWE extraction is Similis [32], previously commercialised by Lingua et Machina and now freely available. According to the official website, "Similis [...] includes a linguistic analysis engine, uses chunk technology to break down segments into intelligent terminological groups (chunks), and automatically generates specific glossaries." The technique implemented in the system is an evolution of the one described in Planas and Furuse (2000). In this article, the authors describe a clever technique for retrieving similar segments in the source language and their correspondences in the target language. Their approach applies a dynamic programming algorithm on a multi-layered structure where sentences are represented as a sequence of surface forms, lemmas and parts of speech. The combination of the matchings in these three layers allows for a good balance between precision and recall for the retrieval of bilingual segments.

3.3 Other MWE tasks

Given that MWE acquisition is one of the main axes of the present thesis, the whole Section 3.2 is dedicated to a detailed review of the state of the art. Here, we overview the state of the art in the other tasks involved in MWE treatment, according to the classification of MWE tasks, namely interpretation (Section 3.3.1), disambiguation (Section 3.3.2), representation (Section 3.3.3) and application (Section 3.3.4).

3.3.1 Interpretation

The interpretation and disambiguation of several types of MWEs are the focus of a large body of literature, even if they received considerably less attention than acquisition. Both can be modelled as classification tasks, so that machine learning algorithms are often employed. Therefore, it is possible to distinguish supervised from unsupervised approaches. In the former, a large effort is usually dedicated to the annotation of a data set that is subsequently used to build classifiers. In the latter, the class attribution is made based on thresholds or rules directly applied to the data features. As for general machine learning problems, supervised methods largely outperform unsupervised methods. However, unsupervised methods may sometimes perform as well as supervised methods when they are applied on very large corpora like, for instance, web-based corpora (Keller and Lapata 2003).

MWE interpretation requires expressions whose meaning does not depend on their occurrence contexts, like compound nouns and and some specific types of phrasal verbs and support verb constructions. However, it is not suitable to interpret ambiguous expressions such as phrasal verbs (*look up a word* vs *look up to the sky*) and idioms (*my grandfather kicked the bucket* vs *the cleaning lady accidentally kicked the bucket*). These are explored in MWE disambiguation tasks. Noun compounds (*traffic light*, *nuclear transcription fac-*

32. http://similis.org/

tor), on the other hand, are rarely ambiguous and their interpretation has been an active research area. We distinguish two types of noun compound interpretation: syntactic and semantic.

The *syntactic interpretation* has been explored by Nicholson and Baldwin (2006), who distinguish three syntactic relations in noun–noun compounds: subject (*product replacement*), direct object (*stress avoidance*) and prepositional object (*side show → show on the side*). For compounds in which the second noun is a nominalisation [33], they used the inflections of the corresponding verb to generate paraphrases that were looked up in Google. The paraphrases and additional features were input in a nearest-neighbour classifier, but the results failed to improve over the state of the art.

Three-word or longer noun compounds like *liver cell line* and *liver cell antibody* require syntactic interpretation of the constituent hierarchy. That is, one needs to distinguish left bracketing like in *(liver cell) antibody* from right bracketing like in *liver (cell line)*. Therefore, Nakov and Hearst (2005) compare two models, based on adjacency and on dependency. They use a set of heuristics to generate surface-level paraphrases and then use search engine counts to estimate model probabilities. They obtain sizeable improvements over state of the art on a set of biomedical compounds.

One of the most challenging interpretation problems is the *semantic interpretation* of the relations involved in noun compounds. The goal is to assign to each noun compound one (or several) tags that describe the semantic relation between the two nouns. Nakov and Hearst (2008) try to solve this task using a methodology similar to the one they employed for syntactic interpretation. First, they generate a large number of paraphrases involving verbs related to the semantic classes (e.g., *causes, implies, generates* for relation *CAUSE*) and the relative *that*. Then, they retrieve web counts for the paraphrases and assign the classes with maximal probability according to the corresponding paraphrases. Their method is completely unsupervised. The resource developed in their work, containing noun compounds and corresponding features, is freely available on the MWE community website (Nakov 2008b). More recently, Kim and Nakov (2011) revisited the problem, this time using a combination of data bootstrapping and web counts. The main difference is that, this time, they generated paraphrases not based on surface forms but on parse trees, thus obtaining more accurate results.

Paraphrases can be used not only as means but also as ends. That is, they may be the *actual* representation of semantic classes instead of a set of (somehow arbitrary) abstract tags. The representation of semantic classes for noun–noun relations is discussed in depth by Girju et al. (2005), who compare Lauer's eight prepositional tags with a proposed classification using 35 abstract tags. Moreover, they annotate a corpus sample using both schemes and investigate the correspondences between them. In addition, paraphrases can be used, for instance, in order to artificially generate new data for training empirical MT systems (Nakov 2008a).

Lapata (2002) focuses on the interpretation of noun compounds involving nominalisations. She reformulates noun compound interpretation as a disambiguation task, recreating missing evidence from corpus. She extracts the counts of the nouns and of the related verb from the BNC, and then uses them as features in a supervised machine learning tool that automatically learns association rules. She also discusses and evaluates several smoothing techniques [34] that help obtaining more realistic counts. Keller and Lapata (2003) used this task as one of their case studies in order to investigate the use of web

33. A noun derived from a verb, like *replacement* is a nominalisation of the verb *replace*.

34. Smoothing techniques are rarely employed in MWE tasks, as opposed to other NLP fields like MT.

counts in NLP disambiguation tasks.

Latent semantic analysis has also been employed for the semantic classification of noun–noun compounds (Baldwin et al. 2003). In order to distinguish compositional from idiomatic constructions, the authors compare the context vectors of the compound with the context vectors of the individual nouns composing it. This approach can be generalised and has also been applied and evaluated on other types of MWEs.

A comprehensive and detailed revision of the semantic interpretation of noun compounds can be found in Nakov's Ph.D. thesis (Nakov 2007). For an up-to-date state of the art, please refer to the proceedings of SemEval 2010, which features a shared task on this topic (Butnariu et al. 2010), and to the corresponding extended paper version (Girju et al. 2009).

Besides noun compounds, other MWE types also require interpretation. English phrasal verbs are ambiguous and can be used both idiomatically (*look up a word*) and literally (*look up to the sky*). However, if we consider only the most usual sense, it is possible to perform type-based interpretation. Cook and Stevenson (2006) use support vector machines to classify the meaning of the particle *up* in English phrasal verbs. According to the verb, it can mean have a sense of vertical, completion, goal or reflexive. These are simplified using a 2-way and a 3-way classification. The features used are standard syntactic slots of the verb, particle characteristics such as distance from the verb, and word co-occurrences.

Considering a larger range of constructions, Bannard (2005) investigates the extent to which the components of a phrasal verb contribute their meanings to the interpretation of the whole. He models compositionality through an entailment task, for instance, *split up* $\Longrightarrow$ *split*? In a comparison between `pmi`, `t-score` and a newly proposed measure based on context cosine similarity, the latter correlates better with human judgements.

A similar work is that of McCarthy et al. (2003). They propose several measures involving an automatically acquired thesaurus in order to estimate the idiomaticity of phrasal verbs. Their annotated data set uses a numeric scale from 0 (totally opaque) to 10 (fully compositional). They show that the best association measure, mutual information, is less correlated to human judgements than a proposed measure which calculates the number of neighbours with the same particle as the phrasal verb minus the equivalent number of simple neighbours.

Venkatapathy and Joshi (2006) explore the type compositionality of verb–noun pairs. They describe the creation of an annotated data set with compositionality judgements ranging from 1 to 6. Then, they present seven distinct features to estimate compositionality which are further combined using a support vector machine. They evaluate the features separately and show that the Spearman correlation between the classifier results and human judgements is around 0.448, which is better than all individual features.

Using a variation of the same data, McCarthy et al. (2007) investigate the use of selectional preferences in this task. They propose three different algorithms to obtain this information from parsed corpora: two based on Wordnet and one based on an automatically constructed thesaurus. They show that the best performance is obtained by combining selectional preferences and a subset of Venkatapathy and Joshi's features through a support vector machine.

3.3.2 Disambiguation

Recall that the disambiguation of MWEs is analogous to their interpretation, except that they are considered together with the context in which they appear (sentences).

Nicholson and Baldwin (2008) present a data set for noun–noun compound disambiguation where a large set of sentences has been manually annotated with syntactic and semantic information about the compounds contained in it. Girju et al. (2005) investigate methods for their disambiguation. They perform a separate analysis of two- and three-noun compounds, annotating their semantics according to two tagging schemes in a training set of around 3K sentences. In addition to a detailed analysis of the coverage and correspondences between the tagging schemes, they apply several supervised learning techniques. Like for the syntactic disambiguation of three-word compounds, they also employ classifiers. They achieves an accuracy of 83.10% by using as features (a) the top three WordNet synsets for each noun, (b) derivationally related forms and (c) a flag telling whether the noun is a nominalisation.

Whereas, for MWE interpretation, the majority of works concerns noun compounds, when it comes to disambiguation a large number of MWE types has been studied. However, English still predominates. One of the rare works on a language different from English concerns the interpretation of German preposition–noun–verb triples (Fritzinger et al. 2010). Constructions like *in Gang kommen* have both a literal interpretation as *to reach the hallway (in den Gang kommen)*, but also idiomatic interpretations as *to be set in motion (in Gang kommen)* and *to get organised (in die Gänge kommen)*. They manually analysed a large set of such constructions retrieved by a parser, classifying them as either literal, compositional or unknown. [35] Then, they investigated the correlation between these classes and morphosyntactic characteristics such as determiners, plural and passivisation. They did not employ machine learning in order to recognise recurrent patterns in the data.

Light/support verbs in Japanese have also been studied in the past. They include sequences like *donari-ageru* (*shout*) and *odosi-ageru* (*threaten*), that is, formed by two lexical units where the verb is usually highly polysemous like *ageru* (*raise*). Uchiyama et al. (2005) propose two disambiguation methods: a statistical approach using a sense inventory, context and a support vector machine; and a rule-based method where the rules were manually defined based on syntax and on the semantics of the first verb. The rule-based method (94.6%) outperforms the statistical method (82.6%) in terms of accuracy, but the latter obtains a surprisingly high performance given its simplicity.

The interpretation of expressions of the type verb–noun has also been explored in English. Cook et al. (2007) explore the idiomaticity of verb–noun pairs, where the noun is the direct object of the verb and may have an idiomatic (*make a face*) or literal (*make a cake*) interpretation. Their basic hypothesis is that idiomatic uses are syntactically more rigid. Thus, they describe a fully unsupervised approach which considers syntactic and context information in order to calculate the similarity with the canonical form of the idiom. In their evaluation, they report results comparable to a supervised approach. The data set used in their experiments is freely available (Cook et al. 2008).

Fazly and Stevenson (2007) propose a more fine-grained classification for light verb–noun constructions. They use a supervised learning strategy based on decision trees in order to perform a 4-way semantic disambiguation. In their scheme, a light verb may be used with its literal meaning (*make a cake*), with its abstract meaning (*make a living*), in light-verb constructions (*make a decision*) or idiomatically (*make a face*). These classes are a mix of syntactic and semantic characteristics and could arguably be improved using more systematic criteria. Even though they perform type-based annotation of their data

35. The context unit used for annotation was the sentence. However, due to anaphora, sometimes it was impossible to know the intended meaning without looking at neighbour sentences.

sets, this work can be considered as disambiguation because the noun is the context used to disambiguate the semantics of a closed set of polysemous light verbs. Considering a random baseline with 25% accuracy, they obtain an overall accuracy of 58.3%. F-measure varies according to the classe: abstract constructions are harder to classify (46%) than light verb constructions (68%).

3.3.3 Representation

The lexical representation of MWEs was one of the main goals of the MWE project at Stanford, and has for a long time haunted lexicographers in the compilation of lexical resources. Most NLP applications contain at least a small amount of MWE entries, specially closed-class expressions. The Stanford parser, for instance, contains a list of several dozens of 2-word and 3-word conjunctions. However, when it comes to open-class expressions, this coverage is too limited and ways to efficiently represent MWEs in computational lexicons are required. Sag et al. (2002) proposed two approaches: words-with-spaces and compositional. However, between these extremes of the compositionality spectrum, there are some other possibilities, sometimes explored in related work. Laporte and Voyatzi (2008), for instance, describe a dictionary containing 6,800 French adverbial expressions like *de nos jours*. A set of 15 flat sequences of parts of speech is used to describe the morphosyntactic pattern of each entry using the lexicon–grammar format.

Graliński et al. (2010) present a qualitative and quantitative comparison between two structured representations for Polish MWEs: Multiflex and POLENG. While the former is designed to be generic and language independent, the latter has a more implicit structure aimed at specific applications. The authors focus on nominal compounds and analyse the power of each formalism to incorporate morphological inflection rules such as case, gender and number agreement. They also measure the time taken by one expert and two novice lexicographers to encode new MWEs. Multiflex does not allow the description of non-contiguous units nor units containing slots and it takes much longer for lexicographers to learn and use it. POLENG offers a complementary approach, allowing a faster description of MWEs including non-contiguous ones.

A less "intuitive" and more corpus-based representation has been proposed for the representation of entries in the Dutch electronic lexicon of MWEs (Grégoire 2007; 2010). She uses an equivalence class method that groups similar expressions according to their syntactic characteristics. In addition to numbers of occurrences and examples, each entry contains a link to a pattern that describes the syntactic behaviour of the expression. This description is quite practical, as the lexicon is aimed for NLP systems such as the Alpino parser and Rosetta MT system.

Izumi et al. (2010) suggest a rule-based method to normalise Japanese functional expressions in order to optimise their representation. They consider two separate problems: the insertion of omitted parts and the removal of satellite parts that do not contribute much to the meaning of the sentence. In a comparison with manually generated paraphrases, they obtain a precision of 77%. If such normalised representation are adopted in the lexicon, the same paraphrasing rules can be applied to running text in order to align it with expressions contained in the lexicon.

The use of tree-rewriting grammars for describing MWEs is proposed by Schuler and Joshi (2011). They provide intuitive arguments and formal proof that this formalism is adequate to represent non-fixed expressions such as *raise X to the Y^{th} power*. The generalisation of their approach to other types of expressions, however, remains to be demonstrated.

Finally, concerning the hierarchical structure among MWEs, SanJuan et al. (2005) explore three strategies (lexical inclusion, Wordnet similarity and clustering) to organise a set of multiword terms manually extracted from the Genia corpus. This kind of representation can be very useful to include extracted expressions in more sophisticated concept nets and ontologies.

When it comes to bilingual and multilingual dictionaries, the problem becomes more complex since it is necessary to represent not only the internal structure of the entries but also cross-language links at global and local levels. To the best of our knowledge, there is little research concerning this problem. Section 3.3.4.4 contains a discussion on the representation of MWEs in MT systems.

In short, due to the modest amount of research in this area and to the complexity of the problem, a model for the efficient lexical representation of MWEs in general remains an open problem.

3.3.4 Applications

A list of potential NLP applications where MWEs are relevant was introduced in Section 1.1.2. Here, we provide a summary of these target applications for which concrete results have been obtained. Many results presented here concern pilot studies and techniques as simple as joining contiguous MWE components with an underscore character as a preprocessing step. From all the MWE tasks discussed in this section, application is by far the one with the least amount of published results.

3.3.4.1 Syntactic analysis

A small set of fixed MWEs like conjunctions are represented in most existing parsers, chunkers and POS taggers. However, the further insertion of additional multiword entries can improve the coverage of the analysis, as more complex MWEs like noun compounds and verbal expressions are valuable information for syntactic disambiguation.

Concerning POS tagging, Constant and Sigogne (2011) present an evaluation on French. They assign special tags to words corresponding to the beginning and to the ending of multiword units. Using a model based on conditional random fields, they learn the MWE-aware tagger from a corpus in which the training data was automatically annotated with entries coming from several lexica containing compounds and proper nouns. This technique obtains 97.7% accuracy, improving considerably over standard taggers like TreeTagger and TnT.

Korkontzelos and Manandhar (2010) obtain impressing improvements by enriching a baseline shallow parser with MWEs. Their method simply consists of joining contiguous nominal expressions with an underscore prior to parsing. This makes the system treat them as unknown tokens and assign them a parse based on the context. They analyse a set of 118 2-word MWEs from WordNet, classifying them by POS sequences and by compositionality. They conclude that, in all cases, the accuracy of the parses was improved, specially for non-compositional adjective–noun pairs, for which the substantial improvements ranged from 15.32% to 19.67%.

As for deep parsing, Zhang and Kordoni (2006) extended the lexicon of an English HPSG parser with 373 MWE entries represented as words-with-spaces. They obtained a significant increase in the coverage of the grammar, from 4.3% to 18.7%. Using a compositional representation, Villavicencio et al. (2007) added 21 new MWEs to the same parser, obtaining an increase in the grammar coverage from 7.1% to 22.7%, without degrading accuracy. However, MWEs do not always improve the performance of the parser,

as shown by Hogan et al. (2011). They try to include a set of named entities in their parsing system, replacing them by placeholders. However, they did not obtain significant improvements over the baseline, even when tuning count thresholds.

As far as we know, the English and Italian parser Fips is one of the few systems dealing with variable MWEs (Wehrli et al. 2010). Its approach is more sophisticated than words-with-spaces, as it dynamically identifies expressions at the same time as it constructs the parse tree. This technique performs better than post-processing the trees after they are produced. The authors demonstrate that MWEs are not a "pain in the neck" but actually a valuable information to reduce syntactic ambiguity. A similar strategy is employed in the translation system ETAP-3 (Apresian et al. 2003).

3.3.4.2 Word sense disambiguation

Given an occurrence of a polysemous word, word sense disambiguation consists of picking up a single sense among those listed in an inventory. For example, the verb *fire* can mean *make somebody lose his/her job*, *pull the trigger of a gun*, or *make something burn*. The sentence in which the verb occurs will determine which of these senses is intended. Although context information is used, MWEs are generally ignored in WSD tasks. As a consequence, not only the correct sense will be ignored but also wrong senses will be inferred for the individual words. For example, in Wordnet, none of the senses of *voice* and of *mail* indicates that *voice mail* means *system that records messages on a telephone when nobody answers*.

As exemplified by Finlayson and Kulkarni (2011), while the word *voice* has eleven senses and the word *mail* has five, the expression *voice mail* only has one. They show that, in Wordnet 1.6, the average polysemy of MWEs is of 1.07 synsets, versus 1.53 for simple words. To the best of our knowledge, their work is the first to report a considerable improvement on word sense disambiguation performance due to the detection of MWEs. Despite its simplicity, their method reaches an improvement of 5 F-measure points given lower and upper bounds of 3.3 and 6.1.

3.3.4.3 Information retrieval (IR)

Let us consider a simplified IR system, modelling documents as bags of words and not keeping track of co-occurrences. For instance, if a document contains the term *rock star*, it will probably be retrieved as an answer to queries on geology (*rock*) and astronomy (*star*). If this MWE was represented as a unit in the index of the system, the precision of the retrieved documents could increase. Most current IR systems allow more sophisticated queries to be expressed through quotes and wildcards. However, representing only relevant MWEs instead of all possible *n*-grams in the documents could speed up the searches.

Joining the words of MWEs before indexation is a simple idea that was put in practice by Acosta et al. (2011). They tested the impact of a large set of automatically and manually acquired MWE dictionaries on standard IR test sets from the CLEF campaign. Their results show that there is a gain in mean average precision when MWEs are tokenised as single words prior to tokenisation.

Choosing the appropriate granularity for units to be indexed is even more complicated in languages like Chinese, which do not use spaces to separate words. In this case, a prior phase of segmentation generally takes place before traditional IR indexation. Xu et al. (2010) propose a new measure for the tightness of 4-character sequences, as well as three procedures for word segmentation based on this measure. Then, they compare a

standard segmentation tool with their methods in an IR system. They show that two of their segmentation strategies improve mean average precision on a test set.

A related task is topic modelling, a popular approach to joint clustering of documents and terms. The standard document representation in this task is a bag of words. However, as presented by Baldwin (2011), it is possible to consolidate the microlevel document representation with the help of MWEs. He argues that recent experimental results demonstrate that linguistically-richer document representations can enhance topic modelling.

3.3.4.4 Machine translation

Related work on the integration of MWEs into MT systems is discussed in Section 7.2.

3.4 Summary

Before diving into the vast literature on MWE processing, let us revise some elementary notions. A *corpus* is simply a body of texts used in empirical language studies (Manning and Schütze 1999, p. 6). *Linguistic analysis* is the process of creating more abstract representations from raw text in corpora. It can be seen as a set of steps going from more concrete to more abstract representations: sentence splitting, tokenisation, lemmatisation, POS-tagging and dependency parsing.

The underlying hypothesis in MWE acquisition is that words that form an expression will co-occur more often than if they were randomly combined. This hypothesis is applied in the design of lexical association measures (AMs) for corpus-based MWE acquisition. There are numerous AMs available for MWE acquisition (Evert 2004, Seretan 2008, Pecina 2008b). For an arbitrary n-gram w_1^n, we estimate its probability under MLE as $p(w_1^n) = \frac{c(w_1)\times c(w_2)\times \ldots \times c(w_n)}{N^n}$. When we scale this estimate by the total number of n-grams in the corpus N, we obtain the expected count $E(w_1^n) = \frac{c(w_1)\times c(w_2)\times \ldots \times c(w_n)}{N^{n-1}}$. AMs are generally based on the difference between the expected count $E(w_1^n)$ and the observed count $c(w_1^n)$, for example:

$$\texttt{t-score} = \frac{c(w_1^n) - E(w_1^n)}{\sqrt{c(w_1^n)}}, \quad \texttt{pmi} = \log_2 \frac{c(w_1^n)}{E(w_1^n)}, \quad \texttt{dice} = \frac{n \times c(w_1^n)}{\sum_{i=1}^n c(w_i)}.$$

More robust and theoretically sound AMs based on contingency tables exist for the special case of 2-grams. Examples of such measures are given below, where $w_i \in \{w_1, \neg w_1\}$ and $w_j \in \{w_2, \neg w_2\}$:

$$\chi^2 = \sum_{w_i,w_j} \frac{[c(w_i w_j) - E(w_i w_j)]^2}{E(w_i w_j)}, \quad \texttt{ll} = 2 \times \sum_{w_i,w_j} c(w_i w_j) \times \log \frac{c(w_i w_j)}{E(w_i w_j)}.$$

MWE acquisition comprises identification (in context) and extraction (out of context). Monolingual MWE acquisition is generally seen as a two-step process.

1. **Candidate extraction**: POS sequences are one of the major approaches, specially for terminology (Justeson and Katz 1995, Daille 2003), but also in noun compounds (Vincze et al. 2011) and verbal expressions (Baldwin 2005b). When a parser is available, syntactic patterns can be much more precise than POS sequences, specially in the extraction of non-fixed MWEs (Seretan and Wehrli 2009, Seretan 2008). Tree substitution grammars (Green et al. 2011) and structural regularities in parsing trees (Martens and Vandeghinste 2010) can also be used in order to learn syntactic MWE models from annotated corpora. The LocalMaxs algorithm performs extraction based on the maximisation of an AM applied to adjacent

word pairs (Silva and Lopes 1999). A string matching algorithm inspired by computational biology has been proposed to extract sequences that occur recurrently throughout the corpus (Duan et al. 2006).

2. **Candidate filtering**: some straightforward procedures are the use of stopword lists and of count thresholds. AMs are also widely employed to rank the candidates and keep only those whose association score is above a certain threshold (Evert and Krenn 2005, Pecina 2005). Supervised learning methods can be used to build a classifier modelling the optimal weights of several AMs and other features (Ramisch et al. 2008b, Pecina 2008b).

Some freely available tools that can be used for monolingual MWE acquisition include LocalMaxs, [36] Text::NSP, [37] UCS, [38] jMWE, [39] and Varro. [40] There are also freely available web services, downloadable tools and numerous commercialised systems for automatic terminology extraction from specialised texts.

As for bilingual acquisition, automatic word alignments can provide lists of MWE candidates by themselves (de Medeiros Caseli et al. 2010). Bai et al. (2009) present an algorithm capable of mining translations for a given MWE in a parallel aligned corpus. The automatic discovery of non-compositional compounds from parallel data has been explored by Melamed (1997). The English-Hindi language pair presents large word order variation, and it has been shown that MWE-based features that model compositionality can help reducing alignment error rate (Venkatapathy and Joshi 2006). Zarrieß and Kuhn (2009) used syntactically analysed corpora and GIZA++ alignments to extract verb-object pairs from a German–English parallel corpus. Daille et al. (2004) performed multiword term extraction from comparable corpora in French and in English, and subsequently used the distances between the context vectors to obtain cross-lingual equivalences.

There is a considerable amount of related work in other tasks related to MWE treatment, as summarised below.

– **Interpretation**: The *syntactic interpretation* of nouns compounds has been explored by Nicholson and Baldwin (2006), who distinguish three syntactic relations in noun–noun compounds: subject, direct object and prepositional object. Three-word or longer noun compounds require syntactic interpretation of the constituent hierarchy. Nakov and Hearst (2005) compare two models, based on adjacency and on dependency, using a set of heuristics to generate surface-level paraphrases and then use search engine counts to estimate model probabilities. Nakov and Hearst (2008) perform unsupervised *semantic interpretation* of noun compounds by generating a large number of paraphrases involving verbs related to the semantic classes and then retrieving their web counts. Kim and Nakov (2011) used a combination of data bootstrapping and web counts, using paraphrases based on parse trees, thus obtaining more accurate results. Besides noun compounds, other MWE types also require interpretation. Cook and Stevenson (2006) use support vector machines to classify the meaning of the particle *up* in English phrasal verbs. Bannard (2005) investigates the extent to which the components of a phrasal verb contribute their meanings to the interpretation of the whole. A similar work is that of McCarthy

36. http://hlt.di.fct.unl.pt/luis/multiwords/
37. http://search.cpan.org/dist/Text-NSP
38. http://www.collocations.de/software.html
39. projects.csail.mit.edu/jmwe
40. http://sourceforge.net/projects/varro/

et al. (2003), who propose several measures involving an automatically acquired thesaurus in order to estimate the idiomaticity of phrasal verbs.

– **Disambiguation**: The disambiguation of MWEs is analogous to their interpretation, except that they are considered together with the context in which they appear. Nicholson and Baldwin (2008) present a data set for noun–noun compound disambiguation where a large set of sentences has been manually annotated. Girju et al. (2005) investigate methods for their disambiguation by applying several supervised learning techniques. Fritzinger et al. (2010) manually analyse a large set of ambiguous German preposition–noun–verb constructions retrieved by a parser, classifying them as either literal, compositional or unknown. Light verbs in Japanese have also been studied by Uchiyama et al. (2005), who proposes two disambiguation methods: a statistical approach and a rule-based method. Cook et al. (2007) explore the idiomaticity of verb–noun pairs, where the noun is the direct object of the verb and may have an idiomatic (*make a face*) or literal (*make a cake*) interpretation. Fazly and Stevenson (2007) propose a more fine-grained classification for light verb–noun constructions, using a supervised learning strategy in order to perform a 4-way semantic disambiguation.
– **Representation**: The lexical representation of MWEs has for a long time haunted lexicographers in the compilation of lexical resources. Sag et al. (2002) proposed two approaches: words-with-spaces and compositional. However, between these extremes of the compositionality spectrum, there are some other possibilities, explored in related work. Laporte and Voyatzi (2008) describe a dictionary of French adverbial expressions and their corresponding morphosyntactic patterns in the lexicon–grammar format. Graliński et al. (2010) present a qualitative and quantitative comparison between two structured representations, Multiflex and POLENG, for Polish MWEs. Grégoire (2007; 2010) uses an equivalence class method that groups similar expressions according to their syntactic characteristics. Izumi et al. (2010) suggest a rule-based method to normalise Japanese functional expressions in order to optimise their representation. Schuler and Joshi (2011) propose the use of tree-rewriting grammars for describing MWEs.
– **Applications**: There are some target applications for which concrete results have been obtained. For instance, concerning syntactic analysis, Constant and Sigogne (2011) present promising results for French POS tagging. Korkontzelos and Manandhar (2010) obtain impressing improvements by enriching a baseline shallow parser with MWEs. Zhang and Kordoni (2006) and Villavicencio et al. (2007) obtain a significant coverage increase by extending the lexicon of an English HPSG parser with MWE entries. Wehrli et al. (2010) demonstrate that MWEs are not a "pain in the neck" but actually a valuable information to reduce syntactic ambiguity. Another example of successful MWE application is information retrieval. Acosta et al. (2011) join the words of MWEs before indexation, showing that there is a gain in mean average precision. Xu et al. (2010) propose a new measure for the tightness of 4-character sequences in Chinese and also improve mean average precision on a test set.

Part II

Automatic MWE acquisition

4 EVALUATION OF MWE ACQUISITION

From a generic point of view, the result of automatic MWE acquisition can be viewed as a list of MWE candidates. The evaluation of the quality of a given approach for MWE acquisition can be thought of as the estimation of the utility of the resulting MWE candidate list for a given application. This list has often an internal structure, and each candidate contains attached information about its properties, coming from corpora or from external resources. However, if we ignore this extra information (which is often the case in related work), it is possible to define objective criteria for determining the quality of the list. Analogously to information retrieval systems, whose result is a list of documents, each MWE candidate is classified as either relevant or irrelevant for the target application. Afterwards, we estimate the proportion of relevant MWEs in the list (precision), which indicates the amount of work that a human expert would need to perform, using this method, to transform a rough list of automatically acquired candidates into a lexical resource that can be used by the application.

However, the problem of MWE acquisition is quite complex because results depend on many parameters of the acquisition context, as we will detail later in this chapter. Precision alone cannot evaluate the quality of acquisition methods. As a consequence, in this chapter our goal is two-fold: first, we would like to introduce a series of background concepts and measures commonly used in the evaluation of MWE acquisition in a given context (Section 4.1). Second, we would like to present the variable parameters of the acquisition context that may have an influence on the evaluation results (Section 4.2). These parameters are the reason why evaluation is hard: they make results obtained in one context difficult to generalise to another context. We close this chapter with a brief discussion of the advantages and inconvenients of different evaluation types, arguing that application-oriented, extrinsic evaluation is required to build solid arguments towards the utility of MWEs in NLP systems in general (Section 4.3).

4.1 Evaluation context

Before starting an evaluation, there are mainly four questions that one should ask:

1. What are the acquisition goals (that is, the target applications) of the resulting MWEs?
2. What is the nature of the evaluation measures that we intend to use?
3. What is the cost of the resources (dictionaries, reference lists, human experts) required for the desired evaluation?
4. How ambiguous are the target MWE types?

The answers to these questions can be modelled as a set of four independent evaluation axes that we describe in Section 4.1.1. These axes constitute a new typology that we propose for the evaluation of MWE acquisition. Since these axes are parameters of the *evaluation context*, they will determine the kind of annotation performed (Section 4.1.3) and the objective evaluation measures that are going to estimate the utility of a resulting MWE list (Section 4.1.2).

4.1.1 Evaluation axes

In the literature of MWE acquisition, there are several prototypical styles of evaluation. First, some work present the results of their methods by showing a list of the top-*k* MWEs returned according to some ranking criterion (da Silva et al. 1999). In terms of quantitative evaluation, it is possible to manually annotate these top-*k* candidates, obtaining an objective estimation of the method's precision (Seretan 2008). Traditional measures based on the information retrieval analogy report precision and recall with respect to a gold standard dictionary, trying to optimise the balance between both in order to obtain a reasonable F-measure (Ramisch 2009). In the evaluation of association measures, in order to avoid setting a hard threshold, it is possible to average precision over all recall points, thus comparing cross-measure quality through mean average precision (Evert and Krenn 2005). Given one or more objective evaluation measures, it is possible to perform a simultaneous comparative evaluation of a set of methods (Pearce 2002, Ramisch et al. 2008a). Finally, the use of the acquired MWEs in a real application can give a concrete measure of the utility of the method. In this case, evaluation of MWE acquisition is performed implicitly through the measures traditionally used to evaluate the target application (Finlayson and Kulkarni 2011, Xu et al. 2010, Carpuat and Diab 2010).

In order to provide a more structured view of the evaluation of MWE acquisition methods, we propose a new typology that classifies existing evaluation styles according to four independent axes. These axes try to bring a systematic answer to the questions asked in the introduction of Section 4.1.

4.1.1.1 According to the acquisition goals

- **Intrinsic**. Most evaluation results reported in related work are intrinsic, that is, they evaluate the MWEs by themselves, directly, as a final product in a process. This is the case, for instance, when one annotates top-*k* candidates or uses a gold standard to automatically calculate precision and recall (defined in Section 4.1.2). The problem with intrinsic evaluation is that, as the definition of MWE depends on the target application (see Definition 2.1), it is often very hard to provide consistent annotation guidelines. Annotation guidelines aim at helping a human judge decide whether a word combination can be considered as a true MWE or whether it is an uninteresting word combination. The coherence and the precision of the guidelines determine the inter-annotator agreement, and a poor agreement makes evaluation of little use as it is highly unreliable. Even though it has numerous limitations, intrinsic evaluation still provides an estimation of the quality of the extracted MWEs that can be compared to related work (assuming the same available dataset).
- **Extrinsic**. Sometimes it is easier to evaluate a NLP application than a list of MWEs. For example, many linguistic tests for detecting light verb constructions use a workaround of trying to translate the expression in another language (Langer 2004). If there is no word-for-word translation can be found, this indicates that the combination needs to be treated as a unit. Therefore, manual or automatic transla-

tion can be considered as an application that is relatively easy to evaluate by a non-expert native speaker according to objective criteria such as accuracy and fluency. If confronted to the analogous problem of judging whether a word combination is a MWE, the same native speaker would probably find it more difficult. Therefore, while intrinsic evaluation often requires expert linguists to judge the data, extrinsic evaluation can be performed using the standard measures used to evaluate the target application. Furthermore, extrinsic evaluation, that is, the use of MWEs inside an external application, can be very conclusive in demonstrating whether acquired MWEs are useful in a given task. Extrinsic evaluation is a current trend in evaluation of MWE acquisition and our work presents two results of extrinsic evaluation applied to computer-aided lexicography (Chapter 6) and to statistical machine translation (Chapter 7). As the evaluation axes in extrinsic evaluation depend on the target application, the remaining three evaluation axes presented below apply only for intrinsic evaluation.

4.1.1.2 According to the nature of measures

- **Quantitative**. A quantitative evaluation assumes the use of objective measures like precision, recall, F-measure, and mean average precision. While many papers only report precision for top-k MWEs, it is important to evaluate recall. This is rarely done but nevertheless of capital importance in assessing the utility of a method. If it extracts only a dozen expressions when there are millions to be retrieved, it will not be more effective than brute force or manual search. The amount of (new) MWEs discovered is as important as their quality, and it is hard to evaluate how many MWEs are "enough" for the automatic acquisition to be useful (Villavicencio et al. 2005b, Church 2011). A summary of the measures most often used in the quantitative intrinsic evaluation of MWE acquisition is provided in Section 4.1.2.
- **Qualitative**. The goal of qualitative evaluation is to obtain a deep understanding of the mistakes done by the acquisition method and, as a consequence, of the target MWEs. Therefore, one tries to extract patterns of correctly/incorrectly acquired MWEs through observation of the resulting lists in terms of criteria such as POS sequences, frequency distributions and context. Qualitative evaluation is often iterative: (i) a first run of the acquisition method provides rough MWE candidates, (ii) a qualitative evaluation allows the identification of problems in the acquisition method (iii) the problems are then corrected if possible, and a new run provides a better set of MWE candidates, and so forth. Qualitative evaluation can be achieved by manual inspection of the data, statistical analysis and questionnaires. It is not impossible to perform both quantitative and qualitative analysis either simultaneously or at different steps of the acquisition.

4.1.1.3 According to the available resources

- **Manual annotation**. Traditionally, after acquisition is performed, one defines criteria to select a representative sample of the output (often a couple of hundred candidates). Then, a group of native speakers will go through the list, making a binary decision on whether the proposed word combination is a true MWE. This process depends on the availability of (volunteer) native speakers to perform the annotation. Ideally, a large sample should be annotated in order to obtain more consistent evaluation measures. Unfortunately, annotation can be quite time consuming and, depending on the type of expression, it may require annotation by

expert native speakers like lexicographers and linguists (in opposition to laymen). Some advanced topics on data annotation for the evaluation of MWE acquisition are presented and discussed in Section 4.1.3.

– **Automatic annotation**. In automatic annotation, one considers that a lexical resource containing the target MWEs already exists. This lexical resource can be a regular dictionary or a simple list of MWEs, and is often referred to as *gold standard* or *reference dictionary*. For performing automatic annotation, it is necessary to assume that the existing gold standard is complete or at least that it has a broad coverage of the target MWEs. Thus, we consider that candidates contained in the gold standard are true positives (genuine/interesting MWEs) while those not contained in the gold standard are considered as false MWEs. This is a strong assumption, as we discuss in Section 4.1.3.

4.1.1.4 According to the type of MWE

– **Type-based evaluation**. Some expressions are non-ambiguous and can be annotated out of context, as entries in a lexicon. Examples include most compound nouns and technical terminology, as well as support verb constructions. The decision of whether a sequence of words is a MWE, in this kind of annotation, is independent from the context in which it occurs. On the MWE community website, several lexicons that can serve as gold standards for type-based evaluation are available. Examples include a lexicon of French adverbial expressions (Laporte and Voyatzi 2008) and a lexicon of German preposition-noun-verb constructions (Krenn 2008). A lexicon for type-based annotation can be a simple list of MWEs or it may contain additional information, useful for other MWE tasks like interpretation. Datasets including additional information contain, for example, information about the syntactic relation between the words (Nicholson and Baldwin 2008) and about semantic relations through paraphrases (Nakov 2008b). In the context where no gold standard data set is available, type-based annotation must be performed manually by human judges.
– **Token-based evaluation**. Token-based evaluation must be performed whenever the target MWEs are ambiguous, such as phrasal verbs and idioms. Out of context, it is impossible to tell whether the words should be treated as a unit or separately. For example, *look up* may be an idiomatic expression meaning to consult a dictionary or a regular verb-adverb combination meaning to look to a higher position. Therefore, token-based evaluation requires manual annotation, and human judges annotate a whole sentence instead of only the MWE candidate. Data sets of sentences with token-level MWE annotations include, for example, English idiomatic verb-noun constructions (Cook et al. 2007; 2008), English verb-particle constructions (Baldwin 2008), and German verb-preposition-noun constructions (Fritzinger et al. 2010).

4.1.2 Evaluation measures

The intrinsic quantitative evaluation of MWE acquisition uses standard measures that were originally proposed in the context of information retrieval systems. The analogy is quite straightforward: ranked candidates can be judged as interesting/uninteresting with respect to a target MWE in the same way as ranked documents are assigned relevance judgements according to a query.

First, let us model the result of MWE acquisition as a list C of MWE candidates sorted

according to some numerical score (typically, AMs as those described in Section 3.1.4). This corresponds to the list of candidates considered as "positive" instances. There are several means to assign a binary value to each element (discussed in Section 4.1.3), judging its relevance as a *true positive* (TP) or as a random/uninteresting word combination. A popular evaluation metric considers the binary annotation of the first k sorted candidates (denoted $C_{[1..k]}$). When we consider the rate of true MWEs among the annotated data, the accuracy of the acquisition is denoted as *precision at k* (P@k):

$$\text{P@k}(C,k) = \frac{|\text{TPs in } C_{[1..k]}|}{k} \tag{4.1}$$

When we set k to a reasonable value (say 100 or 200), annotation by a couple of native speakers is fast. However, it is better if we can evaluate the true precision P(C) of the system by annotating the whole set of returned n-grams. The precision is the proportion of n-grams judged as true MWEs in the set of returned n-grams:

$$\text{P}(C) = \frac{|\text{TPs in } C|}{|C|}$$

Precision measurements indicate the amount of work needed to transform the rough list of automatically acquired MWEs into a final list validated by a specialist (e.g., an experienced lexicographer). However, the use of precision alone oversimplifies the evaluation. What about the elements that should have been returned and that were ignored? If a system only acquires a dozen expressions, even though its precision is close to 100%, this is probably not enough for building a dictionary. Therefore, in addition to precision, it is crucial to calculate the recall R:

$$\text{R}(C) = \frac{|\text{TPs in } C|}{|\text{Total MWEs to acquire}|}$$

The F-measure, that is, the harmonic mean of precision and recall, is frequently used as an overall performance measure:

$$\text{F}(C) = \frac{1}{\frac{1}{2} \times \left(\frac{1}{\text{P}(C)} + \frac{1}{\text{R}(C)} \right)} = \frac{2 \times \text{P}(C) \times \text{R}(C)}{\text{P}(C) + \text{R}(C)}$$

In spite of its importance, R(C) is rarely calculated because it is difficult to estimate the total number of MWEs that should be acquired by a system. If annotation is performed by humans, it means that the whole input corpus must be annotated with the target expressions, which is very onerous for small corpora and impracticable for larger corpora. If the annotation is automatic, based on the comparison with an existing dictionary, then the total number of MWEs to acquire corresponds to the size of the reference dictionary. However, this is very likely to be an underestimation, as new MWEs that were not present in the dictionary will be considered as errors.

Precision, recall and F-measure are independent of any particular ranking. When the list of candidates C is ranked according to a given AM, we can apply the mapping $rank_{AM}(c) : C \rightarrow [1..|C|]$ identically as in Section 3.1.2. Choosing a threshold below which the returned candidates are considered as negative instances is difficult, all the more because there is no systematic way to do it. Thus, it is possible to evaluate the

quality of the ranked candidates through its mean averaged precision (MAP), that is, the mean of the precisions taken at each TP:

$$\mathrm{MAP}(C) = \frac{\sum_{r=1}^{|C|} \mathrm{P@k}(C,r) \times \mathrm{isTP}(C,r)}{|\mathrm{TPs\ in}\ C|},$$

where the binary function $\mathrm{isTP}(C,r)$ is defined as follows:

$$\mathrm{isTP}(C,r) = \begin{cases} 1 & \text{if } (\exists c \in C)[rank_{AM}(c) = r \wedge c \text{ is a TP}], \\ 0 & \text{else.} \end{cases}$$

The precision $\mathrm{P@k}(C,r)$ of a given candidate rank r is defined as in Equation 4.1. It corresponds to the number of TPs up to rank r in C divided by the total number of candidates whose rank is less than or equal to r in C. If we plotted a graph with recall on the abscissa and precision on the ordinate, MAP would correspond to the area below the curve (Evert 2004).

4.1.3 Annotation

There are two types of annotation: manual and automatic. In this section, we will underline some decisions that should be taken when constituting an annotated data set for evaluation. In automatic annotation, one considers that there is a static *gold standard* (GS) or *reference*, that is, a lexicon containing the complete list of MWEs that should have been returned by the acquisition method. The candidates that are found in the GS are referred to as *true positives*. In that case, the interpretation of precision and recall, defined in Section 4.1.2, is as follows:

- Precision (P): proportion of MWE candidates that are present in the gold standard
- Recall (R): proportion of MWEs in the gold standard that are present in the list of candidates

Formally, the measures can be redefined in terms of set operations. For a set of candidate MWEs C and a set GS of true MWEs:

$$\mathrm{P}(C) = \frac{|C \cap \mathrm{GS}|}{|C|}$$

$$\mathrm{R}(C) = \frac{|C \cap \mathrm{GS}|}{|\mathrm{GS}|}$$

Both measures are underestimations as they assume that candidates not in the gold standard are false MWEs, whereas they may simply be absent from dictionaries due to coverage limitations. Conversely, these measures assume that all entries in GS are true MWEs, whereas this may depend on the acquisition goal or context. The automatic evaluation of the candidates will always be limited by the coverage of the reference list. Moreover, when calculating the intersection between C and GS, one needs to be very careful to take into account the normalisation of entries. In previous experiments, for example, *Panama Canal* was considered as a true MWE whereas *US navy* was not. Both are proper names and the latter should also be included in the set of true positives. This could be the case if the lexicon uses a different capitalisation (*US Navy*) or expands acronyms (*United States navy*). Therefore, even in the case of automatic annotation, a careful data inspection must be carried out to assure that such cases are dealt with. Finally, some ambiguous

cases may be difficult to judge, for example, if the set of candidates *C* contains the entry *human right*, should it match the *GS* entry *Human Rights*?

Automatic annotation is used to evaluate the accuracy of the acquisition method in relation to the GS, but it does not necessarily correspond to an informative evaluation of the usefulness of the acquired MWEs. In other words, it is pointless to acquire MWEs that are already known (see Section 4.2.3). In spite of all these disadvantages, automatic annotation is often employed. Its advantages are mainly that it represents a quite cheap and quick way to evaluate a technique. When compared to manual annotation, it provides an underestimation of recall, which is important and cannot be calculated through manual annotation. Indeed, it would be impracticable to ask annotators to go through the whole corpus and manually identify all the MWEs that should be returned by an acquisition method. The use of a GS depends on its availability or cost.

Manual annotation is rarely performed on the whole list of resulting MWEs. These lists can contain several thousand MWE candidates, and manually annotating all of them would be too onerous or infeasible. Hence, a first decision that needs to be made concerns the sample of data to annotate. If the list of MWEs is ranked, the most natural choice is to annotate the top-*k* candidates. However, this can be biased because, if the top of the list contains mostly frequent combinations, they are likely to be known MWEs already present in a lexicon. Indeed, expert lexicographers tend to consider less frequent items as more interesting because they are more likely to be of interest for dictionary users, who are already familiar with the most frequent ones. A fairer evaluation would consider a balanced amount of candidates from high, medium and low frequency ranges. Moreover, this kind of evaluation gives an idea of a method's precision but ignores its recall, regardless of the sampling technique employed.

The second important decision in manual annotation concerns the definition of the target public. Once the data sample is ready, designing the evaluation guidelines for the annotators requires careful planning. As MWEs are complex phenomena, a group of two or three native speakers may be enough. Depending on the availability of native speakers and on their familiarity with computers, one can develop a web interface or use platforms like Amazon Mechanical Turk to gather annotations (Nakov 2008b). However, if the phenomenon is hard to circumscribe, sometimes expert linguistic knowledge is required to perform the annotation. For example, it is difficult to distinguish general-purpose language like *travel photos* from more specialised cases like *lending institution* and *security institution*. For non-experts, it is not clear why the first candidate is not considered as a true MWE while the second and third ones are.

Third, it is necessary to define which candidates should be annotated as true positives, providing precise descriptions and some examples. In this case, it can be useful to define questions that help the annotators, for example: (i) *can the construction be translated word for word in another language?* or (ii) *can the meaning of this expression be derived from the meanings of its parts?*. These questions can be either based on the target application, like question (i), or on known properties of the target MWEs as described in Section 2.3.1, like question (ii). Even though precision and recall require yes/no judgements, when it comes to human annotators it is recommended to avoid binary answers and to allow some room for flexibility, like multiple categories (e.g., *true MWE, maybe a MWE, part of a MWE, random word combination, unknown*) or numerical scales for semantic compositionality (McCarthy et al. 2003). Posteriorly, one can homogenise the answers or keep only those candidates for which a sharp binary class has been assigned.

Fourth, once the data set has been annotated by more than one human judge, native speaker or expert, it is necessary to calculate the agreement between the annotations. Traditionally, Fleiss' kappa agreement score is used to estimate how much annotators agree above of what would be expected by chance (Fleiss 1971). Fleiss' kappa has the advantage that it allows multiple annotations per item (in opposition to Cohen's kappa, which assumes two annotators). However, its use is considered unsafe as its interpretation is controversial (is $\kappa = 0.6$ a good agreement?) and as its value depends on the number of annotations per item and of available annotation categories (Eugenio and Glass 2004). Several heuristics have been used to report evaluation results based on manual annotation: a second pass of annotation can be made in order to solve the disagreements (Fazly and Stevenson 2007), or one can report both results, using the intersection and the union of MWE candidates considered as true positives, as lower and upper bounds, respectively, for the performance of the method (Linardaki et al. 2010).

In short, each annotation strategy has advantages and disadvantages. Automatic annotation is quick and cheap and provides an estimation of recall, but it tends to underestimate evaluation results, while manual annotation ignores recall but provides an accurate estimation of a method's precision. Furthermore, the choice between manual and automatic annotation is not not mutually exclusive. For instance, one of the goals of manual annotation may be the creation of resources for automatic evaluation. Many such data sets exist and are freely available on the MWE community website. [1] It is also possible to use mixed automatic and manual annotation, that is, entries absent from the gold standard are manually annotated.

4.2 Acquisition contexts

In Section 4.1.1, we defined four axes that describe the *evaluation context*. Here, we are interested in the *acquisition context*, that is, the set of parameters that can influence the results of evaluation. Both contexts, of acquisition and of evaluation, are closely correlated. For example, the type of acquired MWE is a characteristic of the acquisition context. Nonetheless, if the type is ambiguous (e.g., idiomatic expressions), the evaluation must be type-based. Similarly, if the acquisition context is in a language for which no gold standards are available, the evaluation must be performed manually. Therefore, the generalisation of evaluation results depends simultaneously on all the parameters of the acquisition context. This implies that a truly extensive evaluation of methods for MWE acquisition should explore all possible values for each parameter, which is impracticable. Generally, comparative evaluations of MWE acquisition tend to use a fixed test set from which conclusions are drawn (Pearce 2002, Ramisch et al. 2008a). The goal of this section is to argue that such evaluations are of limited value, as they are hard to generalise because the results depend on numerous parameters. According to our experience, the most important parameters are the characteristics of the target MWEs (Section 4.2.1), the size and nature of the resources from which the MWEs are acquired (Section 4.2.2) and the existing lexical resources present prior to acquisition (Section 4.2.3).

4.2.1 Characteristics of target constructions

The characteristics of the target MWEs influence the generalisation of evaluation results. The literature in MWE acquisition reports a plethora of methods for MWE acquisi-

1. `http://multiword.sf.net`

tion and they are mostly motivated by different types of target MWEs. As the definition of MWE includes very heterogeneous phenomena, showing that a method performs well for a given MWE type is not enough to conclude that its performance is superior to other methods for other types of MWEs. In this section, we will provide examples of MWE evaluation results that cannot be straightforwardly generalised due to characteristics of the target constructions like their type, language and domain.

4.2.1.1 Type

In the typology proposed in Section 2.3.3.2, we suggest that MWEs can be classified according to the difficulty to deal with them in computational applications. Therefore, it is natural that different MWE types require different acquisition techniques and, as a consequence different evaluations. For example, a method that is usually employed for candidate extraction in the acquisition of noun compounds is the use of sequences of parts of speech. This is not adequate for extracting English verbal expressions (Villavicencio et al. 2012) or to extract flexible "true" collocations (Seretan 2008). Methods based on the flexibility of a word combination need to be adapted to each type of construction: syntactic and semantic variations are not the same for nominal expressions and verbal expressions (Pearce 2002, Ramisch et al. 2008a;b).

4.2.1.2 Language

The language is also an important characteristic of the target MWE that constitutes a parameter of the acquisition context. To start with, for a very simple reason: corpora and preprocessing tools available for different languages are not the same. For instance, one may argue that the acquisition of "true" collocations is much more efficient with the use of a deep syntactic parser, and this claim is justifiable (Seretan 2008). However, if the target language is not a major one (e.g., French, English, Russian, Chinese), for which a deep parser is available, then it is not possible to apply such deep method and shallow alternatives are required. Also, methods that depend on parallel corpora like Europarl or on very large monolingual corpora like the BNC may not be easily adaptable to other languages simply because these resources do not exist in less resourced languages.

There is also another issue with cross-language adaptation, which is more related to the MWEs themselves. Even though existing typologies try to model MWEs in a generic way, so that the types are language independent, MWEs are arbitrary and depend on the language. For example, English and many Germanic languages have a large set of phrasal verbs, which are mostly absent in Romance languages. Compound nouns in German and Swedish are concatenated together as a single lexical unit, while this phenomenon is much less frequent in English (e.g., *database*, *blackboard*) and in Romance languages (e.g., chemical components in French).

4.2.1.3 Domain

Finally, the domain of the expression needs to be taken into account when evaluating MWE acquisition. Justeson and Katz (1995) suggest a list of POS patterns for the automatic acquisition of terms. However, when applied to the biomedical domain, these patterns yielded a poor performance (Ramisch 2009). The original patterns were adapted by considering characteristics of the domain such as the fact that biomedical MWTs are longer than terms in other domains and often contain foreign words and numbers. This improved the performance of the acquisition significantly. Also, methods that are aimed at

MWT will not necessarily perform well for acquiring general-purpose MWEs. For example, contrastive methods, such as the one used by Bonin et al. (2010b) for the acquisition of legal end environmental MWTs from Italian corpora, rely on comparing the distribution of MWEs between specialised and general-purpose corpora. In the case where these counts are similar in both corpora, as it is the case for general-purpose MWEs such as light verb constructions, contrastive methods will not work.

4.2.2 Characteristics of corpora

The characteristics of the target construction are not the only parameter of the acquisition context that can affect evaluation results and their generalisation. For instance, a method that was optimised for a large in-domain English corpus may have a very poor performance in other languages like Portuguese, for which only general-purpose and/or small corpora are available. Among the characteristics of corpora that may heavily influence evaluation results, are its size, its nature (general-purpose, specialised, web as corpus) and the level of linguistic analysis used as preprocessing for candidate extraction.

4.2.2.1 Size

The size of the corpus from which extraction is performed can influence results at two points. First, larger corpora contain more data, so that intuitively a MWE acquisition method will be able to retrieve more candidates, increasing its coverage (recall). Second, statistical methods relying on token counts can be sensitive to data sparsity, and a larger sample allows more precise statistical measures to be deduced from it, potentially increasing the precision of the method.

An evidence for the first affirmation, that is, that larger corpora increase the recall of an acquisition method, is presented in Villavicencio et al. (2005b, p. 425). In these experiments, the use of increasingly larger corpora makes an initial lexicon of around 4,000 verb-particle constructions grow to around 7,000 entries using the BNC and to around 20,000 verified entries using the web as a corpus. Analogously, in the experiments reported in Section 5.3, we use three fragments of increasing sizes of the Europarl corpus. The recall of *n*-gram approaches like `NSP` and the `mwetoolkit` increases from around 83% in the small corpus to more than 89% in a large corpus 100 times larger (see Table 5.4).

The second advantage of using larger corpora is that they are more representative samples of language, thus yielding more reliable statistics. Sparsity is particularly dangerous when it comes to association measures. Dunning (1993), for instance, showed that normal assumptions do not hold for small samples, and that the log-likelihood ratio is much more adequate for these cases because it assumes a LNRE distribution. Pedersen (1996) suggests Fisher's exact test as a very robust measure, and Evert (2004) shows that it approximates quite well the values of Dunning's log-likelihood ratio. In all cases, when applying association measures, one should perform a frequency cut based on a minimal occurrence threshold. Thus, more data means less discarded candidates according to this criterion, in addition to more accurate AMs (and potentially, more precise acquisition methods).

4.2.2.2 Nature

Evaluation results depend on the nature of the corpus. By *nature*, we mean its characteristics, which can be summarised as the domain and genre of the texts. Additionally,

traditional text corpora differ significantly from the use of the web as a corpus. Experimental results show that, in the task of specialised noun compound extraction, the use of the web as a corpus is not recommended (Ramisch et al. 2010d). The counts derived from such a generic resource are too noisy to be used as the base of association measures.

Some techniques use several corpora for acquisition, hence the nature of all of them needs to be taken into account. For example, contrastive measures for multiword term detection require the use of at least two corpora: a specialised one and a general-purpose one (Bonin et al. 2010b). A bad choice in either can decrease the quality of acquired MWEs. For instance, the contrastive measure `simple-csmw` sharply reflects the difference between using a traditional corpus or the web as contrastive corpus, obtaining a MAP of 51.76% for Europarl and 38.5% for Google, even if the latter is several orders of magnitude larger than the former. This is an indication that the source of count information significantly affects the results. A traditional general-purpose corpus yielded good results even when more than 90% of the counts were zero, since these may provide some information about the degree of specialisation of the candidate, while the web was not a good contrastive corpus because of its unboundedness.

On the other hand, the web can be quite useful in tasks that involve the extraction of more generic MWEs. As an example, Section 6.1 and Section 6.2 illustrate its usefulness in the acquisition of Greek nominal expressions and Portuguese verbal expressions, respectively. In particular, the experiments on Portuguese used traditional corpora and the web as a corpus (Duran and Ramisch 2011). The validation of sentiment expressions using the web, similarly to Villavicencio et al. (2005b), was particularly useful in helping to distinguish productive patterns from more rigid expressions. In these experiments, we also noticed that the genre of the corpus has an influence on the results. Since we were interested in sentiment expressions and since our corpus contained newspaper texts of the journalistic genre, most of the acquired expressions have negative polarity. This is probably a consequence of the fact that newspapers report more often bad news like tragedies and crisis, rather than good news involving joy and happiness.

4.2.2.3 Level of analysis

The level of linguistic abstraction used in candidate MWE extraction has an influence on the quality of the results. Existing acquisition methods vary much in the amount of linguistic preprocessing performed: from completely knowledge-free methods based on surface forms only (da Silva et al. 1999), to sophisticated methods depending on a specific type of syntactic formalism (Seretan 2008). See Section 3.1.1 for an overview of some linguistic analysis tools that can be used for MWE acquisition.

For the acquisition of verb-particle constructions in English, for example, Baldwin (2005b) proposes four methods that use increasing levels of linguistic abstraction, according to the preprocessing tools used: a POS tagger, a chunker, a chunk grammar and a syntactic parser. The use of complete (deep) syntactic analysis has been advocated by Seretan (2008), who targets general (non-fixed) collocations such as adverb-adjective and verb-object pairs. Depending on the language, the parts of the collocation may be separated by several intervening words, thus requiring some kind of tree representation. She argues that methods based on shallow POS patterns could not appropriately capture such long-distance relations, whereas syntax-based collocation acquisition has no problem with that. In the same lines, Green et al. (2011) present and evaluate a syntax-based method for the acquisition of French MWEs, showing significant improvements over a baseline based on shallow POS patterns.

The flexibility of the target construction may justify the use of language-dependent linguistic analysis tools, which will increase the usefulness of the resulting MWE list. On the downside, linguistic analysis tools are not readily available for all languages, specially poorly resourced ones. Moreover, it remains to be proven that deeper analysis yields better results: in the experiments of Baldwin (2005b), for instance, the syntactic parser does not systematically obtain the best F-measure. This may be a consequence of the heterogeneous performance of analysis tools themselves. Sometimes, it may be wiser to trust a reasonably good POS tagger than a parser that makes too many attachment errors. In short, the level of linguistic analysis recommended for a given acquisition goal depends on the flexibility of the target construction and on the availability of tools for the target language and domain.

4.2.3 Existing resources

The existence of lexical resources (printed and/or machine-readable dictionaries and thesauri) inventorying the target MWEs is a factor that influences the usefulness of automatic MWE acquisition methods. For example, in Section 6.2, we describe the acquisition of verbal MWEs in Portuguese, given the target application of (manual) semantic role labelling. In this case, the acquisition was motivated by the fact that there was no existing lexical resource containing such constructions for the Portuguese language. Therefore, the novelty of the extracted expressions was 100% and even simple techniques that could help speed up lexicographic work compared to manual corpus inspection was considered as extremely useful by the users. In Chapter 6, we explore the use of automatically acquired MWEs to speed up lexicographic work, given that, in the absence of previously existing lexical resources, "something is better than nothing."

However, when a lexical resource covering the target constructions already exists, there are two possibilities. First, the existing lexical resource can be used as gold standard for automatic annotation, assuming that the method returns no new MWE. Second, the evaluation may report not only classical precision/recall measures but also the novelty of the acquired MWEs. The former is clearly an over-simplification of reality, as it is unreasonable to assume that the previously existing resource has 100% coverage (otherwise, what is the point of performing automatic MWE acquisition?).

In order to estimate the novelty of the acquisition method, it is necessary to perform manual annotation of (a sample of) the candidates that were not present in the dictionary. Then, the novelty of the method can be defined as the ratio between true positives that were not present in the initial dictionary and the total number of true positives. Such measure is rarely performed in the context of MWE acquisition, and to the best of our knowledge none of the works cited in Chapter 3 report novelty results. However, in the context of bilingual lexical extraction from comparable corpora, the novelty of extracted translations is a good indicator of the utility of the method (Lee et al. 2010).

In short, when MWE acquisition is performed in the context of a real NLP application (in opposition to the context of experimental research), the existence of lexical resources containing MWEs must be taken into account.

4.3 Discussion

The evaluation of MWE acquisition is still an open problem. While classical measures like precision and recall based on automatic annotation assume that a complete (or at least broad-coverage) gold standard exists, manual annotation of top-k candidates and

mean average precision (MAP) are labour-intensive even when applied to a small sample, emphasizing precision regardless of the number of acquired *new* expressions. Nonetheless, objective measures provide a lower bound to the ability of a tool or technique to deal with a specific type of MWE.

On the one hand, the results of intrinsic evaluation are of limited value: although they shed some light on the optimal parameters for the given scenario, they are hard to generalise and cannot be directly applied to other configurations. The quality of acquired MWEs as measured by objective criteria depends on the language, domain and type of the target construction, on corpus size and genre, on already available resources, on the preprocessing steps, among others. On the other hand, extrinsic evaluation consists of inserting acquired MWEs into a real NLP application and evaluating the impact of this new data on the overall performance of the system. For instance, it may be easier to ask a human annotator to evaluate the output of a MT system than to ask whether a sequence of words constitutes a MWE.

As pointed out by Pecina (2005), "evaluation of collocation extraction methods is a complicated task. On one hand, different applications require different [...] thresholds. On the other hand, methods give different results within different ranges of their association scores". Efforts for the evaluation of MWE acquisition approaches usually focus on a single technique or compare the quality of association measures (AMs) used to rank a fixed annotated list of MWEs. For instance, Evert and Krenn (2005) and Seretan (2008) specifically evaluate and analyse the lexical AMs used in MWE extraction on small samples of 2-gram candidates.

Some efforts have been made toward comparative evaluations of MWE acquisition techniques. Pearce (2002) systematically evaluates a set of techniques for MWE extraction on a small test set of English collocations, emphasising association measures. Similarly, Pecina (2005) and Ramisch et al. (2008a) present extensive comparisons of individual AMs and of their combination for MWE extraction in Czech, German and English. Punctual comparisons have been performed, for instance, in order to compare candidate extraction based on POS sequences with that based on syntactic models (Schone and Jurafsky 2001). In Section 5.3, we report results of an evaluation of freely available tools compared to the framework proposed in the present work (Ramisch et al. 2012).

One recent initiative aiming at more comparable evaluations of MWE acquisition approaches was in the form of a shared task (Grégoire et al. 2008). However, the experiment presented in Section 5.3 differs from the shared task in its aims. The latter considered only the ranking of precompiled MWE lists using AMs or linguistic filters at the end of extraction. However, for many languages and domains, no such lists are available. In addition, the evaluation results produced for the shared task may be difficult to generalise, as some of the evaluations gave priority to the precision of the techniques without considering the recall or the novelty of the extracted MWEs. To date, little has been said about the practical concerns involving MWE acquisition, like computational resources, flexibility or availability. In our experiment, we hope to help filling this gap by performing a broad evaluation of the *acquisition process as a whole*, considering many different parameters.

There have also been efforts towards the extrinsic evaluation of MWEs for NLP applications such as information retrieval (Doucet and Ahonen-Myka 2004, Xu et al. 2010, Acosta et al. 2011), word sense disambiguation (Finlayson and Kulkarni 2011), MT (Carpuat and Diab 2010, Pal et al. 2010) and ontology learning (Venkatsubramanyan and Perez-Carballo 2004). An original contribution of the present work is application-oriented extrinsic evaluation of MWE acquisition on two study cases: computer-aided lexicogra-

phy (Chapter 6) and statistical machine translation (Chapter 7). Our goal is to investigate (1) how much the MWEs impact on the application, and (2) what is (are) the best way(s) of integrating them in the complex pipeline of the target application.

4.4 Summary

The problem of evaluating MWE acquisition is quite complex because results depend on many parameters of the acquisition context, making results obtained in one context hard to generalise. In related work, several evaluation styles are used: showing a list of ranked top-*k* MWEs (da Silva et al. 1999), manually annotate the top-*k* candidates (Seretan 2008), measure precision and recall with respect to a dictionary (Ramisch 2009), compare the quality of association measures through mean average precision (Evert and Krenn 2005), compare different methods (Pearce 2002, Ramisch et al. 2008a), and measure the impact of acquired MWEs on real NLP applications (Finlayson and Kulkarni 2011, Xu et al. 2010, Carpuat and Diab 2010). In order to provide a more structured view of evaluation, we propose the following typology for classifying the *evaluation context*:

1. **According to the acquisition goals**
   - **Intrinsic**. Results are reported by evaluating the MWEs by themselves, directly, as a final product in a process. Intrinsic evaluation heavily depends on the target application and on the coherence of the annotation guidelines, but it still provides a useful estimation of the quality of the acquired MWEs.
   - **Extrinsic**. Sometimes it is easier to evaluate a NLP application than a list of MWEs. Extrinsic evaluation can be performed by integrating MWEs into an application and then checking whether they improve the its output. It can be very conclusive in demonstrating whether acquired MWEs are useful.
2. **According to the nature of measures**
   - **Quantitative**. This assumes the use of objective measures like precision, recall, F-measure, and mean average precision. While many papers only report precision for top-*k* MWEs, it is important to evaluate recall, because the amount of (new) MWEs discovered is as important as their quality.
   - **Qualitative**. The goal is to understand the mistakes done by the acquisition method. Therefore, one observes the results in terms of POS sequences, frequency distributions, context, etc. Quantitative and qualitative analysis are complementary and can be performed simultaneously and/or iteratively.
3. **According to the available resources**
   - **Manual annotation**. A group of native speakers and/or experts will go through the list, deciding whether the proposed combination is a MWE. Annotation can be quite time consuming and depends on the availability of annotators, thus being performed on a small sample of the output.
   - **Automatic annotation**. In automatic annotation, one considers that a complete or at least broad-coverage dictionary of the target MWEs already exist. Thus, we consider that candidates contained in the dictionary are true positives (genuine/interesting MWEs) while the others are false MWEs.
4. **According to the type of MWE**
   - **Type-based evaluation**. Non-ambiguous expressions like compound nouns, terminology, and support verb constructions can be annotated out of context. Several lexicons that can serve as gold standards for type-based evaluation are available. When no such resource exists, annotation must be performed manually.

– **Token-based evaluation**. Token-based evaluation must be performed for ambiguous MWEs like phrasal verbs and idioms. Out of context, it is impossible to tell whether the words should be treated as a unit. In token-based evaluation human judges annotate a whole sentence instead a MWE candidate.

If we model the result of MWE acquisition as a list C of MWE candidates sorted according to some numerical score, the precision P(C) of the system is the proportion of n-grams judged as true MWEs in the set of returned n-grams, $\mathrm{P}(C) = \frac{|\text{TPs in } C|}{|C|}$. Precision indicates the amount of work needed to transform the rough list of automatically acquired MWEs into a final list validated by a specialist, but it ignores true MWEs that have not been found when they should. Therefore, it is crucial to calculate the recall $\mathrm{R}(C) = \frac{|\text{TPs in } C|}{|\text{Total MWEs to acquire}|}$. In spite of its importance, R(C) is rarely calculated because it is difficult to estimate the total number of MWEs that should be acquired by a system.

There are two types of annotation: automatic and manual. In automatic annotation, there is a static *gold standard*, that is, a lexicon containing the complete list of MWEs that should be found. In automatic annotation, P(C) and R(C) are underestimations as they assume that candidates not in the gold standard are false MWEs. In spite of this simplification, it is often employed, mainly because it is cheap and quick. Manual annotation is rarely performed on the whole list of resulting MWEs, but rather on a sample. If the list is ranked, the top-k candidates can be annotated, but this can bias evaluation towards highly frequent combinations whereas it should include candidates from all frequency ranges. It is important to carefully design evaluation guidelines for the annotators, who are a group of native speakers or, if the target MWEs are very complex, expert linguists. It is recommended to allow some room for flexibility, like multiple categories or numerical scales. Fleiss' kappa agreement score is often used to estimate inter-annotator agreement, even though its interpretation is controversial. Manual and automatic annotation are complementary, and it is possible to use mixed annotation, for example, entries absent from the gold standard are manually annotated.

The *acquisition context* is the set of parameters that can influence the results of evaluation. We argue that evaluation performed on a given acquisition context are hard to generalise because they depend on too many parameters.

Some parameters of the acquisition context are characteristics of the MWEs, such as:

– **Type**. Different MWE types require different evaluations. For example, POS sequences are usually employed for noun compounds acquisition but this is not adequate for verbal expressions (Villavicencio et al. 2012).
– **Language**. Not only MWEs but also NLP resources are not equivalent in different languages. The use of a parser for collocation acquisition like in Seretan (2008) is impossible for poorly resourced languages, requiring shallow alternatives.
– **Domain**. The domain of the expression needs to be taken into account in the evaluation. For example, the list of POS patterns suggested by Justeson and Katz (1995) yield a poor performance when applied to a biomedical corpus (Ramisch 2009).

Some parameters of the acquisition context are characteristics of corpora, such as:

– **Size**. Large corpora contain more data, so intuitively a method will be able to retrieve more candidates, increasing recall. Statistical methods can be sensitive to data sparsity, and larger samples allow more precise measures.
– **Nature**. Results depend on the domain and genre of texts. Experiments show that, in specialised noun compound extraction, the use of the web as a corpus is not recommended (Ramisch et al. 2010d).

– **Level of analysis**. Acquisition methods vary from shallow knowledge-free methods (da Silva et al. 1999) to those depending on a syntactic formalism (Seretan 2008). It remains to be proven that deeper analysis yields better results (Baldwin 2005b).

The evaluation of MWE acquisition remains an open problem. While precision and recall based on automatic annotation assume the existence of a complete gold standard, manual annotation is labour-intensive and emphasises precision regardless of the number of acquired new MWEs. There has been some effort towards comparative evaluation (Schone and Jurafsky 2001, Pecina 2005, Ramisch et al. 2008a) and towards extrinsic evaluation in NLP applications such as information retrieval (Doucet and Ahonen-Myka 2004, Xu et al. 2010, Acosta et al. 2011), word sense disambiguation (Finlayson and Kulkarni 2011), MT (Carpuat and Diab 2010, Pal et al. 2010) and ontology learning (Venkatsubramanyan and Perez-Carballo 2004).

5 A FRAMEWORK FOR MWE ACQUISITION

In the previous chapters, we motivated the importance of MWEs for NLP applications and provided a bibliographic review of past and present research in the area. We are now ready to present our new methodological framework for MWE acquisition. This framework was motivated by the absence of one covering all the steps of MWE acquisition in a systematic and integrated way. Thus, we have developed a methodology in which the process of MWE acquisition is divided into several independent modules that can be chained together in several ways. Each module solves a specific task in MWE treatment, and the modules themselves implement multiple and complementary techniques to solve the task.

We will detail our motivations, guiding principles and methodology in Section 5.1. Then, we will demonstrate how the methodology can be applied to an acquisition context (a corpus in a given language and domain and a target application) through a worked out toy experiment in Section 5.2. In Section 5.3, our methodological framework is systematically compared to other similar available frameworks, underlining their differences in terms of quality but also in terms of computational resources and flexibility.

5.1 Processing overview

The present section is consecrated to a detailed description of a new methodology for MWE acquisition which we baptised `mwetoolkit` for "multiword expressions toolkit". In the first subsection, we provide a general overview of our goals and of the main principles that guided us during the development of the methodology and of the corresponding implementation (Section 5.1.1). In the second subsection, we provide a detailed description of the modules composing our methodological framework and how they can be combined to achieve a given MWE acquisition goal (Section 5.1.2). Finally, we provide a more subjective discussion of the characteristics of the `mwetookit`. This discussion is followed by a thorough and systematic evaluation of the toolkit, which is described in Chapter 6 and in Chapter 7, allowing us to present our ideas for future developments, improvements and extensions.

5.1.1 Goals and guiding principles

The idea of developing a new framework for MWE acquisition originated from a real research need during previous experiments on automatic multiword terminology extraction from specialised corpora (Ramisch 2009). On that occasion, we realised that, in spite of the existence of a certain number of available tools for MWE extraction (see Section 3.2.3), they only dealt with part of the extraction process. For example, while `UCS`

provides several association measures for candidate ranking, the extraction of candidates from the corpus needs to be performed externally, using regular expressions or similar tools. Moreover, it only deals with 2-grams, ignoring larger *n*-grams. `NSP` provides support for larger *n*-grams, but it is impossible to describe more linguistically motivated extraction patterns based on parts of speech, lemmas or syntactic relations. For a detailed comparison of the `mwetoolkit` with other approaches, please refer to Section 5.3.

In a context where existing methods only implemented part of what we needed, our primary goal was to develop a *unified methodology* that would cover the whole acquisition pipeline. However, given that there is no consensus about the best method for a given acquisition context, [1] the new methodology should necessarily allow *multiple solutions* for a given sub-task. Thus, decisions such as the level of linguistic analysis, length *n* of the *n*-grams, filtering thresholds and evaluation measures should not be made by the method itself. Instead, given a large range of available methods, the user should be able to chose and to tune the parameters according to his/her needs. Therefore, one of our guiding principles is *generality*, that is, the relevant decisions in the acquisition should be made by the users. On the one hand, this principle implies that we cannot provide a push-button simplified methodology, but on the other hand the method can be adapted and tuned to a large number of acquisition contexts, maximising its *portability* as a consequence.

The `mwetoolkit` was originally designed to extract multiword terminology from specialised corpora, and later extended to perform automatic acquisition of several types of MWEs in specialised and general-purpose corpora. It implements hybrid knowledge-poor techniques that can be applied virtually to any corpus, independently of the domain and of the language. The main goal of `mwetoolkit` is to aid lexicographers and terminographers in the challenging task of creating language resources that include multiword entries. Therefore, we assume that, whenever a textual corpus of the target language/domain is available, it is possible to automatically acquire interesting groups of lexical units that can be regarded as candidate MWEs. We assume that the existence of targeted lists containing automatically acquired MWEs can speed up the creation and improve both quality and coverage of general-purpose and specialised lexical resources (dictionaries, thesauri) and ontologies.

Basically, we employ a quite standard sub-task definition which consists of two phases: a phase of *candidate extraction* followed by a phase of *candidate filtering*, where we combine association measures (AMs), descriptive and contrastive features and machine learning. In the first phase, one acquires candidates based either on flat *n*-grams or specific morphosyntactic patterns (of surface forms, lemmas, POS tags and dependency relations). Once the candidate lists are extracted, it is possible to filter them by defining criteria that range from simple count-based thresholds, to more complex features such as AMs. Since AMs are based on corpus word and *n*-gram counts, the toolkit provides both a corpus indexing facility and integration with web search engines (for using the web as a corpus). Additionally, for the evaluation phase, we provide validation and annotation facilities. Finally, `mwetoolkit` also allows easy integration with a machine learning tool for the creation of supervised MWE extraction models if annotated data is available.

The `mwetoolkit` methodology was implemented as a set of independent modules [2] that handle an intermediary representation of the *corpus*, the list of MWE *patterns*, the list of MWE *candidates* and the *reference* dictionaries. Each module performs a specific task in the pipeline of MWE extraction, from the raw corpus to the filtered list of MWE

1. See Section 4.2 for a formal definition of an acquisition context.
2. These modules were implemented in Python, with parts in C for efficiency reasons.

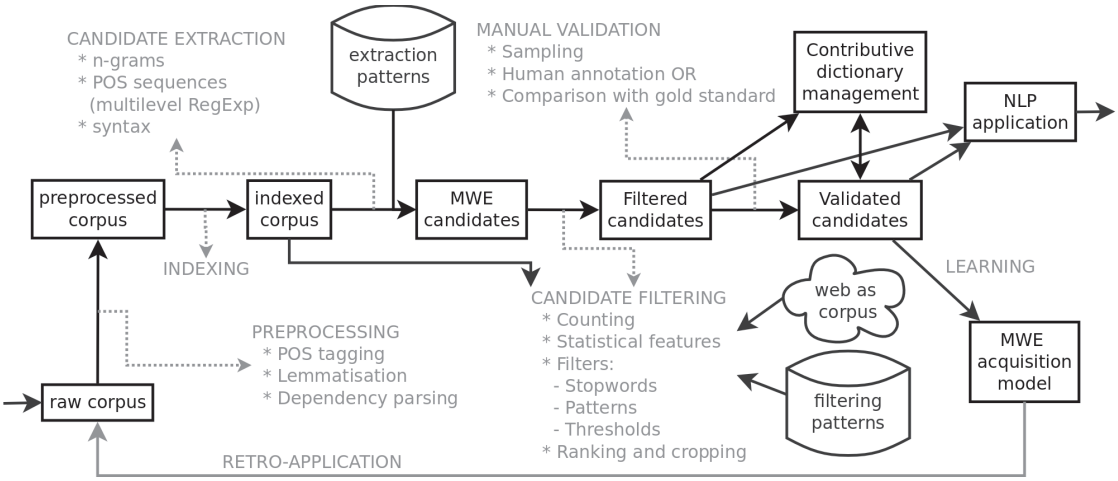


Figure 5.1: Framework for MWE acquisition from corpora, core modules in a prototypical acquisition chain.

candidates, including their automatic evaluation if a reference (gold standard) is given. Figure 5.1 summarises the architecture of `mwetoolkit`, which will be described in detail in the next section. More detailed technical documentation for the implementation is available in Appendix E.

5.1.2 Modules

The general architecture of the methodology is presented in Figure 5.1. Each module is represented as an arrow allowing to convert from one intermediary representation to another. A description of the module's functionalities is provided in light grey. In practice, except for the indexed corpus, each intermediary representation (rectangle) corresponds to and XML file containing the information generated by the modules applied to the preceding file. The figure represents only the core modules and their typical use, but there are countless other ways to combine the modules. In a typical acquisition chain, the input is a raw text corpus which is representative of the language, genre and domain of the target constructions. The core modules are described in detail below.

5.1.2.1 Preproccessing

– *Inputs:* raw textual monolingual corpus
– *Outputs:* preprocessed corpus

Preprocessing is actually not a module of the `mwetoolkit` methodology. It should be performed by external tools such as parsers and POS taggers. Preprocessing with external tools includes:

1. consistent tokenisation
2. lemmatisation
3. POS tagging
4. dependency parsing

The last three steps are optional, but higher level analysis can be crucial for determining the quality of the acquired MWEs.

Additionally, case homogenisation can be performed through `mwetoolkit`'s heuristic lowercasing rules, that tend to preserve the case of words that occur with different

capitalisation throughout the corpus. However, POS taggers and parsers often already perform adequate case processing. Lowercasing all words may not be a good idea because valuable information can be lost. For example, one should keep capitalisation of chemical component names (*NaCl*), person names (*Bill Gates* and *bill gates* are not the same entity) and acronyms (*SOAP* is not *soap*, *US* is a country, *us* is a pronoun).

Syntax information must be represented as dependency trees. Traditional constituent parsing trees must be somehow converted into dependency trees for used with the `mwetoolkit`. Our data format allows the definition of attributes by token. Thus, a dependency relation can be represented as a pair ⟨parent, type⟩ where the first element is the position of the token on which the current token depends and the second element is the type of relation (e.g., object, subject, modifier, determiner).

The preprocessed corpus should be converted from the format used by the preprocessing tool to XML. The XML file contains a sequence of sentences, and each sentence is composed of a sequence of tokens. Each token has the following optional attributes: surface form, lemma, part of speech and dependency relation. An example of XML file is provided in Figure E.1, in Appendix E.7. The `mwetoolkit` provides useful scripts to easily convert the output formats of the TreeTagger and of the RASP parser to XML. A practical tutorial on creating XML files from raw corpora using the TreeTagger and the RASP parser is included in Appendix E.5 and Appendix E.6. For more details on preprocessing in general, please refer to Section 3.1.1, where we show the application of the TreeTagger and of the RASP parser on an example sentence.

5.1.2.2 Indexing

– *Inputs:* preprocessed corpus
– *Outputs:* indexed corpus

Processing a large corpus through its XML representation is far from fast. Even with the use of a minimalist library for XML parsing like SAX, the time taken to load the file and parse its elements and attributes implies in a prohibitive overhead. Therefore, the first operation performed by the `mwetoolkit` is the creation of an index based on suffix arrays (see Section 3.3.3). A suffix array is a memory-efficient data structure that allows for the counts of *n*-grams of arbitrary length to be accessed quickly in very large corpora. For each attribute at the token level (surface form, lemma, POS and syntax), we generate a separate suffix array. Each suffix array is composed of three files: a vocabulary containing the mappings between strings and integers, a corpus file containing the sequence of integers and the suffix array itself, containing the suffix indices of the corpus sorted in lexicographical order.

The *n*-gram counts are later retrieved during candidate extraction and filtering. During the step of candidate extraction, we use the index corpus file to match regular expressions on integers, which is faster than character string operations. We use the index again to count the occurrences of the whole candidate sequence as well as the individual words. The older version of the index routine was a Python script that allowed a static index to be created from the XML corpus, but it was not scalable. Thus, the current implementation contains index routines rewritten in C. We assume that the index must fit in main memory, but the current routines provide faster indexing with reasonable memory consumption, proportional to the corpus size. For instance, with the C index routines, indexing the BNC corpus (100 million words) took about 5 minutes per attribute on a 3GB RAM computer.

One of the advantages of performing candidate extraction and *n*-gram counting as independent steps is that, in addition to the original corpus, we can obtain counts from

```
<pat id="1">
   <pat repeat="?"><w pos="DT"/></pat>
   <pat repeat="*"><w pos="J"/></pat>
   <pat repeat="+"><w pos="N"/></pat>
</pat>
<pat id="2">
   <w pos="N" id="@noun1"/>
   <w pos="P"/>
   <backw lemma="@noun1" pos="@noun1"/>
</pat>
```

Figure 5.2: Pattern 1 matches noun phrases of the form DT? J* N+, pattern 2 matches sequences N_1 P N_1.

other corpora as well. In other words, it is possible not only to use counts coming from the original corpus, as in traditional MWE acquisition, but also to count the *n*-grams in other sources like smaller domain corpora and the Web as a corpus. The `mwetoolkit` provides full integration with Yahoo!'s API [3] and with Google's API [4]. Both search engines provide page hit counts that allow us to see the web as a huge corpus, thus offering an alternative solution to overcome data sparseness (Ramisch et al. 2010d). Since web queries can be quite time-consuming, we keep a cache file with recent queries, and this avoids some delay caused by redundant network requests.

5.1.2.3 Candidate extraction

– *Inputs:* indexed corpus, extraction patterns
– *Outputs:* MWE candidates

Once the corpus has been preprocessed and indexed, we generate a first list of candidates based either on raw *n*-grams or on morphosyntactic patterns. The former is a straightforward method to extract all possible word combinations using no linguistic analysis, and could be used as a backoff strategy when no linguistic information is available. [5]

Morphosyntactic patterns allow the definition of fine-grained morphosyntactic constraints on the extracted sequences. For example, suppose we want to extract *noun–noun* and *adjective-noun* pairs, or collocations involving the adjective *strong*, or direct objects of the verb *remove*. It is possible to define patterns containing wildcards and to extract semi-fixed expressions with intervening words, using a formalism similar to regular expressions. Such patterns are multilevel, that is, it is possible to match simultaneously one or more token attributes among surface forms, lemmas, parts of speech and syntax. Multilevel patterns are correctly handled during pattern matching, in spite of individual per-attribute indices. Some scripts may merge the individual indices on the fly, producing a combined index (e.g., for *n*-gram counting).

We support all the operators shown in Figure 5.2 plus repetition interval (`{2,3}`), multiple choice (`either`) and in-word wildcards like *writ** matching *written* and *writing*. We use a notation derived from standard Python regular expressions.

– Repetition of items: an arbitrary number of times (`*`), once or more (`+`), between `a`

3. `http://developer.yahoo.com/download/download.html` — In 2012, Yahoo! announced that this service would no longer be supported.

4. `http://code.google.com/apis/ajaxsearch/`

5. If tools like a POS tagger are not available for a given language and/or domain, it is possible to generate simple *n*-gram lists, but the quality will most probably be poor. In this case, a relatively cheap workaround would be to filter out candidates on a keyword basis, for example, from a list of stopwords.

and `b` times (`{a,b}`), at least `a` times (`{a,}`), at most `b` times (`{,b}`)
- Optional items (?) and multiple choice (`either`)
- Backreference to previously matched words (`backw`)
- Wildcard words and wildcard word attributes

In Figure 5.2, the use of the two first items is illustrated in pattern 1 while the third item is illustrated in pattern 2. Pattern 1 returns noun phrases, such as *the duck*, *children* and *the big green apple tree*, pattern 2 returns a pair of equal nouns linked by a preposition, such as *hand in hand*, *word for word* and *little by little*.

The optimal set of patterns for a given domain, language and MWE type can be defined based on several factors. First, it is possible to define patterns based on linguistic intuition and/or expert knowledge about the target MWE type. Second, it is possible to perform empirical observation of some positive and negative examples in order to match only the positive ones. Finally, a combination of these two steps is often required. Initial intuition can be validated by performing a first extraction step and an evaluation of a sample of extracted candidates. Then, the patterns can be improved and a second round of candidate extraction is performed. The process is repeated until a good trade-off is obtained.

Technical documentation and a manual showing how to define morphosyntactic patterns can be found in Appendix E.3.

5.1.2.4 Candidate filtering

- *Inputs:* MWE candidates
- *Outputs:* filtered MWE candidates

In a first step, the initial candidate list can be filtered in order to exclude candidates that contain spurious punctuation, *n*-grams occurring less frequently than a given threshold or specific words and POS. This first filtering step contains mostly heuristics that will help cleaning the data. For example, statistics calculated on events occurring only once are unreliable, thus they can be excluded from the candidate list.

In a second step, each candidate is enriched with a set of *features*. These features can represent any information that helps distinguish true MWEs from random word combinations that were accidentally captured by the morphosyntactic patterns. Features can be used either directly for setting threshold values or indirectly through the application of machine learning models. In the current implementation, we provide four kinds of features: descriptive features, association measures, contrastive measures and variation entropy.

Descriptive features are simply a structured representation of the properties of the MWE candidate itself. Examples of descriptive features are: length of *n*-gram, sequence of POS tags, presence of dashes/slashes and capitalisation. Many other descriptive features can be added according to the type of target expression. Even if these features are not directly interpretable, their presence may correlate with the class of the candidate (true MWE or random word combination). Thus, a machine learning algorithm could use them to build an automatic MWE classifier.

The *association measures* estimate the degree of independence between the counts of the MWE candidate and the counts of the individual words that compose it. The following AMs are available:

- `mle`: relative frequency, that is, the *n*-gram count divided by the number of tokens in the corpus, as defined in Section 3.1.2;
- `t-score`: score derived from Student's *t* test, as defined in Equation 3.8;

– `pmi`: pointwise mutual information, as defined in Equation 3.9;
– `dice`: Dice's similarity coefficient, as defined in Equation 3.10;
– `ll`: log-likelihood measure, based on contingency tables, as defined in Equation 3.13 (applicable only to 2-grams).

Contrastive measures estimate how specialised a MWT candidate is with respect to its occurrences in general-purpose texts (the *contrast* corpus). The main difference between AMs and contrastive measures is that the former are designed for general MWE identification whereas the latter aim at automatic term recognition. The measure implemented in the `mwetoolkit` is inspired by the `CSmw` measure proposed by Bonin et al. (2010a;b), only we simplify the original function into a rank-equivalent variant:

$$\texttt{simple-csmw} = \log_2 c(w_1^n) \times \frac{c(w_1^n)}{c_{contrast}(w_1^n)} \tag{5.1}$$

We denote as $c_{contrast}(w_1^n)$ the number of occurrences of the MWT candidate in a contrastive frequency source. In a typical configuration, $c(\cdot)$ is the count in the original specialised corpus while $c_{contrast}(\cdot)$ corresponds to the count in a larger general-purpose corpus.

Finally, *variation entropy* estimates the degree of syntactic and/or semantic variability of the candidate. In order to calculate variation entropy, it is necessary to first artificially generate variations of the original *n*-gram. In order to create syntactic variations, it is possible to generate simple permutations by randomly changing word order (Zhang et al. 2006, Villavicencio et al. 2007), or alternatively to use knowledge about the syntactic behaviour of the target constructions in order to generate syntactically valid permutations (Ramisch et al. 2008a). In order to create semantic variations, it is possible to replace parts of the candidate by a semantically equivalent word, for instance a Wordnet synonym (Pearce 2001, Ramisch et al. 2008b) or a word in the same lexical field (Duran et al. 2011).

Once we have gathered a set of variations $\{v_1, \ldots, v_m\}$ for a given candidate w_1^n, we obtain their counts in a corpus. The sum of all the counts over the variations is denoted as $M = \sum_{i=1}^{m} c(v_i)$ and then we the entropy is computed as follows:

$$H(w_1^n) = -\sum_{i=1}^{m} \frac{c(v_i)}{M} \log \frac{c(v_i)}{M} \tag{5.2}$$

The interpretation of variation entropy is as follows: high values close to the maximum $\log m$ indicate a homogeneous distribution, that is, variations are roughly equiprobable, while lower values closer to zero show that one of the variations is much more likely than the others, showing a pronounced preference for that variation. In other words, low values indicate less variability (syntactic or semantic, depending on the type of variation employed), and flexibility is related to low productivity, which is one of the characteristics of MWEs (see Section 2.3).

5.1.2.5 Manual validation

– *Inputs:* filtered MWE candidates
– *Outputs:* validated MWE candidates

The output of candidate filtering is still a rough stone: a set of MWE candidates that require further validation. However, the toolkit can provide a straightforward starting point for lexicographic work, speeding up the construction of language resources, especially for poorly resourced languages and domains.

Numerical features can be used as sorting keys, in order to rank a list of candidates. Thus, given a single feature, it is possible to filter out all candidates whose feature value is below a given threshold. As there is no systematic way of deciding which features should be used to rank candidates, this is often performed on a trial-and-error basis.

Depending on the acquisition context, a list of filtered MWE candidates can have different uses. The more traditional method is the manual validation of candidates by an expert human lexicographer, thus generating a list of validated MWEs directly. This manual annotation task can be defined, in a first moment, as a binary classification task whose goal is to separate interesting and genuine target MWEs from random word combinations. The interesting MWEs will further be included in a lexical resource, which can be a paper or machine-readable dictionary, thesaurus or ontology. In addition to the validation, the expert lexicographer may want to use the features to annotate the distributional characteristics of the MWE. The evaluation of MWEs acquired automatically by expert lexicographers is explored in Chapter 6.

An alternative to direct annotation by a single lexicographer is the use of a software platform for the collaborative creation and management of lexical resources. An example of such platform is the Jibiki system (Mangeot and Chalvin 2006). Collaborative annotation allows for multiple users to create the dictionary as a joint collective effort, thus speeding up and optimising the process. The system can incrementally and interactively generate a list of validated MWEs, possibly enriched with further lexical information. This corresponds to the bidirectional arrow on Figure 5.1.

Finally, the resulting list of validated MWEs can be used by a real NLP application. Examples of such applications are described in Section 3.3.4. In Chapter 7, we describe the integration of MWEs into a statistical machine translation system. Depending on the target application, the phase of manual validation can be skipped. This means that, despite potential noise in the candidate list, the target application can appropriately use the automatically acquired MWEs, particularly when quantity is more important than quality.

5.1.2.6 Learning and retro-application

– *Inputs:* validated MWE candidates
– *Outputs:* MWE acquisition model

If a gold standard data set is available, the toolkit can automatically annotate each candidate to indicate whether it is contained in the gold standard. Therefore, an evaluation facility is provided so that, if a (potentially limited) reference gold standard is present, the class of the candidate is automatically inferred. That is, if the candidate is contained in the reference list, it is a true MWE, otherwise we assume that it is a random word combination. [6]

The class annotation is not used to filter the lists, but only by a machine learning algorithm that builds a classifier based on the relation between the features and the MWE class of the candidate in the training set. This is particularly useful because, to date, it remains unclear which features perform best for a particular MWE type or language, and the classifier applies measures according to their efficiency in filtering the candidates.

The `mwetoolkit` package provides a conversion facility that allows the importation of a candidates list into the machine learning package WEKA [7]. Once the data set is imported into WEKA, a plethora of machine learning algorithms and models can be

6. This assumption is strong, as a candidate absent from the gold standard can simply be a newly acquired MWE. The pros and cons of automatic evaluation are discussed in Section 4.1.3.

7. `http://www.cs.waikato.ac.nz/ml/weka/`

applied, our problem being handled as a traditional classification problem in artificial intelligence. Some preliminary experiments have shown that polynomial support vector machines perform quite well in the task of automatic MWE candidate filtering (Ramisch 2009). Equally good candidates for efficient machine learning technique are logistic linear regression and artificial neural networks (Pecina 2008b).

If we have an existing machine learning model for MWE acquisition, we can perform a retro-application on a new data set. Thus, once each candidate has a set of associated features, we can apply an existing machine learning model to distinguish true and false positives based on the characteristics of another MWE data set. However, this should be performed carefully, as we cannot assume that a model that works well on a given language and domain will work well on other languages and domains. The generalisation of a model for MWE acquisition is a hypothesis that remains to be validated.

5.1.2.7 Auxiliary modules

In addition to the core modules, some auxiliary tools are available in the `mwetoolkit` software package. These include scripts for performing simple operations on XML files, like counting the number of words, lines and characters (like Unix's `wc`), keeping only the first or last *n* lines of a file (like Unix's `head` and `tail`), or sorting a list of candidates according to the numeric or lexicographic order of a given feature domain (like Unix's `sort`). Finally, some useful scripts perform the conversion of XML files into several formats including TXT, CSV, [8] ARFF, [9] UCS [10] and OWL. [11]

5.1.3 Discussion

To date, there is little agreement on whether there is a single best method for MWE acquisition, or whether a different subset of methods is better for a given MWE type. Most of recent work on MWE treatment focuses on candidate extraction from preprocessed text (Seretan 2008) and on the automatic filtering and ranking through association measures (Evert 2004, Pecina 2010), but few authors provide a whole picture of the MWE treatment pipeline. The main contribution of our methodology, rather than a revolutionary approach to MWE acquisition, is the systematic integration of the processes and tasks required for acquisition, from sophisticated corpus queries, like in CQP (Christ 1994) and Manatee (Rychlý and Smrz 2004), to candidate extraction, like in Text::NSP (Banerjee and Pedersen 2003), filtering, like in UCS (Evert 2004), and machine learning.

One of the advantages of the framework proposed here is that it models the whole acquisition process in a modular approach that can be configured in several ways, each task having multiple available alternatives. Therefore, it is highly customisable and allows for a large number of parameters to be tuned according to the target MWE types. For instance, one of the advantages of our candidate extraction step is that we separate pattern matching from *n*-gram counting. Therefore, it is possible to match the patterns in a corpus *A* and then use count information from sources *B* and *C*.

The framework proposed in the `mwetoolkit` can be used not only to speed up the work of lexicographers and terminographers in the creation of lexical resources for new domains and languages, but also to contribute to the porting of NLP systems such as ma-

8. Comma-separated values, readable by most spreadsheet software like Microsoft Excel and OpenOffice Calc.

9. Format supported by WEKA.

10. Special CSV format supported by the UCS toolkit.

11. Web ontology language, standard format in the web semantic community.

chine translation and information extraction across languages and domains. The methodology employed in the toolkit is not based on symbolic knowledge or pre-existing dictionaries, and the techniques that are incorporated in it are language independent. Moreover, the techniques that we have developed do not depend on a fixed length of candidate expression nor on adjacency assumptions, as the words in an expression might occur several words away. Thanks to this generality, this methodology can be applied to virtually any language, MWE type and domain, not strictly depending on a given formalism or tool. [12] Intuitively, for a given language, if some preprocessing tools like POS taggers and/or parsers are available, the results will be much better than running the methods on raw text. But since such tools are not available for all languages, the methodology was designed to be applicable even in the absence of preprocessing.

In sum, the `mwetoolkit` methodology allows users to perform systematic MWE acquisition with consistent intermediary files and well defined modules and arguments (avoiding the need for a series of ad hoc separate processes). Even if some basic knowledge about how to run Python scripts and how to pass arguments to the command line is necessary, the user is not required to be a computer programmer.

We believe that there is room for improvement at several points of the `mwetoolkit` acquisition methodology. Nested MWEs are a problem in the current approach. For example, if the two 2-grams *International Cooperation* and *Cooperation Agreement* were acquired, both would be evaluated separately. However, they could be considered as parts of a larger MWE *International Cooperation Agreement*. If the reference dictionary only contains the larger expression, the shorter sub-expressions will count as negative results even though they are part of a MWE. With the current methodology, it is not possible to detect this kind of situation. Another problematic case would be the inverse case, that is, the candidate contains a MWE, like in the example *pro-human right*. In this case, it would be necessary to separate the prefix from the MWE, that is, to re-tokenise the words around the MWE candidate. In the case of multiple overlapping candidates matching a pattern, the current strategy returns all possibilities.

We expect, in the future, to integrate a higher number of features of the MWE candidates into the classifiers. Other features that could potentially improve classification results are new descriptive features, deep syntax, semantic classes, semantic relations, domain-specific keywords, context-based measures and context words. In addition, we would like to integrate information coming from peripheral sources such as parallel corpora (word alignments) and general-purpose or domain-specific simple word dictionaries. While for poor-resourced languages we can only count on shallow linguistic information, it is unreasonable to ignore available information for other languages like English, Spanish, French and German.

Related work showed that association measures based on contingency tables are more robust to data sparseness (Evert and Krenn 2005). However, they are based on pairwise comparisons and their application on arbitrarily long n-grams is not straightforward. A heuristic to adapt these measures consists in applying them recursively over increasing n-gram lengths. In the future, we would like to test several heuristics to handle nested candidates and longer n-grams.

Moreover, we would like to provide better integration between the candidate extraction step and the classifier construction step. Currently, the latter is performed externally using WEKA, but we believe that if this step were integrated into the toolkit's pipeline,

12. However, it is designed to deal with languages that use spaces to separate words. Thus, when working with Chinese, Japanese, or even with German compounds, some additional preprocessing is required.

we would increase its ease of use. Still under the perspective of usability, we would like to develop or adapt an interface for manual evaluation of the candidates and for testing the results in the context of lexical resources construction.

One of our goals for future versions is to be able to automatically extract bilingual MWEs from parallel or comparable corpora. This could be done through the inclusion of automatic word alignment information. Some previous experiments show, however, that this may not be enough, as automatic word alignment uses almost no linguistic information and its output is often quite noisy (Ramisch et al. 2010a). Combining alignment and linguistic information seems a promising solution for the automatic extraction of bilingual MWEs. Another method that we would like to explore is the generation of compositional translations to be validated against corpora evidence. The potential uses of bilingual MWE lexicons are multiple, but the most obvious applications are machine translation and multilingual technical writing. On the one hand, MWEs could be used to guide the word alignment process. For instance, this could solve the problem of aligning a language having a writing system where compounds are made of separate words, like French, with a language that joins compound words together, like German. In statistical machine translation systems, MWEs could help to filter phrase tables or to boost the scores of phrases whose words are likely to be multiwords.

We would like to evaluate our method on several data sets, varying the languages, domains and target MWE types. This extensive evaluation could allow the development of standard machine learning models for MWE acquisition in different domains. Thus, we would be able to compare the similarities and differences between domains based on the models that are created for them. Additionally, we could evaluate how well the classifiers perform across languages and domains.

The `mwetoolkit` is an important first step toward robust and reliable MWE treatment by NLP applications. It is a freely available core application providing powerful tools and coherent up-to-date documentation, and these are essential characteristics for the extension and support of any computer tool. Thus, we would like to keep making periodical releases of a stable software version. Therefore, we would need extensive testing and constant documentation update.

5.2 A case study

In the present section, we describe a step-by-step example of MWE acquisition from a corpus. In the following toy experiment, we used the `mwetoolkit` to extract *multiword terms* (MWTs) from the Genia corpus (described in Appendix D) In order to train machine learning models and test them, the original corpus was divided into a training set and a test set, with the latter containing 895 sentences ($\approx 4.9\%$ of the corpus), and the former containing all other sentences (17,543).

5.2.1 Candidate extraction

In order to unify the spelling of the words throughout the corpus, we preprocessed it uniformly according to the following criteria:

- Capitalised words were lowercased using the heuristics described in Section 5.1.2.1.
- POS tags were simplified to match a set of patterns (e.g., NN, NNS, NP... $\rightarrow$ N)
- Words containing dashes and slashes were retokenised, as these symbols are not used consistently in the Genia corpus (e.g., *T cell* and *T-cell*). Therefore, any word that contained these symbols was split into independent subparts as the symbols

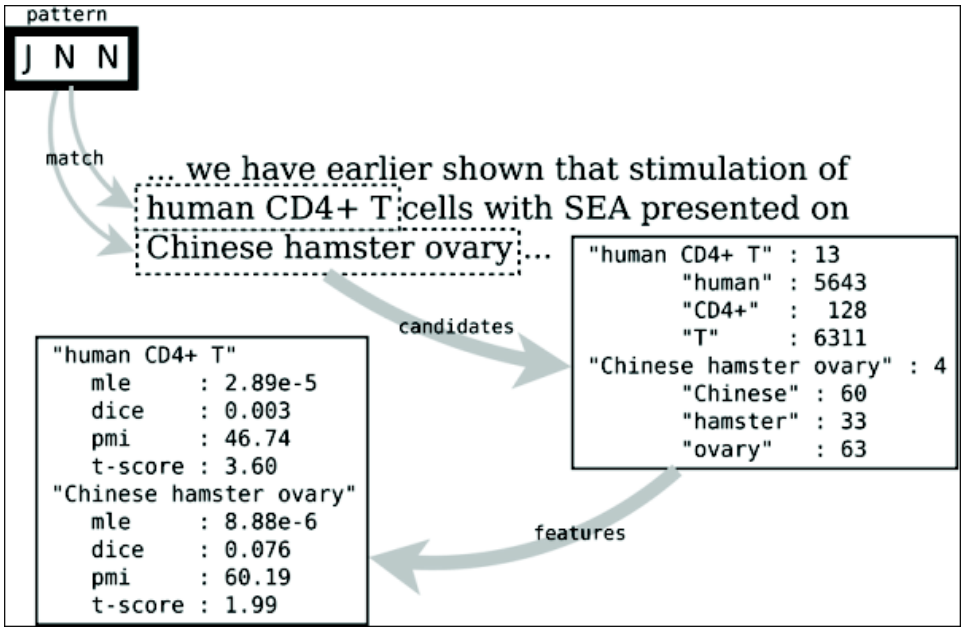


Figure 5.3: Example of MWT candidates extracted from the Genia corpus.

were removed (e.g., *T-cell* becomes *T cell*). [13]

– Acronyms were recognised and removed when they occurred between parentheses (e.g., *human immunodeficiency virus (HIV) type 1* was changed to *human immunodeficiency virus type 1*).
– Nouns were lemmatised to their singular form.

These preprocessing steps aim to reduce the problem of data sparseness, which is particularly acute for MWEs and specific domains, and they have a significant impact on the quality of the results. We estimate, for instance, that precision and recall are reduced by more than 50% if the lemmatisation and retokenisation steps are not performed. This happens because the reference dictionary only contains canonical forms and because the counts of lemmatised words are less sparse than those of inflected ones.

In this experiment, we used a set of 57 morphosyntactic patterns based on the POS sequences defined by Justeson and Katz (1995). Their original set of patterns was augmented through the use of a heuristic that enables the extraction of longer sequences of contiguous nouns and adjectives than originally defined. For instance, using these patterns, it is possible to extract candidates that match POS patterns containing sequences of two to seven adjacent nouns and adjectives (e.g., *T cell*, *thromboxane receptor gene*), foreign words (e.g., *in vitro*) and numbers (e.g., *nucleotide 46*).

From the Genia corpus sentence shown in Figure 5.3, we selected two candidates that match the sequence adjective-noun-noun (J N N): *human CD4+ T* and *Chinese hamster ovary*. The former, although part of a longer MWT in this sentence (*human CD4+ T cell*) is a false positive if seen as a 3-gram. [14]

13. In the future, we would like to apply existing techniques to unify the spelling of words around dashes and slashes.

14. This sentence exemplifies the problem that arises from ignoring nested MWTs. Here, each part of a MWT is treated independently from any other part. As the original MWT (*human CD4+ T cell*) matches different POS patterns, it forms 3 different candidates which are treated independently: *human CD4+ T* (J N N), *CD4+ T cell* (N N N) and *human CD4+ T cell* (J N N N).

	$t = 0$	$t = 1$	$t = 5$
# cand	763	739	174
# ref	2,009	2,009	2,009
# TP	401	420	129
P	52.56%	56.83%	74.14%
R	19.96%	20.91%	6.42%
F	28.93%	30.57%	11.82%

Table 5.1: Performance of the `mwetoolkit` considering (a) no filtering threshold, (b) a threshold of $t = 1$ occurrence and (c) a threshold of $t = 5$ occurrences.

5.2.2 Candidate filtering

This initial list of candidates can be further validated using some criteria, in order to, insofar as possible, remove false positives from the list, and only keep genuine MWTs. This validation is done using a set of AMs as basis for building a classifier. In order to calculate the AMs for each candidate, the `mwetoolkit` determines the corpus counts for the candidate as well as for the individual words that compose it. In Figure 5.3, the *n*-gram and word counts of the Genia corpus are represented.

After obtaining the corpus counts, the toolkit uses this information as input to the formulae that calculate four association scores for each candidate in each corpus (the `ll` measure was not used because it can only be applied to 2-grams). All AMs are used as features for the classifier and it then decides on the best feature combination to use in order to choose whether a candidate should be kept in the list or be discarded as noise.

Figure 5.4 shows an example of XML representation obtained for one of our example candidates extracted from the Genia corpus: *Chinese hamster ovary*. For each individual word and for the whole candidate, the `freq` elements show their corpus counts in two different corpora: Genia (as `genia`) and Yahoo! (as `yahoo`). The idea is to use two heterogeneous data sources so that we do not loose in accuracy because of the sparseness of the former or because of the rough approximations done by the latter. The first two features are descriptive properties of the candidate such as the number of words and the POS sequence and the remainder of the features correspond to the AMs in the Genia corpus and in Yahoo!. After the list of features, the special element `tpclass` indicates the class of the candidate with respect to the reference list. This information, when available, can be used to build a new classifier for a given language or domain. In our toy experiment, its utility is two-fold: (1) on the training corpus, it is used as class information for a supervised learning algorithm that will build our MWT classifier; (2) in the test corpus, it determines whether a candidate is correctly classified as a true positive (or as a true negative), helping us evaluate the performance of the `mwetoolkit`.

5.2.3 Results

We evaluate the performance of the MWT identification in terms of precision (P), recall (R) and F-measure, using the Genia ontology as MWT gold standard (see Section 4.1). The Genia ontology is a manually-built resource that contains, among other information, the set of terms found in the Genia corpus (Kim et al. 2006). For a given portion of the Genia corpus, the MWT *reference list* is composed of the multiword entries of the Genia ontology that occur in that portion of the corpus.

```
<cand candid="4582">
    <ngram>
        <w lemma="Chinese" pos="A" >
            <freq name="genia" value="60" />
            <freq name="yahoo" value="1460000000" />
        </w>
        <w lemma="hamster" pos="N" >
            <freq name="genia" value="33" />
            <freq name="yahoo" value="42600000" />
        </w>
        <w lemma="ovary" pos="N" >
            <freq name="genia" value="63" />
            <freq name="yahoo" value="12300000" />
        </w>
        <freq name="genia" value="4" />
        <freq name="yahoo" value="723000" />
    </ngram>
    <occurs>
        <ngram>
            <w surface="Chinese" pos="A" />
            <w surface="hamster" pos="N" />
            <w surface="ovary" pos="N" />
            <freq name="corpus" value="4" />
        </ngram>
    </occurs>
    <features>
        <feat name="pos_pattern" value="A#S#N#S#N" />
        <feat name="n" value="3" />
        <feat name="mle_genia" value="8.8833220071e-06" />
        <feat name="pmi_genia" value="60.193312488" />
        <feat name="t_genia" value="1.99999969239" />
        <feat name="dice_genia" value="0.0769230769231" />
        <feat name="mle_yahoo" value="1.31454545455e-05" />
        <feat name="pmi_yahoo" value="82.8386600941" />
        <feat name="t_yahoo" value="849.996644814" />
        <feat name="dice_yahoo" value="0.00143177767509" />
    </features>
    <tpclass name="genia-reference" value="True" />
</cand>
```

Figure 5.4: XML fragment describing a MWT candidate extracted from the Genia corpus with `mwetoolkit`.

The candidates were fed into a learning algorithm that produced a *support vector machine* (SVM) classifier. In previous experiments performed, among all tested machine learning models, SVM with polynomial kernel presented the best balance between precision and recall (Ramisch 2009). We applied this model to the test corpus (the remaining unannotated 895 sentences of the Genia corpus) and evaluated the output in terms of precision and recall.

Table 5.1 shows three different filtering configurations applied (both during training and testing) to the candidates extracted by the `mwetoolkit` from the test portion of the Genia corpus. In the first condition, we considered all candidates without any frequency threshold. In the second, we considered all the candidates which occurred more than once in the test corpus, while in the third, we kept all the candidates that occurred at least five times. The results show us that, as expected, statistical AMs calculated including candidates that occur only once are not reliable ($t = 0$), and discarding them helps to improve precision and recall ($t = 1$). A higher threshold like $t = 5$ provides even better precision at the price of drastically reducing recall, but even so recall and F-measure in this configuration are still higher than those of the baseline systems with which we compared the `mwetoolkit`.

For a given application, the exact value of the threshold can be customised according to whether the preference is for a higher recall or for a higher precision. For instance, if the goal is to create a terminological dictionary, a higher recall may be desirable with manual validation of the results. The `mwetoolkit` allows parametrisation and customisation of its various modules according to a particular application without being language- or domain-dependent. Therefore, its performance could be improved even further with better tuning to the domain or postprocessing of the results.

A detailed tutorial explaining the application of the scripts, the parameters and intermediary files can be found in Appendix E.3.

5.3 Comparison with related approaches

In this section, we compare the `mwetoolkit` methodology, presented in the previous sections, with three other similar approaches. We consider only freely available, downloadable and openly documented tools. Therefore, outside the scope of this comparison are proprietary tools, terminology and lexicography tools, translation aid tools and published techniques for which no available implementation is provided. The experimental setup used in our comparison is presented in Section 5.3.2. In Section 5.3.3, we evaluate the following acquisition dimensions: quality of extracted candidates and of association measures, use of computational resources and flexibility. Thus, this comparative investigation indicates the best cost-benefit ratio in a given context (language, type, corpus size).

5.3.1 Related approaches

The goal of this section is to compare the `mwetoolkit` methodology with other approaches for the automatic acquisition of MWEs from corpora, examining as parameters of the experimental context: the language (English and French), the type of target MWE (verbal and nominal) and the size of corpus (small, medium, large).

We focus our comparative evaluation on MWE acquisition methods that follow the general trend in the area of using shallow linguistic (lemmas, POS, stopwords) and/or statistical (counts, AMs) information to distinguishing ordinary sequences (e.g., *yellow*

	Small	**M**edium	**L**arge
# sentences	5,000	50,000	500,000
# `en` words	133,859	1,355,482	13,164,654
# `fr` words	145,888	1,483,428	14,584,617

Table 5.2: Number of sentences and of words of each fragment of the Europarl corpus in `fr` and in `en`.

dress, *go to a concert*) from MWEs (e.g., *black box*, *go by a name*). Our evaluation compares the `mwetoolkit` with the three first approaches described in Section 3.2.3.1, namely:

- the LocalMaxs reference implementation (`LocMax`);
- the *N*-gram statistics package (`NSP`); and
- the `UCS` toolkit.

In addition to the brief description provided in Section 3.2.3.1, Section 5.3.3.4 underlines the main differences between the `mwetoolkit` and these approaches.

As the focus of our comparison is on MWE acquisition, other tasks related to MWE treatment are not considered in this thesis. This is the case, for instance, of approaches for dictionary-based in-context MWE token identification requiring an initial dictionary of valid MWEs, like jMWE.

5.3.2 Experimental setup

We investigate the acquisition of MWEs in two languages, English (`en`) and French (`fr`), analysing nominal and verbal expressions in English and nominal expressions in French. As French does not present many verb-particle constructions and due to the lack of availability of resource for other types of French verbal expressions (e.g., light verb constructions), only nominal expressions are considered. The candidate MWEs were obtained through the following patterns:

- **Nominal expressions** `en`: a noun preceded by a sequence of one or more nouns or adjectives (e.g., *European Union*, *clock radio*, *clown anemone fish*).
- **Nominal expressions** `fr`: a noun followed by either an adjective or a prepositional complement (with the prepositions *de*, *à* and *en*) followed by an optionally determined noun (e.g., *algue verte*, *aliénation de biens*, *allergie à la poussière*).
- **Verbal expressions** `en`: verb-particle constructions formed by a verb (except *be* and *have*) followed by a prepositional particle [15] not further than 5 words after it, [16] (e.g., *give up*, *switch the old computer off*).

To test the influence of corpus size on performance, three fragments of the English and French parts of the Europarl corpus v3 (described in Appendix D), were used as test corpora: (**S**)mall, (**M**)edium and (**L**)arge, summarised in Table 5.2.

The extracted MWEs were automatically evaluated against the following gold standards: WordNet 3, the Cambridge Dictionary of Phrasal Verbs, and the VPC (Baldwin

15. *up, off, down, back, away, in, on.*

16. In theory, a particle could occur further than 5 positions away, like in the example **take** *patient risk factors and convenience* **into account** (googled on May 6, 2012). However, such cases are rare and, for verb-particle constructions, empirical studies showed that the longest noun phrase separating a verb from a particle contains 3 words (Baldwin 2005b).

type	lang.	# entries			
		total	S	M	L
Nominal	en	59,683	122	764	2,710
Nominal	fr	69,118	220	1,406	4,747
Verbal	en	1,846	699	1,846	1,846

Table 5.3: Dimensions of the reference gold standards used and of the respective number of entries that occur at least twice in the S, M and L corpora.

2008) and CN (Kim and Baldwin 2008) datasets [17] for en; the Lexique-Grammaire [18] for fr. The total number of entries is listed in Table 5.3, along with the number of entries occurring at least twice in each corpus, which was the denominator used to calculate recall in Section 5.3.3.1.

5.3.3 Comparison results

We performed MWE acquisition using four tools: mwetoolkit, LocMax, NSP and UCS. We includes both versions of LocMax: *LocalMaxs Strict*, which gives priority to high precision (henceforth LocMax-S), and *LocalMaxs Relaxed* which focuses on high recall (henceforth LocMax-R). As approaches differ in the way they allow the description of extraction criteria, we present the results of candidate extraction (Section 5.3.3.1) separately from the results of AMs (Section 5.3.3.2). Additionally, we go beyond traditional evaluation by presenting the trade-off between the usefulness of the acquired MWEs and the computational resources used (Section 5.3.3.3). We close this section with a discussion about the flexibility of the techniques in each extraction context (Section 5.3.3.4).

5.3.3.1 Extracted candidates

We consider as *MWE candidates* the initial set of sequences before any AM is applied. Candidate extraction is performed through the application of patterns describing the target MWEs in terms of POS sequences, as described in Section 5.3.2.

The quality of candidates extracted from the medium-size corpus (M) varies across MWE types/languages, as shown in Figure 5.5. UCS is unable to process candidates longer than 2 words. Therefore, the candidates for UCS are obtained by keeping only the 2-grams in the candidate list returned by the mwetoolkit. For nominal MWEs, the approaches have similar patterns of performance in the two languages, with high recall and low precision yielding an F-measure of around 10 to 15%. The variation between en and fr can be partly explained by the differences in size of the gold standards for each of these languages. Further research would be needed to determine to what degree the characteristics of these languages and the set of extraction patterns influence these results. For verbal expressions, LocMax has high precision (around 70%) but low recall while the other approaches have more balanced P and R values around 20%. This is partly due to the need for simulating POS filters for extraction of verbal MWE candidates with LocMax. The filter consists of keeping only contiguous *n*-grams in which the first and the last words matched verb+particle pattern and removing intervening words.

The techniques differ in terms of extraction strategy: (i) mwetoolkit and NSP allow

17. The latter are available from http://multiword.sf.net/
18. http://infolingu.univ-mlv.fr/

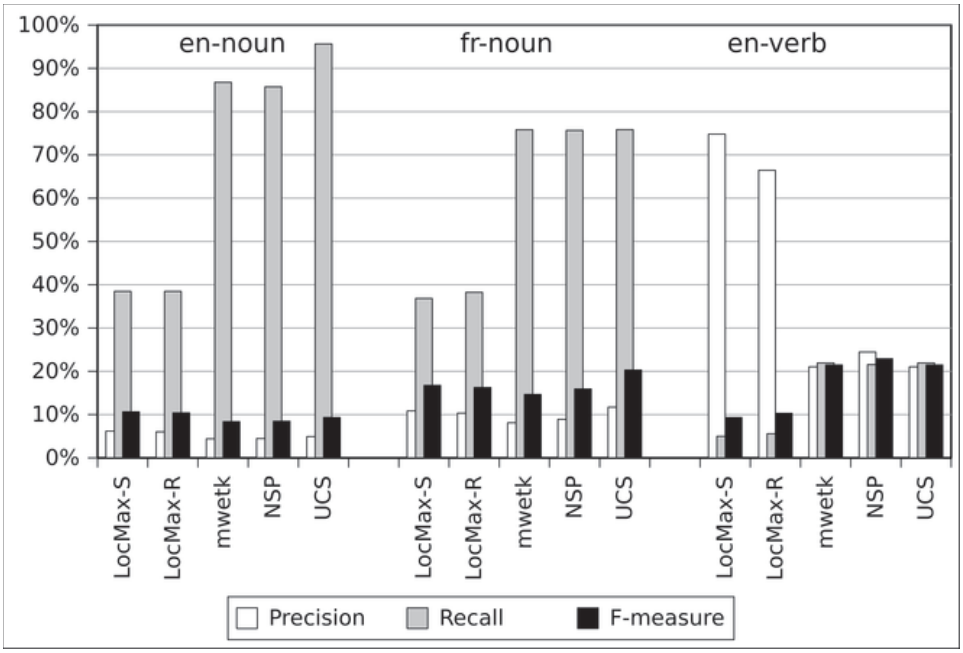


Figure 5.5: Quality of candidates extracted from medium corpus, comparison across languages/MWE types.

	LocMax-S		LocMax-R		mwetoolkit		NSP		UCS	
	P	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
S	7.53	42.62	7.46	42.62	6.50	83.61	6.61	83.61	6.96	96.19
M	6.18	38.48	6.02	38.48	4.40	86.78	4.46	85.73	4.91	95.65
L	4.50	37.42	—	—	2.35	89.23	2.48	89.41	2.77	96.88

Table 5.4: (P)recision and (R)ecall of en nominal candidates, comparison across corpus sizes: (S)mall, (M)edium and (L)arge.

the definition of linguistic filters while LocMax only allows the application of grep-like filters after extraction; (ii) there is no preliminary filtering in mwetoolkit and NSP, they simply return all candidates matching a pattern, while LocMax filters the candidates based on the local maxima criterion; (iii) LocMax only extracts contiguous candidates while the others allow discontiguous candidates. The way mwetoolkit and NSP extract discontiguous candidates differs: the former extracts all verbs with particles no further than 5 positions to the right. NSP extracts 2-grams in a window of 5 words, and then filters the list, keeping only those in which the first word is a verb and that contain a particle. However, the results are similar, with slightly better values for NSP.

The evaluation of en nominal candidates according to corpus size is shown in Table 5.4. [19] For all approaches, precision decreases when the corpus size increases as more false MWEs are returned, while recall increases for all except LocMax. This may be due to the latter ignoring shorter *n*-grams when longer candidates containing them become sufficiently frequent, as is the case when the corpus increases. Table 5.5 shows that the candidates extracted by LocMax are almost completely covered by the candidates extracted by the other approaches. The relaxed version extracts slightly more candidates, but still much less than mwetoolkit, NSP and UCS, which all extract a similar set of

19. It was not possible to evaluate LocMax-R on the large corpus as the provided implementation did not support corpora of this size.

	LocMax-S	LocMax-R	mwetk	NSP	UCS	**Total verbs**
LocMax-S	—	124	124	122	124	124
LocMax-R	4,747	—	156	153	156	156
mwetoolkit	4,738	4,862	—	1,565	1,926	1,926
NSP	4,756	4,879	14,611	—	1,565	1,629
UCS	4,377	4,364	13,407	13,045	—	1,926
Total nouns	4,760	4,884	15,064	14,682	13,418	

Table 5.5: Intersection of the candidate lists extracted from medium corpus. Nominal candidates `en` in bottom left, verbal candidates `en` in top right.

candidates. In order to distinguish the performance of the approaches, we need to analyse the AMs they use to rank the candidates.

5.3.3.2 Association measures

Traditionally, to evaluate an AM, the candidates are ranked according to it and a threshold value is applied, below which the candidates are discarded. However, if we take the average of precision considering all true MWEs as threshold points, we obtain the mean average precision (MAP) of the measure without setting a hard threshold (see Section 4.1.2).

Table 5.6 presents the MAP values for the tested AMs applied to the candidates extracted from the large corpus (L), where the larger the value, the better the performance. We used as baseline the assignment of a random score and the use of the raw relative frequency for the candidates. Except for `mwetoolkit`'s `t-score` and `pmi`, all MAP values are significantly different from the two baselines, with a two-tailed t test for difference of means assuming unequal sample sizes and variances (p-value < 0.005).

`LocMax`'s `glue` performs best for all types of MWEs, suggesting local maxima as a good generic MWE indicator and `glue` as an efficient AM to generate highly precise results (considering the difficulty of this task). On the other hand, this approach returns a small set of candidates and this may be problematic for some tasks (e.g., for building a wide-coverage lexicon). For `mwetoolkit`, the best overall AM is `dice`; the other measures are not consistently better than the baseline, or perform better for one MWE type than for the other. The Poisson-Stirling (`Poisson`) measure performed quite well, while the other two measures tested for `NSP` performed below baseline for some cases. Finally, the AMs applied by `UCS` perform all above baseline and, for nominal MWEs, are comparable to the best AM (e.g., `Poisson` and `local.MI`). The MAP for verbal expressions varies much for `UCS` (from 30% to 53%), but none of the measures comes close to the MAP of `glue` (87.06%).

5.3.3.3 Computational resources

In the decision of which AM to adopt, factors like the degree of MWE variability and computational performance may be taken into account. For instance, `dice` can be applied to n-grams of any length quite fast while more sophisticated measures like `Poisson` can be applied only to 2-grams and sometimes use considerable computational resources. Even if one could argue that we can be lenient towards a slow offline extraction process, the extra waiting may not be worth a slight quality improvement. Moreover, memory

	en noun	fr noun	en verb
Baseline			
rand	2.75	6.11	17.21
freq	4.75	8.79	22.72
LocMax-S			
glue	6.99	12.94	87.06
mwetoolkit			
dice	5.78	9.54	46.36
t	5.09	8.68	26.42
pmi	2.76	2.92	53.56
ll	3.17	5.52	45.88
NSP			
pmi	2.99	7.68	62.17
ps	5.40	12.38	57.62
tmi	2.108	4.89	19.80
UCS			
z.score	6.12	11.77	46.87
Poisson	6.59	12.82	32.77
MI	5.15	9.34	53.56
rel.risk	5.10	9.29	46.67
odds	5.04	9.21	50.22
gmean	6.01	11.52	45.61
local.MI	6.43	12.78	29.99

Table 5.6: Mean average precision of AMs in the large corpus.

	LocMax	mwetoolkit	NSP	UCS
Candidate extraction	Yes	Yes	Yes	No
N-grams with $n > 2$	Yes	Yes	Yes	No
Discontiguous MWE	No	Yes	Yes	—
Linguistic filter	No	Yes	No	No
Robust AMs	No	No	Yes	Yes
Large corpora	Partly	Yes	Yes	No
Availability	Free	Free	Free	Free

Table 5.7: Summary of tools for MWE acquisition.

limitations are an issue if no large computer clusters are available, like it is often the case in real lexicographic environments.

In Figure 5.6, we plotted in log-scale the time in seconds used by each approach to extract nominal and verbal expressions in en, using a dedicated 2.4GHz quad-core Linux machine with 4Gb RAM. For nominal expressions, time increases linearly with the size of the corpus, whereas for verbal expressions it seems to increase faster than the size of the corpus. UCS is the slowest approach for both MWE types while NSP and LocMax-S are the fastest. However, it is important to emphasize that NSP consumed more than 3Gb memory to extract 4- and 5-grams from the large corpus and LocMax-R could not handle the large corpus at all. In theory, all techniques can be applied to arbitrarily large corpora if we used a map-reduce approach (e.g., NSP provides tools to split and join the corpus). However, the goal of this evaluation is to discover the performance of the techniques with no manual optimisation. In this sense, mwetoolkit provides an average trade-off between quality and resources used.

5.3.3.4 Generality

Table 5.7 summarises the characteristics of the approaches. Among them, UCS does not extract candidates from corpora but takes as input a list of 2-grams and their counts.

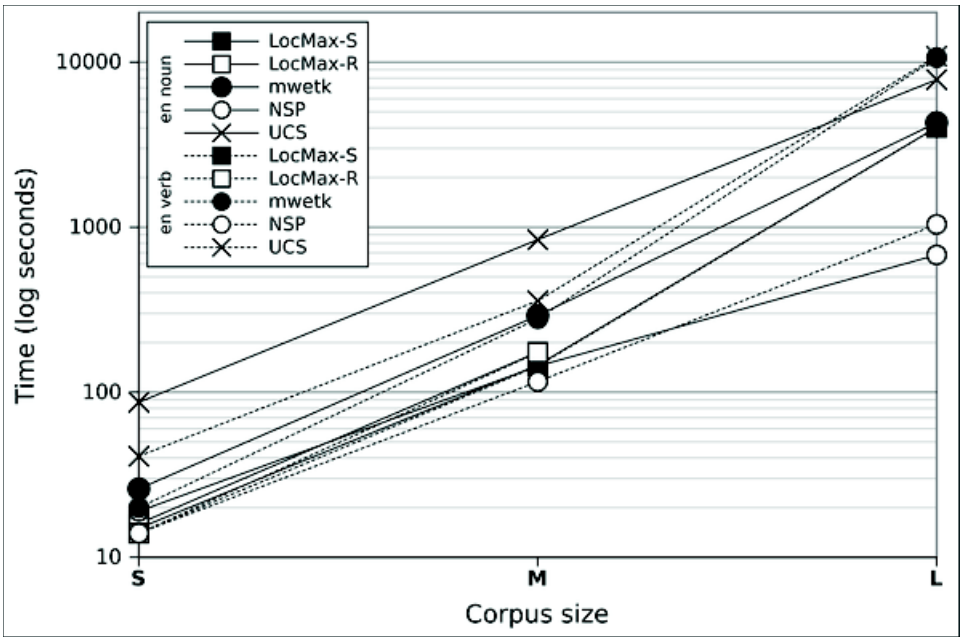


Figure 5.6: Time (seconds, log scale) to extract `en` nouns (bold line) and verbs (dashed line) from corpora.

While it only supports *n*-grams of size 2, `NSP` implements some of the AMs for 3 and 4-grams and `mwetoolkit` and `LocMax` have no constraint on the number of words. The AMs implemented by `LocMax` and `mwetoolkit` are thus less statistically sound than the clearly designed measures used by `UCS` and, to some extent, by `NSP` (Fisher test). `LocMax` extracts only contiguous MWEs while `mwetoolkit` allows the extraction of unrestrictedly distant words and `NSP` allows the specification of a window of maximum *w* ignored words between each two words of the candidate. Only `mwetoolkit` integrates linguistic filters on the lemma, POS and syntax, but this was performed using external tools (sed/grep) for the other approaches with similar results. The large corpus used in our experiments was not supported by `LocMax-R` version, but `LocMax-S` has a version that deals with large corpora, as well as `mwetoolkit` and `NSP`. Finally, all of these approaches are freely available for download and documented on the web.

5.4 Summary

We introduce a new framework called `mwetoolkit`, which integrates multiple techniques and covers the whole pipeline of MWE acquisition. One can preprocess a raw monolingual corpus, if tools are available for the target language, enriching it with POS tags, lemmas and dependency syntax. Then, based on expert linguistic knowledge, intuition, empiric observation and/or examples, one defines multilevel patterns in a formalism similar to regular expressions to describe the target MWEs. The application of these patterns on an indexed corpus generates a list of candidate MWEs. For filtering, a plethora of methods is available, ranging from simple frequency thresholds to stopword lists and sophisticated association measures. Finally, the resulting filtered candidates are either directly injected into a NLP application or further manually validated before application. An alternative use for the validated candidates is to train a machine learning model which can be applied on new corpora in order to automatically identify and extract MWEs based

on the characteristics of the previously acquired ones. This is summarised in the schema of Figure 5.1. Further details are provided on the tool website and in previous publications (Ramisch et al. 2010b;c).

To date, there is little agreement on whether there is a single best method for MWE acquisition, or whether a different subset of methods is better for a given MWE type. The main contribution of our methodology is the systematic integration of the processes and tasks required for acquisition. One of its main advantages is that it models the whole acquisition process in a modular approach, thus being customisable and allowing for a large number of parameters to be tuned. The `mwetoolkit` can be used to speed up lexicographic and terminographic work and contribute to the porting of NLP systems across languages and domains. The methodology employed in the toolkit is not based on symbolic knowledge or pre-existing dictionaries, and the techniques are language independent. Moreover, they do not depend on fixed candidate length nor adjacency. Thanks to this generality, this methodology can be applied to virtually any language, MWE type and domain, not strictly depending on a given formalism or tool. In sum, the `mwetoolkit` methodology allows users to perform systematic MWE acquisition with consistent intermediary files and well defined modules and arguments.

We compared the `mwetoolkit` methodology, with three other freely available, downloadable and openly documented tools: the LocalMaxs reference implementation (`LocMax`), the *N*-gram statistics package (`NSP`), and the `UCS` toolkit. We investigated the acquisition of MWEs in two languages, English (`en`) and French (`fr`), analysing nominal and verbal expressions in English and nominal expressions in French. The extracted MWEs were automatically evaluated against existing gold standards.

The quality of candidates extracted from the medium-size corpus (M) varies across MWE types/languages, as shown in Figure 5.5. For nominal MWEs, the approaches have similar patterns of performance, with high recall and low precision. For verbal expressions, `LocMax` has high precision (around 70%) but low recall while the other approaches have more balanced P and R values around 20%. The techniques differ in terms of extraction strategy: (i) `mwetoolkit` and `NSP` allow the definition of linguistic filters while `LocMax` only allows the application of grep-like filters after extraction; (ii) there is no preliminary filtering in `mwetoolkit` and `NSP`, they simply return all candidates matching a pattern, while `LocMax` filters the candidates based on the local maxima criterion; (iii) `LocMax` only extracts contiguous candidates while the others allow discontiguous candidates. The evaluation of `en` nominal candidates according to corpus size is shown in Table 5.4. For all approaches, precision decreases when the corpus size increases, while recall increases for all except `LocMax`.

Table 5.6 presents the evaluation of the AMs. `LocMax`'s `glue` performs best for all types of MWEs, suggesting local maxima as a good generic MWE indicator and `glue` as an efficient AM to generate highly precise results. For `mwetoolkit`, the best overall AM is `dice`; the other measures are not consistently better than the baseline. The Poisson-Stirling (`Poisson`) measure performed quite well, while the other two measures tested for `NSP` performed below baseline for some cases. Finally, the AMs applied by `UCS` perform all above baseline and, for nominal MWEs, are comparable to the best AM.

Aspects like the degree of MWE variability and computational performance influence the decision of which AM to adopt. For instance, `dice` can be easily applied to any *n*-gram, while more sophisticated measures like `Poisson` are defined only for 2-grams and are sometimes computationally heavy. `UCS` does not extract candidates from cor-

pora but takes as input a list of 2-grams. NSP implements some of the AMs for 3 and 4-grams and mwetoolkit and LocMax have no constraint on the number of words. LocMax extracts only contiguous MWEs while mwetoolkit and NSP allow the extraction of non-adjacent words. Only mwetoolkit integrates linguistic filters on the lemma, POS and syntax. This can be performed using external tools (sed/grep) for the other approaches.

The mwetoolkit is an important first step toward robust and reliable MWE treatment by NLP applications. It is a freely available core application providing powerful tools and coherent up-to-date documentation. These are essential characteristics for the extension and support of any computational tool.

Part III

Application-oriented evaluation

6 APPLICATION 1: LEXICOGRAPHY

This chapter shows the results of the evaluation of the methodology proposed in the `mwetoolkit` for the creation of MWE dictionaries. First, we explore the creation of a dictionary containing Greek nominal expressions (Section 6.1). Second, we present the creation of two lexical resources for Brazilian Portuguese. They contain complex predicates (verbal expressions) and are aimed at two real applications: semantic role labelling and sentiment analysis (Section 6.2). These two languages were chosen because: (a) they are poorly resourced in terms of MWE lexicons, and (b) there was a real need to build MWE lexicons for a given application.

6.1 A dictionary of nominal MWEs in Greek [1]

The main goal of this section is to evaluate the effectiveness of the MWE acquisition approach proposed in Chapter 5 for the automatic construction of a MWE dictionary for Greek. We present the results of experiments carried out in order to create a dictionary of MWEs for Greek using a combination of automatic extraction and human validation. In Section 6.1.1 we discuss some related work on the construction of language resources for the Greek language. We performed extraction using the `mwetoolkit`, based on POS patterns applied to the Greek portion of the Europarl corpus (Section 6.1.2). The results obtained by AMs on the Greek Europarl corpus are compared and contrasted with those obtained by the same measures using the web as a corpus (Section 6.1.3). The manual evaluation of the results by Greek native speakers led to the creation of a lexical resource that was later made available on the MWE community website.

6.1.1 Greek nominal MWEs

In the state of the art presented in Chapter 3, the performance of techniques for the automatic acquisition of MWEs has been tested on languages like English, Spanish, French and German. As a consequence, the construction of MWE resources for these languages is picking up pace, whereas for languages like Greek, computational approaches for the automatic or semi-automatic construction of language resources are still underexploited. However, the Greek language is as rich in MWEs as main European languages. Some examples of MWEs in Greek are: κάλιο αργά παρά πατέ (*better late than ever* — idiom), πλυντήριο πιάτων (*washing machine* — noun compound), οπτική ίνά (*optical fiber* —

1. Work reported in this section was previously published in the paper *Towards the Construction of Language Resources for Greek Multiword Expressions: Extraction and Evaluation* (Linardaki et al. 2010). It was carried out with the collaboration of Evita Linardaki, Carlos Ramisch, Aline Villavicencio and Aggeliki Fotopoulou.

Greek source	Result of MT	English reference	Count
. . . , όπως αυτό ορίζεται από την ανθρώπινη οπτική γωνία	*. . . , as this is fixed by the human* ***optical corner***	*. . . , as seen from the human* ***point of view***	131
Το ξέπλυμα βρώμικου χρήματος αντιπροσωπεύει το 2 έως 5% . . .	*The* ***rinsing of dirty money*** *represents the 2 until 5% . . .*	***Money laundering*** *represents between 2 and 5% . . .*	21
Για τα εργοστάσια ατομικής ενέργειας η Ευρωπαϊκή Ένωση έχει αναλάβει δράσεις για την υψηλότερη ασφάλεια,. . .	*For the* ***factories of individual energy*** *the European Union has undertaken action for the higher safety,. . .*	***Nuclear power stations*** *in the European Union have the highest safety standards. . .*	8

Table 6.1: Example sentences in Greek where MWEs can be at the root of translation problems. The source and reference texts were taken from the Europarl corpus. The last column shows the number of occurrences of the highlighted Greek MWE in the corpus.

terminology). These examples indicate the wide range of linguistic structures that can be classified as MWEs in Greek.

Table 6.1 illustrates the importance of MWE treatment in the context of MT. It shows a set of sentence fragments taken from the Greek portion of the Europarl corpus along with an English translation generated by a commercial MT system. [2] The corresponding reference translations from the English portion of the Europarl corpus show that the expected translations of the highlighted MWEs in the source text are clearly not equivalent to the actual output of the system.

The linguistic properties of MWEs in Greek have been the focus of considerable work (Fotopoulou 1993, Moustaki 1995, Fotopoulou 1997). However, published results about a purely computational treatment are still very limited. One of the few works concerning the acquisition of MWEs for Greek is the one of Fotopoulou et al. (2008). Their approach combines grammar rules and statistical measures in an attempt to extract from a 142,000,000-word collection of Greek texts as many nominal MWEs as possible while at the same time ensuring consistency of results. The said collection of texts is a combination of the Hellenic National Corpus and the Greek corpus maintained by the Université de Louvain. Once the corpus is tagged and lemmatised, the initial list of candidates is extracted based on a set of predefined part-of-speech patterns. This is then filtered using a set of more specific rules and word lists that identify possible, less likely and impossible MWE combinations. Depending on the type of list, a given word belongs to, the candidate can either be rejected or marked to be assigned extra weight in the statistical analysis stage. During this final step, the remaining candidates are ranked based on their

2. The result of MT was obtained through Systran's online translation service, available at `http://www.systranet.com/`. The goal of this table is to show the importance of MWEs in multilingual applications. We do not intend to compare Systran with other MT systems or to evaluate its quality. This means that other MT systems could translate these examples correctly, as well as Systran could correctly translate other MWEs in different contexts.

log-likelihood scores.

Another approach is that of Michou and Seretan (2009). They describe a Greek version of the *FipsCo* system that is able to extract collocations from corpora. Their method uses a hand-crafted generative parser for Greek built upon the *Fips* framework to analyse the sentences of the Europarl corpus and then extract MWE candidates based on syntactic patterns. The candidates are further filtered according to their association strength through the log-likelihood measure. Their system also allows the potential extraction of bilingual Greek–French MWEs when parallel corpora is available.

Despite the methodological similarities, our experiments differ from these works not only in the techniques used in each extraction step, but also in its goal: instead of building a hand-crafted specialised deep analysis tool aimed at the identification of Greek MWEs, we use the language-independent `mwetoolkit` methodology to extract shallow MWE candidates. Then, we evaluate the effectiveness of several AMs implemented by the toolkit using both textual corpora and the World Wide Web as a corpus.

The general characteristics of Greek MWEs are the same as those described in Section 2.3.1. They also vary to a great extent in terms of the fixedness of their morphosyntactic structure and of their semantic interpretation, that can be more or less transparent depending on the type of MWE (idioms tend to be less transparent than specialised terms, for example). The decision to investigate nominal MWEs (as opposed to verbal ones) was largely based on the fact that they are less heterogeneous in nature and can, therefore, be more easily encoded (Mini and Fotopoulou 2009).

The most common types of Greek nominal MWEs identified in the literature are (Anastasiadi-Symeonidi 1986, Fotopoulou et al. 2008): [3]

- J + N: In this case we have an adjective followed by a noun which constitutes the head of the phrase, for example, φορητός υπολογιστής (*laptop*), ομφάλιος λώρος (*umbilical cord*).
- N + N: MWEs of this type consist of two nouns that:
  - carry the same weight and have the same case, for example, κράτος μέλος (*member state*), παιδί θαύμα (*child prodigy*).
  - the second is in genitive and modifies the first, for example, σύνοδος κορυφής (*summit*), Υπουργείο Εξωτερικών (*ministry of foreign affairs*).
- N + DT + N: These MWEs have a noun phrase modifying a preceding noun, for example, κοινωνία της πληροφορίας (*information society*), μήλο της Έριδος (*apple of discord*).
- N + P + N: In this case we have a prepositional phrase modifying a preceding noun, for example, σκλήρυνση κατά πλάκας (*multiple sclerosis*), φόνος εκ προμελέτης (*premeditated murder*).
- P + N + N: MWEs in this category are very similar to those in the previous one in terms of their grammatical composition, the only difference being that the modifier precedes the noun it modifies, for example, διά βίου μάθηση (*lifelong learning*), κατά κεφαλήν εισόδημα (*per capita income*).

In addition to these, we are going to examine two more categories:

- N + J + N: MWEs in this category consist of an adjectival phrase in the genitive case modifying a preceding noun, for example, ξέπλυμα βρώμικου Χρήματος (*money laundering*), εμπόριο λευκής σαρκός (*white slavery*).
- N + CC + N: In this last category we come across phrases that consist of two conjoined nouns, for example, σάρκα και οστά (*[take] shape*), τελεία και παύλα

3. See the list of acronyms in the preamble of the thesis for a description of the POS tags.

```
<CHAPTER ID=1>
Έγκριση των συνοπτικών πρακτικών της
προηγούμενης συνεδρίασης
<SPEAKER ID=1 NAME="Πρόεδρος" >
Τα συνοπτικά πρακτικά της χθεσινής συνεδρίασης
έχουν διανεμηθεί.
<P>
Υπάρχουν παρατηρήσεις
<P>
<SPEAKER ID=2 LANGUAGE"IT" NAME="Speroni">
Κύριε Πρόεδρε, χτες, στο τέλος της ψηφοφορίας
σχετικά
```

Figure 6.1: Excerpt of Greek EP from 17/12/1999.

```
        (SENT       <S>
1\1     TOK         Έγκριση     έγκριση       N
1\9     TOK         των         ο             DT
1\13    TOK         συνοπτικών  συνοπτικός    J
1\24    TOK         πρακτικών   πρακτικός     J
1\34    TOK         της         ο             DT
1\38    TOK         προηγούμενης προηγούμενος J
1\51    TOK         συνεδρίασης συνεδρίαση    N
1\62    PTERM_P     .           .       PTERM_P
1\63    CHUNK       _
        )SENT       </S>
```

Figure 6.2: Tagger output containing surface form, lemma and simplified POS tag.

(*full stop*).

6.1.2 Materials and methods

The candidate extraction process was carried out on the Greek portion of the Europarl (EP) v3 corpus, described in Appendix D. It consists of 962,820 sentences and 26,306,875 words making it one of the largest Greek corpora widely available. Even though EP does not contain a great variation of text types, it can be assumed to constitute a relatively representative sample of general-purpose Greek language, mainly due to its size. An excerpt of the corpus is shown in Figure 6.1.

Before feeding the corpus into the `mwetoolkit`, we preprocessed it using external tools. In order to tag and lemmatise the corpus, we first had to remove the XML tags and split the text so that each file contained one sentence per line. Then, we used the Greek POS tagger developed at ILSP by Papageorgiou et al. (2000). Since Greek is a morphologically rich language, the tagset used for the description of the various morphosyntactic phenomena is very large compared to tagsets used by annotation schemata in other languages (584 vs 36 tags in the Penn Treebank). These labels were reduced to simplified POS tags, as those in the example of Figure 6.2. The word lemmas were determined using the ILSP morphological dictionary which contains around 80,000 lemmas corresponding to approximately 2,500,000 fully inflected entries.

The tagged corpus contains a relatively small number of errors like πρακτ ικών which has been misclassified as an adjective (ο πρακτικός — *practical*) rather than a noun (τα πρακτικά — *proceedings*). These errors may affect the extraction process since the patterns for MWE candidate extraction are defined in terms of POS tags. In this context,

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE patterns SYSTEM "mwttoolkit-patterns.dtd">
<patterns>
<pattern><w pos="J"/> <w pos="N"/></pattern><!--φορητός υπολογιστής -->
<pattern><w pos="N"/><w pos="N"/></pattern><!--κράτος μέλος, Υπουργείο Εσωτερικών-->
<pattern><w pos="N"/><w pos="DT"/><w pos="N"/></pattern><!--φαινόμενο του θερμοκηπίου-->
<pattern><w pos="N"/><w pos="J"/> <w pos="N"/></pattern><!--εμπόριο λευκής σαρκός-->
<pattern><w pos="N"/><w pos="P"/><w pos="N"/></pattern><!--σκλήρυνση κατά πλάκας-->
<pattern><w pos="P"/><w pos="N"/><w pos="N"/></pattern><!--κατά κεφαλήν εισόδημα-->
<pattern><w pos="N"/><w pos="CC"/><w pos="N"/></pattern><!--τελεία και παύλα-->
</patterns>
```

Figure 6.3: XML file containing the description of the relevant POS patterns for extraction.

```
<cand candid="13421">
  <ngram>
    <w lemma="αχίλλειος" pos= "J" >
      <freq name="EP" value="14" /><freq name="WWW" value="16700" /></w>
    <w lemma="πτέρνα" pos="N" >
      <freq name="EP" value="14" /><freq name="WWW" value="49900" /></w>
    <freq name="EP" value="14" /><freq name="WWW" value="15400" />
  </ngram>
  <occurs>
    <ngram><w surface="αχίλλειος" lemma="αχίλλειος" pos="J" />
           <w surface="πτέρνα" lemma="πτέρνα" pos="N" />
           <freq name="EP" value="8" /></ngram>
    <ngram><w surface="Αχίλλειος" lemma="aq'illeios" pos="J" />
           <w surface="πτέρνα" lemma="pt'erna" pos="N" />
           <freq name="EP" value="1" /></ngram>
    <ngram><w surface="αχίλλειο" lemma="aq'illeios" pos="J" />
           <w surface="πτέρνα" lemma="pt'erna" pos="N" />
           <freq name="EP" value="5" /></ngram>
  </occurs>
  <features>
    <feat name="pos-pattern" value= "J#S#N#S" /><feat name="n" value="2" />
    <feat name="mle-EP" value="7.4773e-07" /><feat name="pmi-EP" value="44.5092" />
    <feat name="t-EP" value="3.7416" /><feat name="dice-EP" value="1.0" />
    <feat name="mle-WWW" value="3.08e-07" /><feat name="pmi-WWW" value="55.3587" />
    <feat name="t-WWW" value="124.0966" /><feat name="dice-WWW" value="0.4624" />
  </features>
</cand>
```

Figure 6.4: Extract of the XML output file with MWE candidates and their AM scores.

tagging errors cause some candidates to be incorrectly kept (false positives) while others are incorrectly removed (false negatives). This, however, cannot be avoided in situations where large quantities of automatically POS tagged data are employed and their manual checking is not feasible, as is the case here.

Once the corpus was cleaned, tagged and lemmatised, it was fed as input to the `mwetoolkit`. The seven POS patterns in Figure 6.3 are defined on the basis of the types discussed in Section 6.1.1. Its application on the Greek EP corpus produced 526,012 word sequences. In order to reduce the effects of data sparseness and avoid computational overhead, we disregarded *n*-grams that occurred less than 10 times. The size of the list of candidates reduced to 25,257 word sequences, which constitute our list of MWE candidates.

For each candidate entry, `mwetoolkit` gets the individual word counts both in EP and in the web through Yahoo! search API. These, combined with the *n*-gram joint count, are used to calculate four statistical AMs for each MWE candidate: pointwise mutual information (`pmi`), maximum likelihood estimator (`mle`), Student's *t* score (`t-score`) and Dice's coefficient (`dice`), as described in Section 3.1.4.

The `mwetoolkit` outputs a file containing the following information on each MWE candidate: the lemma forms and POS tags of its individual words, the counts of these

words as well as of the entire *n*-gram sequence both in EP and in the web, all the surface forms of each candidate together with their counts in the original corpus (EP) and a set of features that correspond to the candidate's score for each AM. An example of an extracted candidate is showed in Figure 6.4. The candidates are sorted into eight lists, according to each AM based on the EP and on the web counts.

6.1.3 Results

Since there is, to our knowledge, no gold standard containing a considerable number of MWE entries in Greek, there is no way of automatically evaluating which are the interesting MWEs among the candidates. Consequently, evaluation was performed manually by three native speakers. In terms of the typology proposed in Chapter 4, our evaluation is intrinsic and quantitative, involves manual annotation, and is type-based. Our evaluation is based on precision as performance measure, in spite of its limitations and oversimplification, as discussed in Section 4.3. In order to calculate recall, however, we would need to know how many MWEs exist in EP, in the web, or more generally in the Greek language. Given that it is impossible to know and very difficult to estimate these values, our evaluation procedure will be based on precision only.

Due to the size of the candidate list (25,257 candidates), it was not possible to perform exhaustive manual judgement of all the candidates. Instead, the human judges annotated a sample containing the first 150 candidates proposed by each measure. From these, we manually removed the most striking cases of noise (introduced by the tagger) such as single words or candidates that appeared more than once based on a different grammatical classification. In short, each annotator classified around 1,200 entries in one of the following categories:

1. `mwe`: the candidate is a MWE, that is, a true positive;
2. `maybe`: the candidate is ambiguous, but it may be considered as a MWE depending on the context of use;
3. `part`: the candidate includes a or is part of a MWE or;
4. `not`: the candidate is not a MWE, but a regular sequence of words with no particularity.

In the following evaluation steps, we considered MWEs to be those that were classified as such (`mwe`) by at least two out of three of our judges. This is a conservative evaluation scheme that does not take into account other categories such as `maybe` and `part`. Therefore, we also propose a scoring scheme that will be described later in this section to be used in the creation of the final dictionary of Greek MWEs.

Our initial anticipation, given evaluation results reported in the literature (Evert and Krenn 2005), was that the `dice` coefficient or the `pmi` would be the first in rank, followed by `t-score` and then `mle`. As Figure 6.5 shows, this was not exactly the case. Considering only EP counts, the $\texttt{dice}_{\texttt{EP}}$ coefficient did indeed have the highest score, 81.08%. This level of precision surpassed all our expectations since it is one of the highest reported in the Greek literature. The second highest precision (58.21%) is achieved by the $\texttt{t-score}_{\texttt{EP}}$, followed by the $\texttt{mle}_{\texttt{EP}}$ at approximately the same levels (57.43%), leaving $\texttt{pmi}_{\texttt{EP}}$ behind with a precision of 52.66%.

The most surprising result obtained, however, was the level of precision achieved by `mle`. This measure does not take into account individual token frequencies, which led us to believe that it would be a very poor judge of MWEness. Surprisingly enough though, it did turn out not to be the case.

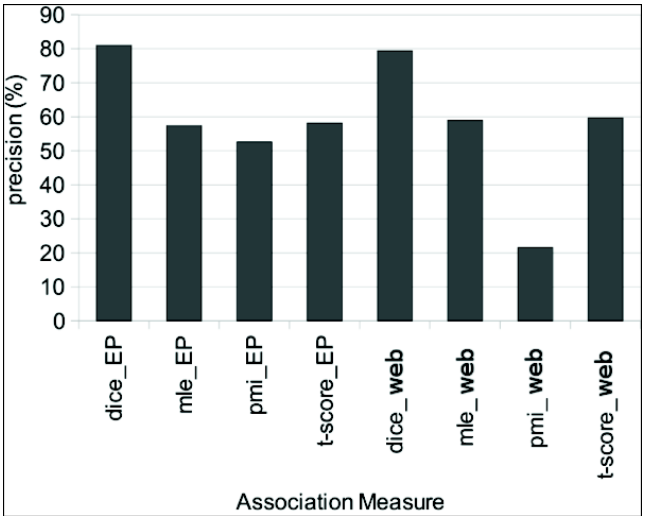


Figure 6.5: Precision based on the EP counts.

	mwe	**maybe**	**part**	**not**	κ	$s \geq 4$
$\texttt{dice}_{\text{EP}}$	78%	10%	3%	9%	40%	82%
$\texttt{mle}_{\text{EP}}$	55%	9%	3%	33%	65%	60%
$\texttt{pmi}_{\text{EP}}$	50%	9%	16%	26%	52%	57%
$\texttt{t-score}_{\text{EP}}$	56%	9%	3%	32%	61%	61%
$\texttt{dice}_{\text{web}}$	78%	8%	2%	12%	56%	84%
$\texttt{mle}_{\text{web}}$	58%	6%	1%	36%	74%	60%
$\texttt{pmi}_{\text{web}}$	21%	7%	36%	36%	63%	24%
$\texttt{t-score}_{\text{web}}$	58%	6%	1%	35%	70%	61%

Table 6.2: Inter-annotator agreement for each of the four categories and each evaluated AM in both corpora, as well as Fleiss' kappa coefficient (κ) and proportion of true positives according to score $s \geq 4$.

The web-based precision for each AM other than the $\texttt{pmi}_{\text{web}}$ reached the same levels as the EP-based one. More precisely, the $\texttt{dice}_{\text{web}}$ coefficient yielded a precision of 79.43%, corresponding to a marginal decrease of approximately 2%. $\texttt{mle}_{\text{web}}$ and $\texttt{t-score}_{\text{web}}$, on the other hand, did show an increase of $2.6\% - 2.7\%$, with their exact precision values being 58.99% and 59.71% respectively. These values seem to confirm our earlier assumption that EP, despite its lack of textual genre variation, can reasonably be assumed to contain a representative sample of the Greek language, mainly due to its size. The most striking result about the web-based results, however, is the dramatic decrease (almost 60% lower) in the precision the $\texttt{pmi}_{\text{web}}$ measure achieves (a very unimpressive 21.62%).

These values seem to both verify and contradict some of the arguments presented in the literature about the use of the web as a corpus. The slight increase in the precision achieved by the `t-score` and the `mle` measures seems to indicate that the size of the web makes it an invaluable tool for the MWE extraction process and possibly for other NLP applications as well. The magnitude of the precision decrease of `pmi`, however, seems to indicate that the threshold of 10 *n*-gram occurrences, which was more than satisfactory in the case of EP, turned out to be a serious underestimate in the case of the web, where almost all of the proposed candidates were wrong.

At the same time, `pmi` seems to overestimate the importance of the size of the word sequence since the candidate lists consisted entirely of three-word candidates both in the case of EP and web as opposed to, say, the `dice` coefficient whose candidate lists consisted of entirely of 2-grams (something that can be attributed to their higher number of occurrences in general language use).

A large number of the candidates proposed by `pmi` included partial MWEs, which were not proposed as a unit by themselves, but in combination with some other word. To be more precise, out of the 148 candidates evaluated, 32 were classified as MWEs while 50 included some MWE, which in the majority of cases was εν λόγω (*in question*). Indeed, some of the candidates classified as `part` or `maybe` should be manually analysed for deciding whether to include them as entries in the dictionary, as they could constitute interesting MWEs.

Therefore, to evaluate the MWE list, we propose a scoring scheme where each candidate is assigned a value *s*. This score depends on the number of judges that classified the candidate as an instance of a category (`mwe`, `maybe` or `part`). The precise formulation of the scheme to be adopted depends on which criterion one wants to emphasise: precision or coverage. To emphasise precision, one could consider as genuine MWEs only those candidates classified as `mwe` by most judges. On the other hand, to emphasise coverage, one can also consider those candidates classified as `maybe` and `part`. In addition, a preference on the categories can also be taken into account in the scoring scheme, where each category could be assigned a specific weight depending on how much influence it has. For instance, for unambiguous MWEs to be given more weight than ambiguous or partial cases, `mwe`, `maybe` and `part` can be given decreasing weights.

The scoring scheme adopted in our evaluation is:

$$s = 2 \times \#(\texttt{mwe}) + \#(\texttt{maybe}) + \#(\texttt{part})$$

For this evaluation, we consider as interesting MWE candidates those that have a score greater than or equal to 4, including cases which were classified as `mwe` by at least one judge and as ambiguous/partial by the others. We did not chose among the evaluated AMs, but combined the four EP-based lists into a single one since the candidate lists retrieved

by each measure are very heterogeneous. The web-based results were disregarded, since they did not bring performance improvements over EP-results (this does not mean that they could not be useful in the case of smaller corpora, for example).

As an additional evaluation, we quantified the difficulty of the classification task for the human judges. Therefore, we calculated the inter-judge agreement rate using Fleiss's kappa coefficient. The results for each analysed AM are summarised in Table 6.2: the first four columns correspond to the individual agreement proportions for each of the categories while the last two columns of the table contain respectively the kappa value and the proportion of instances that were considered as true MWEs according to the scoring scheme proposed above. The values in the last column are slightly greater than the performance values showed in Figure 6.5, mainly because the scoring scheme is less conservative than the majority vote used to perform the preliminary evaluation of each AM independently.

The agreement coefficients are very heterogeneous, ranging from $\kappa = 40\%$ to $\kappa = 74\%$. A coefficient of 40%, for example, means that there is a probability of 40% that this agreement was not obtained by chance. This explains such low κ values despite the high agreement for category `mwe`, which is also the most frequent in this data set. The coefficient is, therefore, unable to assign more importance to a given category. Moreover, although there is no general agreement on how to interpret these results, it is believed that kappa values should be above 70%. Our results show, however, that there is no high agreement among annotators according to this criterion. If we look in detail at the proportion of agreement for each category, we can see that annotators are quite at ease to identify true MWEs, whereas, for the other classes, the agreement is much lower (e.g., annotators cannot truly distinguish `maybe` from `part`). While, on one hand, this might be caused by ambiguous annotating guidelines, on the other hand, it is also an indicator of how difficult it is for a human annotator to identify and classify MWEs.

We also discovered that there is high correlation ($r \approx .99$) between the agreement on category `mwe` and the precision of the method. That is, it is easy to identify true MWEs in a high-quality list, whereas it is much more difficult to select useful MWEs when the list contains a lot of noise. While this might seem obvious, it corroborates the hypothesis that precise automatic methods can considerably speed up lexicographic work in the process of language resources creation. Additionally, the agreement is always higher when web-based AMs are analysed, and this is not in direct correlation to the performance of the AM. At first glance, we can suppose that it is easier to interpret the results coming from a web-based method than the results from EP, even if the former does not necessarily improve precision. This issue, however, needs further investigation, since it is not clear to date what benefits one could take from the web combining with or replacing well-formed corpora like EP.

The results of manual evaluation by the three native speakers were joined and resulted in a lexicon of 815 nominal MWEs in Greek. The dictionary was made freely available on the MWE community website. [4]

4. `http://multiword.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets`

6.2 Acquisition and analysis of Portuguese complex predicates [5]

In this section, we describe the creation of two related lexical resources for concrete NLP applications dealing with Brazilian Portuguese. Both resources are dictionaries including complex predicates, that is, expressions which act as a predicate in a sentence and which are composed of a verb and a complement.

The first dictionary was constructed based on a concrete need of a semantic role labelling task. Semantic role labelling annotation depends on the correct identification of predicates, before identifying arguments and assigning them role labels. However, most predicates are not constituted only by a verb: they constitute *complex predicates* (CPs) not yet available in a computational lexicon. In order to create a dictionary of CPs aimed at semantic role labelling (henceforth, CP-SRL), we employed the `mwetoolkit` using POS tag sequences instead of a limited list of verbs or nouns, in contrast to similar studies. The resulting CPs include (but are not limited to) light and support verb constructions.

The second lexicon constructed using a similar methodology also contains CPs, but is aimed at a different application: sentiment analysis. Therefore, our experiments investigates how sentiments are expressed in Brazilian Portuguese. Sentiment verbs like *temer, (fear), odiar (hate)* and *invejar (envy)* are examples of lexical units specifically used to express the respective feelings. The same meaning may be conveyed through other verbs associated to sentiment nouns. Our experiments firstly identify 7 recurrent patterns of sentiment expression through CPs and then employ these patterns to identify sentiment nouns associated to them. We will refer to the lexical resource resulting from these experiences as CP-SENT.

The remainder of this section is structured as follows: we start by introducing and exemplifying the characteristics of CPs in Brazilian Portuguese in Section 6.2.1. Then, in Section 6.2.2 we present the corpus, the POS patterns employed and the acquisition methodology using the `mwetoolkit`. Then, we present the analysis of the results and the creation of the CP-SRL lexicon in Section 6.2.3.1, and analogously, for CP-SENT in Section 6.2.3.2. We conclude with a discussion on the role of the `mwetoolkit` in the creation of both resources (Section 6.2.4).

6.2.1 Portuguese complex predicates

Complex predicates can be defined as "predicates which are multi-headed: they are composed of more than one grammatical element" (Alsina et al. 1997, p. 1), like *give a try, take care, take a shower*. The correct identification of CPs is a crucial step in *semantic role labelling* (SRL) and for sentiment analysis. We examine the behaviour and importance of CPs for these two applications separately in Section 6.2.1.1 and Section 6.2.1.2.

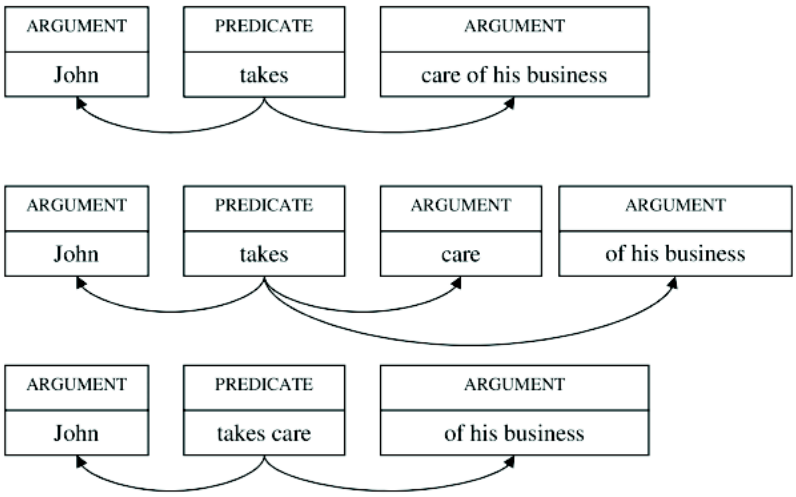
6.2.1.1 Complex predicates and semantic role labelling

Independently of the approach adopted, SRL comprehends two steps before the assignment of role labels: (a) the delimitation of argument takers and (b) the delimitation of arguments. If the argument taker is not correctly identified, the argument identification will propagate the error. Argument takers are predicates, represented by a verb or by a CP.

5. Work reported in this section was previously published in the papers *Identifying and Analysing Brazilian Portuguese Complex Predicates* (Duran et al. 2011) and *How do you feel? Investigating lexical-syntactic patterns in sentiment expression* (Duran and Ramisch 2011). It was carried out with the collaboration of Magali Sanchez Duran, Sandra Maria Aluisio and Aline Villavicencio.

The verbal phrases identified by a parser are usually used to automatically identify argument takers, but do no suffice. A lexicon of CPs, as well as the knowledge about verbal chains composition, are required for the fully automatic identification of argument takers. Consequently, the possibility of disagreement between SRL annotators would rely only on the assignment of role labels to arguments. The first part of our experiments reports the creation of the CP-SRL lexicon, in order to meet the needs arisen from a SRL annotation task in a corpus of Brazilian Portuguese [6].

To stress the importance of these CPs for SRL, consider the sentence *John takes care of his business* in three alternatives of annotation:



The first annotation shows *care of his business* as a unique argument, masking the fact that this segment is constituted of a predicative noun, *care*, and its internal argument, *of his business*. The second annotation shows *care* and *of his business* as arguments of *take*, which is incorrect because *of his business* is clearly an argument of *care*. The third annotation is the best for SRL purposes: as a unique predicate — *take care*, *take* shares its external argument with *care* and *care* shares its internal argument with *take*.

One of the goals of this section is to describe our computer-aided corpus-based method used to build a comprehensive machine-readable dictionary of CPs for SRL. In addition to the lexicon creation, we analyse these expressions and their behaviour in order to shed some light on the most adequate lexical representation for further integration of our resource into a SRL annotation task. To the best of our knowledge, to date, there is no similar study regarding these complex predicates in Brazilian Portuguese, focusing on the development of a lexical resource for NLP tasks, such as SRL.

In CP-SRL, we classify CPs into two groups: idiomatic CPs and less idiomatic CPs. Idiomatic CPs are those whose sense may not be inferred from their parts. Examples in Portuguese are *fazer questão (insist on), ir embora (go away), dar o fora (get out), tomar conta (take care), dar para trás (give up), dar de ombros (shrug), passar mal (get sick, faint)*. On the other hand, we use "less idiomatic CPs" to refer to those CPs that vary in a continuum of different levels of compositionality, from fully compositional to semi-compositional sense, that is, at least one of their lexical components may be literally understood and/or translated. Examples of less idiomatic CPs in Portuguese are: *dar instrução (give instructions), fazer menção (mention), tomar banho (take a shower), tirar foto (take a photo), entrar em depressão (get depressed), ficar triste (become sad)*.

6. CPs constituted by verbal chains (e.g., *have been working*) are not considered here.

Less idiomatic CPs headed by a predicative noun have been called in the literature *light verb constructions* (LVCs) or *support verb constructions* (SVCs). Although both terms have been employed as synonyms, *light verb* is, in fact, a semantic concept and *support verb* is a syntactic concept. The term light verb is attributed to Jespersen (1965) and the term support verb was already used by Gross in 1981. A light verb is the use of a polysemous verb in a non-prototypical sense or "with a subset of their [its] full semantic features", North (2005). Common light verbs in English are *give*, *get*, *set*, *make* and *take*, but many other verbs which have a concrete meaning but act as light verbs when combined with some nouns, like *experience a development*. On the other hand, a support verb is the verb that combines with a noun to enable it to fully predicate, given that some nouns and adjectives may evoke internal arguments, but need to be associated with a verb to evoke the external argument, that is, the subject. As the function of support verb is almost always performed by a light verb, attributes of LVCs and SVCs have often been merged, making them near synonyms. Against this tendency, our analysis shows cases of SVCs without light verbs (*trazer prejuízo* = *damage*, lit. *bring damage*) and cases of LVCs without support verbs (*dar certo* = *work well*, lit. *give correct*).

Part of the CPs focused on here are represented by LVCs and SVCs. These CPs have been studied in several languages from different points of view: diachronic (Ranchhod 1999, Marchello-Nizia 1996), contrastive (Danlos and Samvelian 1992, Athayde 2001), descriptive (Butt 2003, Langer 2004; 2005) and for NLP purposes (Salkoff 1990, Stevenson et al. 2004, Barreiro and Cabral 2009, Hwang et al. 2010). Work focusing on the automatic extraction of LVCs or SVCs often take as starting point a list of recurrent light verbs (Hendrickx et al. 2010) or a list of nominalisations (Teufel and Grefenstette 1995, Dras 1995, Hwang et al. 2010). These approaches are not adopted here because our goal is precisely to identify which are the verbs, the nouns and other lexical elements that take part in CPs.

Closer to our study, Hendrickx et al. (2010) annotated a Treebank of 1M tokens of European Portuguese with almost 2,000 CPs, which include LVCs and verbal chains. This lexicon is relevant for many NLP applications, notably for automatic translation, since in any task involving language generation they confer fluency and naturalness to the output of the system. Similar motivation to study LVCs/SVCs (that is, for SRL) is found within the scope of Framenet (Atkins et al. 2003) and Propbank (Hwang et al. 2010). These projects have taken different decisions on how to annotate such constructions. Framenet annotates the head of the construction (noun or adjective) as argument taker (or frame evoker) and the light verb separately; Propbank, on its turn, first annotates separately light verbs and the predicative nouns (as ARG-PRX) and then merges them, annotating the whole construction as an argument taker.

Regarding Portuguese LVCs/SVCs in both European (Athayde 2001, Rio-Torto 2006, Barreiro and Cabral 2009, Duarte et al. 2010) and Brazilian Portuguese (Neves 1996, Conejo 2008, Silva 2009, Abreu 2011), we verified differences in combination patterns of both variants beyond the variations due to dialectal aspects. Brazilian Portuguese studies do not aim at providing data for NLP applications, whereas in European Portuguese there are at least two studies focusing on NLP applications: Barreiro and Cabral (2009), for automatic translation and Hendrickx et al. (2010) for corpus annotation.

6.2.1.2 Complex predicates and sentiment analysis

Sentiment analysis and opinion mining are a growing topic of interest in the last few years due to the large amount of texts produced through web facilities, like social net-

working, blogs, e-mail and chats. These texts contain information about what people think and feel, which constitute valuable information. However, it is humanly impossible to deal with such increasing amount of data. In order to facilitate human analysis or even substitute it, computer-based techniques are required. For this reason, sentiment analysis has become a popular research subject and a challenge for the NLP community.

Whatever the strategy used, it is essential to count on a sentiment lexicon. However, even when they contain sentiment words, some utterances are not instances of sentiment expression. In the sentence *overcoming fear is a skill that anyone can learn*, for example, the sentiment noun *fear* is a topic of discourse. In this example, nobody can be identified as feeling fear, as well as nothing can be identified as causing fear. In order to identify this kind of utterances, it is possible to associate morphosyntactic features to the sentiment lexicon and use only sentiment verbs instead of nouns when searching for sentiment expression. But this is not a complete solution. Although sentiment verbs are lexical items specifically used to express feelings, they are not the only way to do this. In Portuguese, it is possible and frequent to express feelings using other verbs associated to sentiment nouns. For example, in the sentence *João tem inveja de você. (lit. João has envy of you = João envies you)*, the sentiment expressed is *inveja (envy)*, *João* is the one who feels envy and *você (you)* is the cause (or stimulus) for *João* feeling envy.

It would be interesting, indeed, that a Portuguese sentiment lexicon includes collocations like *ter inveja*, which corresponds to the verb *invejar (to envy)*. Analogously to the problem of SRL described above, it is relevant for sentiment data mining to know how to determine who is feeling the expressed sentiment and what is causing the expressed sentiment. Hence, our experiments aim to explore recurrent patterns used to express feelings in Portuguese, using verbs other than sentiment verbs, in order to provide new lexical syntactic inputs for sentiment analysis.

A comprehensive review of sentiment analysis and opinion mining as a research field for NLP is presented in Pang and Lee (2008). The review provides guidance for those interested in developing opinion mining search engines. The authors address the problem of deciding where to mine opinion and sentiment expression, how to gather information and how to present the information gathered.

Due to the role played by the lexicon in sentiment analysis systems, the NLP related tasks are highly language dependent. An ontological approach, as proposed by López et al. (2008) and Mathieu (2005) may benefit the semantic description of the sentiment lexicon and pave the way for multilingual approaches.

Besides the identification of sentiment words, there are studies dedicated to enriching the description of these words, aggregating features that enable clustering the gathered information. Up to this date, features regarding sentiment words are almost always related to their polarity, as may be seen in Kim and Hovy (2004), in SentiWordNet (Esuli and Sebastiani 2006) and in SentiLexPT, [7] this latter being a lexical resource for Portuguese.

In Portuguese, there are few reported studies related to sentiment analysis (Silva et al. 2009, Carvalho et al. 2011). Due to their role in political and marketing decisions, sentiment analysis and opinion mining systems constitute a competitive advantage. This fact encourages private financial support for developing new resources that remain undisclosed. The growing need for lexical resources aimed at sentiment analysis and the role played by CPs in sentiment expression motivate our efforts toward the creation of the CP-SENT lexicon, using the `mwetoolkit` methodology.

7. `http://dmir.inesc-id.pt/reaction/SentiLex-PT_01`

6.2.2 Materials and methods

We employ a corpus-based methodology in order to create the dictionaries of CPs. After a first step in which we use the `mwetoolkit` to automatically acquire candidate *n*-grams from the corpus, the candidate lists have been analysed by a linguist to distinguish CPs from fully compositional word sequences. For the automatic acquisition step, the PLN-BR-FULL [8] corpus was used. The corpus was first preprocessed for sentence splitting, case homogenisation, lemmatisation and POS tagging using the PALAVRAS parser, described in Appendix D.

Differently from the studies referred to in Section 6.2.1, we did not presume any closed list of light verbs or nouns as starting point to our searches. The search criteria we used in order to acquire CPs for SRL are composed of seven POS patterns observed in examples collected during previous corpus annotation tasks: [9]

1. V + N + P: *abrir mão de* (*give up*, lit. *open hand of*);
2. V + P + N: *deixar de lado* (*ignore*, lit. *leave at side*);
3. V + DT + N + P: *virar as costas para* (*ignore*, lit. *turn the back to*);
4. V + DT + R: *dar o fora* (*get out*, lit. *give the out*);
5. V + R: *ir atrás* (*follow*, lit. *go behind*);
6. V + P + R: *dar para trás* (*give up*, lit. *give to back*);
7. V + J: *dar duro* (*work hard*, lit. *give hard*).

We will refer to this set of POS patterns as PAT-SRL. This strategy is suitable to extract occurrences from active sentences, both affirmative and negative. Cases which present intervening material between the verb and the other elements of the CP are not captured. This does not seem to be a serious problem, considering the size of our corpus, although it influences the frequencies used in candidate selection. After generating separate lists of candidates for each pattern with the `mwetoolkit`, we filtered out all those occurring less than a certain threshold. This threshold was set based on the analysis presented in Figure 6.6. The graphic shows that precision increases logarithmically while the drop in recall is roughly linear, so that a good compromise of the F-measure can be obtained by filtering out all entries that occur less than 10 times in this corpus. This value was retained as a filter for the experiments with unrestricted verbal combinations.

In order to create the CP-SRL lexicon, we manually analysed the candidates generated by the `mwetoolkit`. During this analysis, constructions with sentiment nouns were found. These findings motivated the creation of the second lexicon, CP-SENT, using the following patterns: [10]

1. *sentir* N *de* — *to feel* N *of*
2. *sentir* N *por* — *to feel* N *for*
3. *ter* N *de* — *to have* N *of*
4. *ter* N *por* — *to have* N *for*
5. *ficar com* N *de* — *to become with* N *of*
6. *estar com* N *de* — *to be with* N *of*
7. *dar* N *em* — *to give* N *in*

The identification of these syntactic patterns was performed empirically based on data observation of the data in the CP-SRL lexicon and on trial and error. We will refer to

8. See the corpus description in Appendix D
9. See the list of acronyms in the preamble of the thesis for a description of the POS tags.
10. The placeholder N stands for a sentiment noun.

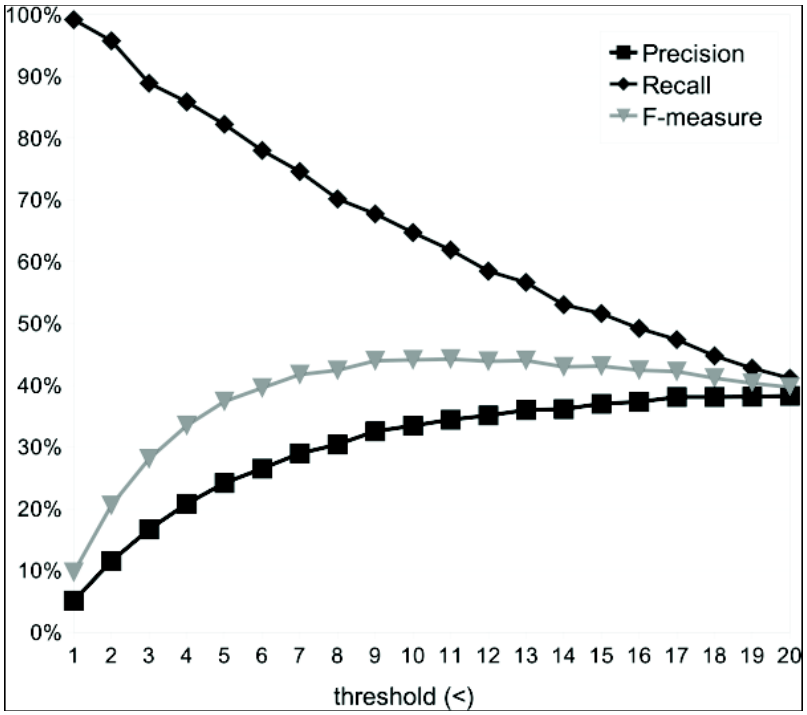


Figure 6.6: Quality comparison of threshold values.

this set of patterns as PAT-SENT. Notice that, instead of using abstract POS like for CP-SRL, we used the identified lemmas of the support verbs of sentiment nouns. This was necessary because these patterns correspond to syntactic configurations in which, in most cases, the sentiment noun is part of a CP instead of being the topic of conversation, as discussed in Section 6.2.1.2. If we had used POS instead of lemmas, the resulting list would be too noisy to be useful for lexicographic purposes.

The patterns in PAT-SENT allowed us to manually identify 98 sentiment nouns. We combine all the nouns with all the patterns in PAT-SENT, thus artificially generating 686 variations that were automatically looked up in the web using the `mwetoolkit`. By using the web, we can estimate the acceptability of a given pattern based on a very large sample of current language use. The original corpus itself is not large nor representative enough to allow the distinction between unacceptable constructions and constructions that were not found in the corpus due to sparsity or limited representativity. Additionally, as Portuguese has verb inflections and in the web we cannot search for lemmas, we generated inflected forms for each variation. For instance, the candidate *ter medo de (to have fear of)* became *ter* | *tem* | *teve* | *tinha medo de (to have* | *has* | *had* | *was having fear of)*, where the vertical bar | denotes the alternative. That is, this query retrieves any sequence containing one of the forms of the verb *ter* in infinitive, present, past perfect or imperfect followed by the target sentiment noun and the corresponding preposition.

The example below shows the queries and the resulting number of hits generated for the target sentiment word *consciência (conscience)*. The query in bold corresponds to the preferred pattern, that is, the pattern that maximises the hit counter for the target noun. The underlined queries are acceptable patterns, that is, patterns that return three hits or more:

pattern	extracted	analysed	less idiomatic	idiomatic
V + N + P	69,264	2,140	327	8
V + P + N	74,086	1,238	77	8
V + DT + N + P	178,956	3,187	131	4
V + DT + R	1,537	32	0	0
V + R	51,552	3,626	19	41
V + P + R	5,916	182	0	2
V + R	25,703	2,140	145	11
Total	407,014	12,545	699	74

Table 6.3: Number of candidates extracted from the corpus and analysed.

dar | *dá* | *deu* | *dava consciência em*	0
ficar | *fica* | *ficou* | *ficava com consciência de*	3
estar | *está* | *esteve* | *estava com consciência de*	2
sentir | *sente* | *sentiu* | *sentia consciência de*	6
sentir | *sente* | *sentiu* | *sentia consciência por*	0
ter | tem | teve | tinha consciência de	47,600
ter | *tem* | *teve* | *tinha consciência por*	179

6.2.3 Results

In spite of using a common methodology that uses morphosyntactic patterns for MWE acquisition, both lexicons, CP-SRL and CP-SENT have different purposes. Therefore, the analysis of the results of automatic acquisition is presented in two parts. First, we analyse each of the patterns used for CP acquisition in the context of SRL, focusing on idiomaticity and single-verb paraphrases (Section 6.2.1.1). Second, we analyse the patterns used for sentiment analysis, in terms of their precision and of the polarity and source of the acquired sentiment nouns (Section 6.2.1.2).

6.2.3.1 Analysis of the CP-SRL lexicon

Each of the POS patterns contained in the PAT-SRL set returned a large number of candidates. Our expectation was to identify CPs among the most frequent candidates. First we annotated "interesting" candidates and then, in a deep analysis, we judged their idiomaticity. In Table 6.3, we show the total number of candidates extracted before applying any threshold (extracted), the number of analysed candidates using a threshold of 10 (analysed) and the number of CPs correctly identified divided into two columns: idiomatic and less idiomatic CPs. In addition to the idiomaticity judgement, each CP was annotated with one or more single-verb paraphrases. Sometimes, it is not a simple task to decide whether a candidate constitutes a CP, specially when the verb is a very polysemous one and is often used as support verb. For example, *fazer exame em/de alguém/alguma coisa* (lit. *make exam in/of something/somebody*) is a CP corresponding to *examinar* (*exam*). But *fazer exame* in another use is not a CP and means to submit oneself to someone else's exam or to perform a test to pass examinations (*take an exam*).

The pattern V + N is very productive, as every direct object of a transitive verb not introduced by preposition takes this form (*buy tickets, make plans, write letters*). For this reason, we restricted the pattern, adding a preposition after the noun with the aim of

capturing only nouns that have their own complements (*have fear of, have pride of, take advantage of*).

We identified 335 CPs, including both idiomatic and less idiomatic ones. For example, *bater papo* (*shoot the breeze*, lit. *hit chat*) or *bater boca* (*have an argument*, lit. *hit mouth*) are idiomatic, as their sense is not compositional. On the other side, *tomar consciência* (*become aware*, lit. *take conscience*) and *tirar proveito* (*take advantage*) are less idiomatic, because their sense is closer to the meanings of the nouns. The candidates selected with the pattern V + N + P presented 29 different verbs, as shown in Figure 6.7. [11]

Sometimes, causative verbs, like *causar* (*cause*) and *provocar* (*provoke*) give origin to constructions paraphrasable by a single verb. In spite of taking them into consideration, we cannot call them LVCs, as they are used in their full sense while light verbs have a bleached semantic contribution. Examples:

– *provocar alteração* (*provoke alteration*)= *alterar* (*alter*);
– *causar tumulto* (*cause riot*) = *tumultuar* (*riot*).

Some of the candidates returned by this pattern take a deverbal noun, that is, a noun created from a verb, as stated by most works on LVCs and SVCs. But the opposite also occurs: some constructions present denominal verbs as paraphrases, like *ter simpatia por* (*have sympathy for*) = *simpatizar com* (*sympathise with*) and *fazer visita* (lit. *make visit*) = *visitar* (*visit*). These results contradict the hypothesis stating that LVCs result only from the combination of a deverbal noun with a light verb. In addition, we have identified idiomatic LVCs that are not paraphrasable by verbs of the same word root, like *fazer jus a* (lit. *make right to*) = *merecer* (*deserve*).

Moreover, we have found some constructions that have no correspondent paraphrases, like *fazer sucesso* (lit. *make success*) and *abrir exceção* (lit. *open exception*). These findings evidence that the most popular test to identify LVCs and SVC — the existence of a paraphrase formed by a single verb — has several exceptions.

We have also observed that, when the CP has a paraphrase by a single verb, the prepositions that introduce the arguments may change or even be suppressed, like in:

– *dar apoio* **a** *alguém* = *apoiar alguém* (*give support* **to** *somebody* = *support somebody*);
– *dar cabo* **de** *alguém ou* **de** *alguma coisa* = *acabar* **com** *alguém ou* **com** *alguma coisa* (*give end* **of** *somebody or* **of** *something* = *end* **with** *somebody or* **with** *something*).

Finally, some constructions are polysemic, like:

– *dar satisfação a alguém* (lit. *give satisfaction to somebody*) = make somebody happy or provide explanations to somebody;
– *chamar atenção de alguém* (lit. *call the attention of somebody*) = attract the attention of somebody or reprehend somebody.

The results of the pattern V + P + N contain too much noise, as many transitive verbs share with this CP class the same POS tags sequence. We found constructions with 12 verbs, as shown in Figure 6.8. We classified seven of these constructions as idiomatic CPs: *dar de ombro* (*shrug*), *deixar de lado* (*ignore*), *pôr de lado* (*put aside*), *estar de olho* (*be alert*), *ficar de olho* (*stay alert*), *sair de férias* (*go out on vacation*). The latter example is very interesting, as *sair de férias* is a synonym of *entrar em férias* (*enter on vacation*), that is, two antonym verbs are used to express the same idea, with the same syntactic frame. In the remaining constructions, the more frequent verbs are used to give an aspectual meaning to the noun: *cair em*, *entrar em*, *colocar em*, *pôr em* (*fall in*, *enter*

11. We provide one possible (most frequent sense) English translation for each Portuguese verb.

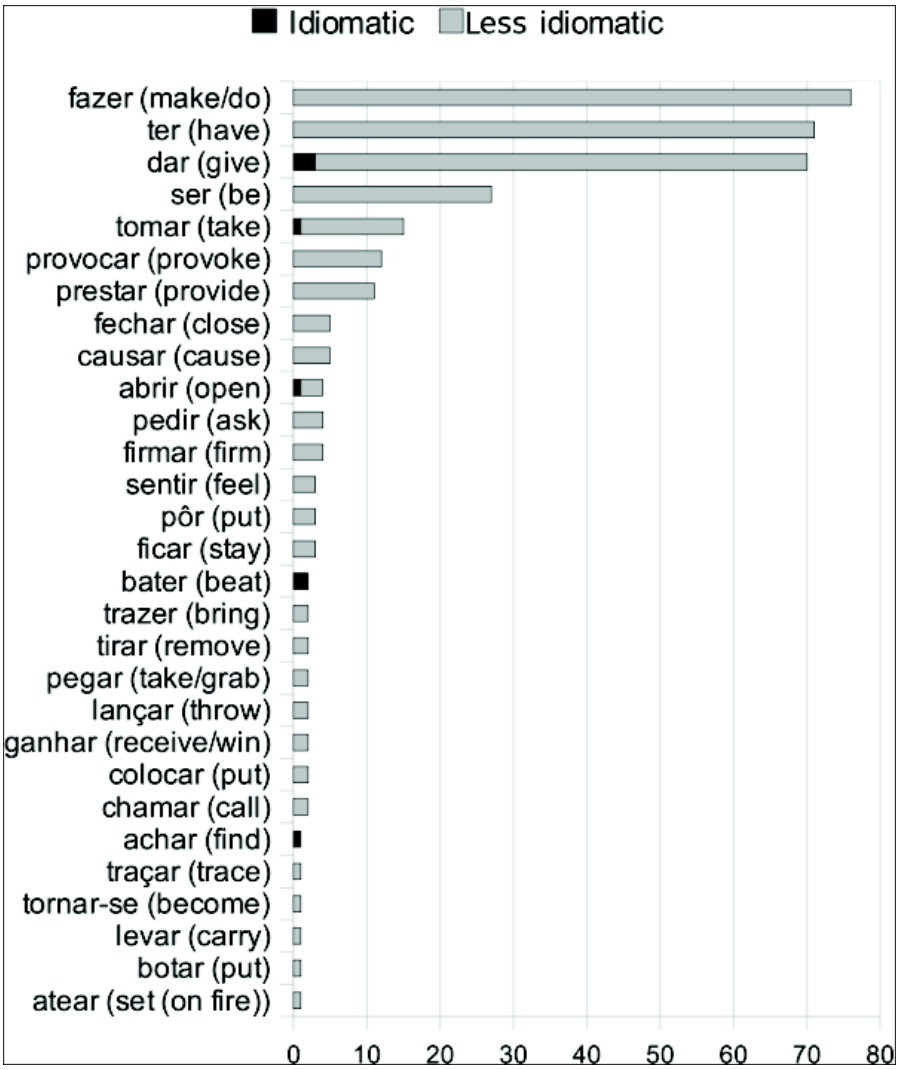


Figure 6.7: Distribution of verbs involved in CPs, pattern V + N + P.

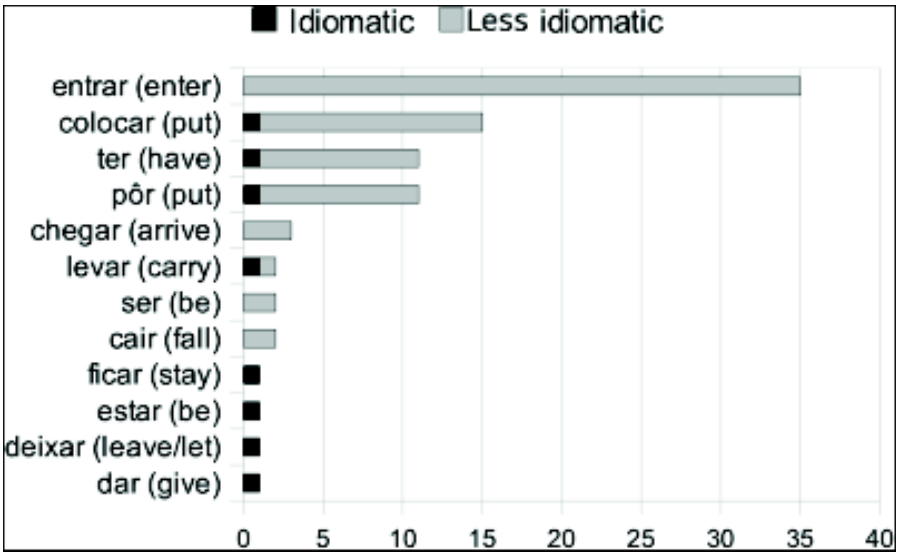


Figure 6.8: Distribution of verbs involved in CPs, pattern V + P + N.

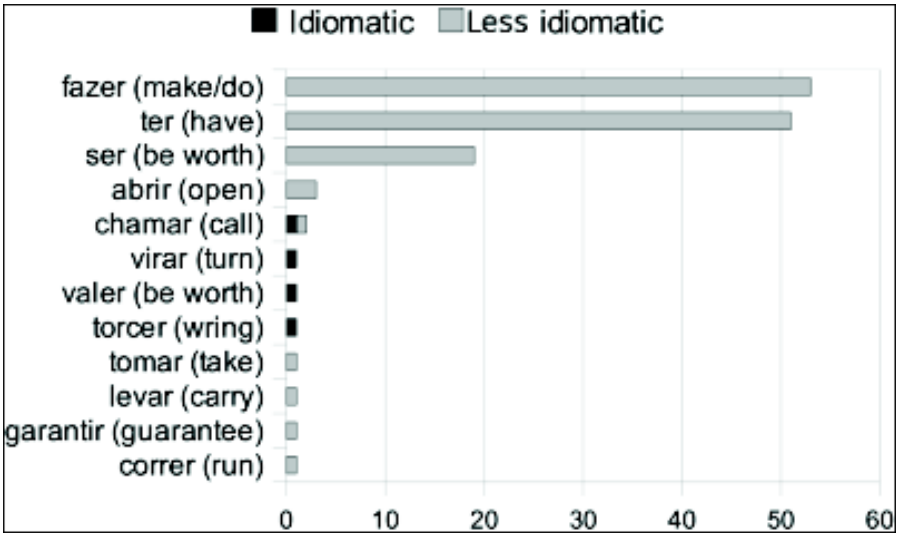


Figure 6.9: Distribution of verbs involved in CPs, pattern V + DT + N + P.

in, *put in*) have inchoative meaning, that is, indicate an action starting, while *chegar a* (*arrive at*) has a resultative meaning.

The results of the pattern V + DT + N + P are very similar to the pattern V + N + P, proving that it is possible to have determiners as intervening material between the verb and the noun in less idiomatic CPs. The verbs involved in the candidates validated for this pattern are presented in Figure 6.9.

The verbs *ser* (*be*) and *ter* (*have*) are special cases. Some *ter* expressions are paraphrasable by an expression with *ser* + J, for example:

– *Ter a responsabilidade por* = *ser responsável por* (*have the responsibility for* = *be responsible for*);
– *Ter a fama de* = *ser famoso por* (*have the fame of* = *be famous for*);
– *Ter a garantia de* = *ser garantido por* (*have the guarantee of* = *be guaranteed for*).

Some *ter* expressions may be paraphrased by a single verb:

– *Ter a esperança de* = *esperar* (*have the hope of* = *hope*);

– *Ter a intenção de = tencionar* (*have the intention of = intend*);
– *Ter a duração de = durar* (*have the duration of = last*).

Most of the *ser* expressions may be paraphrased by a single verb, as in *ser uma homenagem para = homenagear* (*be a homage to = pay homage to*). The verb *ser*, in these cases, seems to mean *to constitute*. These remarks indicate that the patterns *ser* + DT + N and *ter* + DT + N deserve further analysis, given that they are less compositional than they are usually assumed in Portuguese.

We have not identified any CP following the pattern V + DT + R. This pattern was inspired by the complex predicate *dar o fora* (*escape*, lit. *give the out*). Probably this is typical in spoken language and has no similar occurrences in our newspaper corpus. Similarly, the pattern V + P + R is not productive, but helped identify two expressions: *deixar para lá (put aside)* and *achar por bem (decide)*.

The pattern V + R is the only one that returned more idiomatic than less idiomatic CPs, for instance:

– *vir abaixo = desmoronar* (lit. *come down = crumble*);
– *cair bem = ser adequado* (lit. *fall well = be suitable*);
– *pegar mal = não ser socialmente adequado* (lit. *pick up bad = be inadequate*);
– *estar de pé* [12] *= estar em vigor* (lit. *be on foot = be in effect*);
– *ir atrás (de alguém) = perseguir* (lit. *go behind (somebody) = pursue*);
– *partir para cima (de alguém) = agredir* (lit. *leave upwards = attack*);
– *dar-se bem = ter sucesso* (lit. *give oneself well = succeed*);
– *dar-se mal = fracassar* (lit. *give oneself bad = fail*).

In addition, some CPs identified through this pattern present a pragmatic meaning: *olhar lá* (*look there*), *ver lá* (*see there*), *saber lá* (*know there*), *ver só* (*see only*), *olhar só* (*look only*), provided they are employed in restricted situations. The adverbials in these expressions are expletives, not contributing to the meaning, exception made for *saber lá*, (lit. *know there*) which is only used in present tense and in first and third persons. When somebody says *Eu sei lá* the meaning is *I don't know*.

Here we identified three interesting clusters concerning the pattern V + J: attributive verbs, expressions involving predicative adjectives and idiomatic CPs.

1. **Attributive verbs**, that is, an object and an attribute assigned to the object. These verbs are: *achar* (*find*), *considerar* (*consider*), *deixar* (*let/leave*), *julgar* (*judge*), *manter* (*keep*), *tornar* (*make*) as in: *Ele acha você inteligente* (lit. *He finds you intelligent = He considers you intelligent*). For SRL annotation, we will consider them as full verbs with two internal arguments. The adjective, in these cases, will be labeled as an argument. However, constructions with the verbs *fazer* and *tornar* followed by adjectives may give origin to some deadjectival verbs, like *possibilitar = tornar possível* (*possibilitate = make possible*). Other examples of the same type are: *esclarecer* (*make clear*), *evidenciar* (*make evident*), *inviabilizar* (*make unfeasible*), *popularizar* (*make popular*), *responsabilizar* (*hold responsible*), *viabilizar* (*make feasible*).

2. **Expressions involving predicative adjectives**, in which the verb performs a functional role, in the same way as support verbs do in relation to nouns. In contrast to predicative nouns, predicative adjectives do not select their "support" verbs: they combine with any verb of a small set of verbs called copula. Examples of copula verbs are: *acabar* (*finish*), *andar* (*walk*), *continuar* (*continue*), *estar* (*be*), *ficar*

12. The POS tagger classifies *de pé* as R.

(*stay*), *parecer* (*seem*), *permanecer* (*remain*), *sair* (*go out*), *ser* (*be*), *tornar-se* (*become*), *viver* (*live*). Some of these verbs add an aspect to the predicative adjective: durative (*andar*, *continuar*, *estar*, *permanecer*, *viver*) and resultative (*acabar*, *ficar*, *tornar-se*, *sair*).
   - The resultative aspect may be expressed by an infix, substituting the combination of V + J by a full verb: *ficar triste* = *entristecer* (*become sad*) or by the verbalisation of the adjective in reflexive form: *ficar tranquilo* = *tranquilizar-se* (*calm down*); *estar incluído* = *incluir-se* (*be included*).
   - In most cases, adjectives preceded by copula verbs are formed by past participles and inherit the argument structure of the verb: *estar arrependido de* = *arrepender-se de* (lit. *be regretful of* = *regret*).

3. **Idiomatic CPs**, like *dar duro* (lit. *give hard* = *make an effort*), *dar errado* (lit. *give wrong* = *go wrong*), *fazer bonito* (lit. *make beautiful* = *do well*), *fazer feio* (*make ugly* = *fail*), *pegar leve* (lit. *pick up light* = *go easy*), *sair errado* (lit. *go out wrong* = *go wrong*), *dar certo* (lit. *give correct* = *work well*).

In total, we identified 699 less idiomatic CPs and observed the following recurrent pairs of paraphrases:

- V = V + deverbal N, for example, *tratar* = *dar tratamento* (*treat* = *give treatment*);
- Denominal V = V + N, for example, *amedrontar* = *dar medo* (*frighten* = *give fear*);
- Deadjectival V = V + J, for example, *responsabilizar* = *tornar responsável* (lit. *responsibilise* = *hold responsible*).

Further extensions of the CP-SRL lexicon can consider this fact, as we may search for denominal and deadjectival verbs (which may be automatically recognised through infix and suffix rules) to manually identify corresponding CPs. Moreover, the large set of verbs involved in the analysed CPs, summarised in Figure 6.10, shows that any study based on a closed set of light verbs will be limited, as it cannot capture common exceptions and non-prototypical constructions. The CP-SRL lexicon containing idiomaticity and paraphrase information is available at the MWE community website. [13]

6.2.3.2 Analysis of the CP-SENT lexicon

The result of applying the patterns PAT-SENT on the PLN-BR-FULL corpus consist of 7 candidate lists, one for each pattern, with the collocated nouns and their respective count in the corpus. The 1,774 candidates are distributed as described in Table 6.4. The noisy occurrence lists have been carefully analysed by human annotators in order to distinguish nouns denoting sentiments from other nouns, for example *ter ódio de* vs *ter camisa de* (lit. *to have hate of* vs *to have shirt of*). The analysis of these lists identified 173 combinations of sentiment nouns used with the patterns. Comparing the quantity of candidates analysed (column 1) with the quantity of candidates validated (column 2), we found the precision of each pattern (column 3). This measure indicates how much a pattern is associated with sentiment nouns or, in other words, how specific is a pattern to express feelings.

The pattern *ter* N *de* returned the largest amount of validated candidates, but, at the same time, it is the one that presented one of the largest amounts of noise. This is most probably due to the high polysemy of the verb *ter (to have)*. In this sense, the patterns *sentir* N *de* and *sentir* N *por* are much less ambiguous and their precision ranges from 44.9% to 72.22%, respectively. Patterns 5 and 6 have a similar profile; both are responsi-

13. `http://multiword.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets`

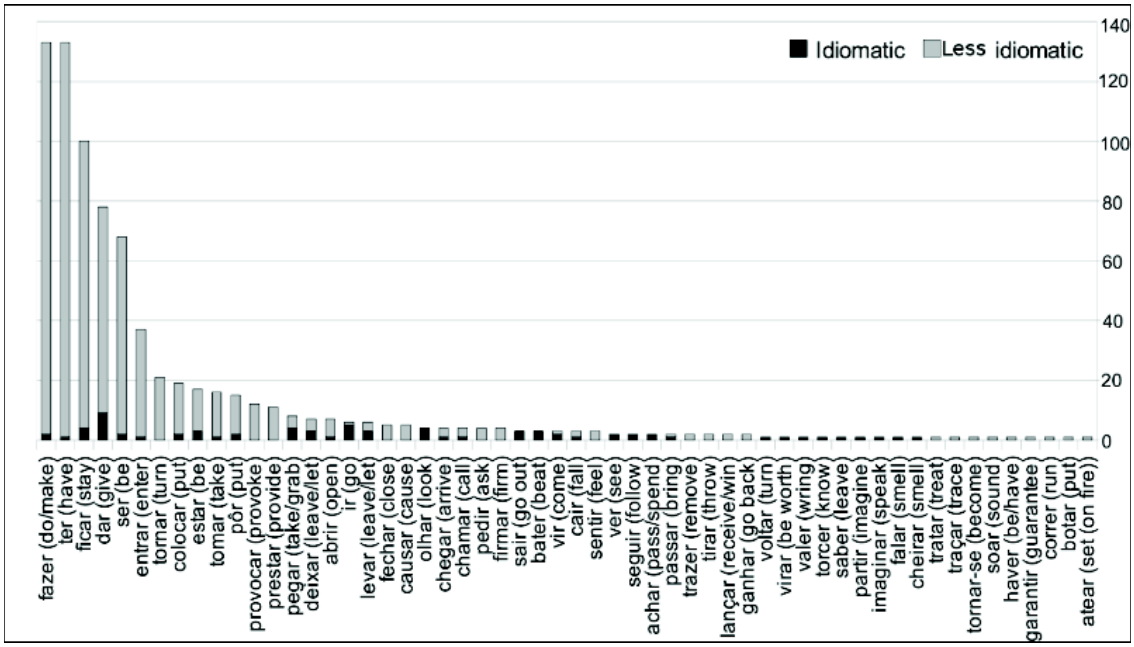


Figure 6.10: Distribution of verbs involved in CPs, total number of CPs (all patterns).

	Pattern	Candidates	TPs	Precision	Coverage
1	*sentir* N *de*	49	22	44.90%	12.72%
2	*sentir* N *por*	18	13	**72.22%**	7.51%
3	*ter* N *de*	1,218	69	5.67%	**39.88%**
4	*ter* N *por*	131	29	22.14%	16.76%
5	*ficar com* N de	51	14	27.45%	8.09%
6	*estar com* N *de*	92	16	17.39%	9.25%
7	*dar* N *em*	215	10	4.65%	5.78%

Table 6.4: Number of candidates extracted and validated per pattern.

Polarity	# Expressions	Examples
negative	45	*hate, contempt, grudge*
positive	29	*love, tenderness, compassion*
neutral	15	*interest, impression, curiosity*
context dependent	9	*pride, ambition, anxiety*

Table 6.5: Distribution of sentiment nouns according to their polarity.

Source	# Expressions	Examples
psychological-emotional	67	*jealousy, sympathy, anger*
psychological-rational	18	*confidence, respect, concern*
physical	13	*cold, thirst, hunger, pain*

Table 6.6: Distribution of sentiment nouns according to their source.

ble for 8% and 9% of the final list, with a precision between 17.39% for *estar* and 27.45% for *ficar*. Pattern 7 presents the lowest precision, 4.5%, which is expected as the verb *dar* is highly polysemous in Portuguese.

The 173 validated candidates present some repetitions of nouns which occur in more than one pattern. Eliminating the redundancies, we obtained a list of 98 sentiment nouns. We observed and annotated two features associated to these sentiment nouns: polarity and source. Polarity was annotated by two human judges, as it involves subjectivity (Kim and Hovy 2004, Esuli and Sebastiani 2006, Silva et al. 2009). The result is shown in Table 6.5. We notice that most of the expressions found actually express negative emotions. We propose two hypotheses to explain this fact: either this is a bias from our newspaper corpus (there are often more bad news than good news in general newspapers) or Brazilian Portuguese native speakers prefer to use the identified patterns instead of sentiment verbs, because they somehow diminish/blur the impact of the negative emotion expressed.

The second feature that we annotated is the source of the feeling expressed by the sentiment noun. This made it possible to distinguish physical sensations, expressed through the same patterns, from more psychological feelings. Furthermore, we separated rational feelings from emotional feelings, as shown in Table 6.6.

After the corpus-based extraction, we generated web-based variations for each identified sentiment noun, as described in Section 6.2.2. Results showed some variations with zero occurrences. This may be due to the implausibility of the combination or due to limitations of our search arguments, which should be refined. For example, we realised that the pattern *dar* N *em* is almost always presented with a personal pronoun taking the place of the experiencer, avoiding the preposition *em* and preceding the verb: *isso me dá medo* (lit. *this gives me fear)*. The same pattern may be used without the experiencer, in utterances like *dá medo pensar nisso* (lit. *give fear thinking about this = thinking about this causes fear)*.

Aiming to verify whether the preferred way to express feelings varies according to the feeling expressed, we plotted the graphic shown in Figure 6.11. It shows how many sentiment nouns take each pattern as preferred pattern. This table evidences the pattern *ter* N *de* as the preferred one for expressing 61 of a total of 98 sentiment nouns. Therefore, this pattern is extensively used to express feelings. All patterns but one (*ficar com* N *de*) are preferred by at least one sentiment noun.

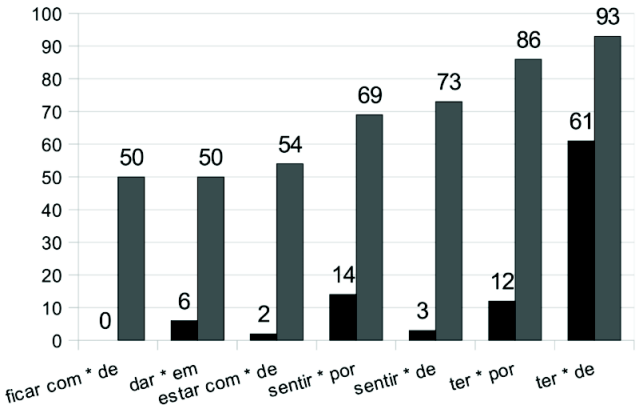


Figure 6.11: Number of sentiment nouns (*y* axis) that prefer (dark bars) and accept (light bars) each pattern.

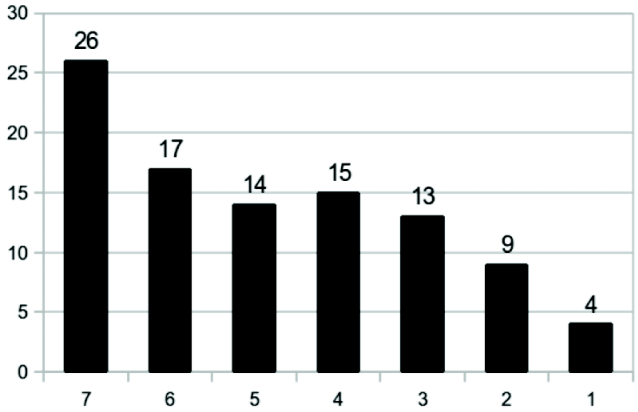


Figure 6.12: Number of nouns (*y* axis) vs number of patterns accepted (*x* axis).

In Figure 6.12, we present the quantity of sentiment nouns that accept [14] one or more patterns. With these data, we are able to distinguish more variable constructions from more fixed ones. Lexicalised constructions present zero count of all alternative patterns except for the preferred one. This is the case of 4 sentiment nouns. Most of the nouns, however, are quite variable and accept several patterns, although it is not clear whether alternative patterns express the same sentiment with the same connotation and use.

6.2.4 Discussion

The growing importance of sentiment analysis encourages further developments of this work. Our analysis identified a large number of CPs useful for SRL and for sentiment analysis. The automatic approach proved to be very useful to identify verbal MWEs, notably with POS tag patterns that have not been explored by other studies (patterns not used to identify LVCs/SVCs). However, due to the cost of manual annotation, we use an arbitrary threshold of 10 occurrences that removes potentially interesting candidates. Our hypothesis is that, in a machine-readable dictionary, as well as in traditional lexicography, rare entries are more useful than common ones, and we would like to explore alternatives to address this issue.

Second, we strongly believe that our patterns are sensitive to corpus genre, because the CPs identified are typical of colloquial register. Thus, a limitation of our work is using a corpus of news. The same patterns should be applied on a corpus of spoken Brazilian Portuguese, as well as other written genres like blog posts. A corpus of speech transcriptions, blogs (Gill et al. 2008) or social networking posts would more likely present CPs and sentiment expression. Due to its size and availability, web-based corpora would also allow us to obtain better frequency estimators.

We underline, however, that we should not underestimate the value of our original corpus, as it contains a large amount of unexplored material. We observed that only the context can tell us whether a given verb is being used as a full verb or as a light and/or support verb. [15] As a consequence, it is not possible to build a comprehensive lexicon of light and support verbs, because there are full verbs that function as light and/or support verbs in specific constructions, like *correr* (*run*) in *correr risco* (*run risk*). As we discarded a considerable number of infrequent lexical items, it is possible that other unusual verbs participate in similar CPs which have not been identified by our study.

For the moment, it is difficult to assess a quantitative measure for the quality and usefulness of our resource, as no similar work exists for Portuguese. Moreover, the lexical resource presented here is far from being complete. A standard resource for English like DANTE, [16] for example, contains 497 support verb constructions involving a fixed set of 5 support verbs, and was evaluated extrinsically with regard to its contribution in complementing the FrameNet data (Atkins 2010). Likewise, we would like to evaluate our resource in the context of SRL annotation, to measure its contribution in automatic argument taker identification. It would also be interesting, for instance, to compare, across genres, utterances using sentiment verbs with utterances using the patterns we have identified. For this purpose, one may use the list of sentiment verbs from Brazilian Wordnet (da Silva 2010), provided in Appendix G.2, and the sentiment nouns obtained in this study,

14. We say that a noun "accepts" a pattern if the count returned by the web search engine is greater than 3 pages, thus avoiding noise probably due to typos and artificial results.

15. A verb is not light or support in the lexicon, it is light and/or support depending on the combinations in which it participates.

16. `http://www.webdante.com`

listed in Appendix G.3, associated with the patterns here discussed. Another alternative to obtain more constructions is the use of serious lexical games such as JeuxDeMots, as described in Section 8.2.

There are many possible extensions to the present research, which could help build a broad-coverage lexicon of CPs in Brazilian Portuguese. This lexicon may contribute to different NLP applications, in addition to SRL and sentiment analysis. We believe that computer-assisted language learning systems and other Portuguese as second language learning material may take great profit from it. Analysis systems like automatic textual entailment may use the relationship between CPs and paraphrases to infer equivalences between propositions. Computational language generation systems may also want to choose the most natural verbal construction to use when generating texts in Portuguese. Furthermore, these MWEs may be used to improve bilingual dictionaries with information on how to express sentiments from the point-of-view of a Brazilian speaker.

6.3 Summary

A first quantitative and qualitative evaluation of the framework for MWE acquisition proposed was performed in the context of computer-aided lexicography. We have collaborated with colleagues who are experienced linguists and lexicographers in order to create new lexical resources containing MWEs in Greek and in Portuguese. The created data sets are freely available. [17]

For Greek, considerable work has been done to study the linguistic properties of MWEs, but computational approaches are still limited (Fotopoulou et al. 2008). In our experiments, we used the `mwetoolkit` to extract an initial list of MWE candidates from the Greek Europarl corpus. We extracted words matching the following patterns: adjective-noun, noun-noun, noun-determiner-noun, noun-preposition-noun, preposition-noun-noun, noun-adjective-noun and noun-conjunction-noun. For filtering these candidates, we applied a set of statistical association measures using counts collected both from the corpus and from the web. The top-150 ranked candidates produced by four AMs applied on two different corpora were manually evaluated by three native speakers. Each annotator judged around 1,200 candidates and in the end the annotations were joined, creating a lexicon with 815 Greek nominal MWEs.

Based on these judgements, we analysed the precise contribution of the different AMs to the number of correct MWEs retrieved. The AM that produced better results was `dice`, which significantly outperformed the other measures, followed by the `t-score`. The performance of the latter, however, is surprisingly similar to the performance of raw *n*-gram counts, suggesting that sophisticated measures are not needed when enough data is available. In relation to the use of the web as a corpus, it has a number of advantages over standard corpora, the most salient being its availability and accessibility. However, in our experiments, the results obtained with web counts did not bring considerable improvements. In sum, our results indicate that automatic methods can indeed be used for extending NLP resources with MWE information, and improving the quality of NLP systems that support Greek.

The goal of the work with Portuguese complex predicates (CPs) was to perform a qualitative analysis of these constructions. We generated two lexical resources based on two target applications: CP-SRL is aimed at semantic role label annotation while CP-

17. `http://multiword.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets`

SENT is aimed at sentiment analysis. For both resources, we POS-tagged the PLN-BR-FULL corpus and extracted sequences of words matching specific POS patterns using the `mwetoolkit`.

Semantic role label annotation depends on the correct identification of predicates, before identifying arguments and assigning them role labels. However, many predicates are not constituted only by a verb: they constitute CPs not available in a computational lexicon. In order to create the dictionary CP-SRL, we used POS sequences instead of a limited list of verbs or nouns: verb-[determiner]-noun-preposition, verb-preposition-noun, verb-[preposition/determiner]-adverb and verb-adjective. The extraction process resulted in a list of 407,014 candidates which were further filtered using statistical AMs. An expert human annotator manually validated 12,545 candidates, from which 699 were annotated as compositional verbal expressions and 74 as idiomatic verbal expressions. Results include (but are not limited to) light and support verb constructions. We observed the following recurrent pairs of paraphrases:

- V = V + DEVERBAL N: *tratar* = *dar tratamento* (*treat* = *give treatment*);
- DENOMINAL V = V + N: *amedrontar* = *dar medo* (*frighten* = *give fear*);
- DEADJECTIVAL V = V + ADJ: *responsabilizar* = *tornar responsável* (lit. *responsibilise* = *hold responsible*).

For the creation of CP-SENT, our goal was to investigates how sentiments are expressed in Brazilian Portuguese. Sentiment verbs like *temer, (fear), odiar (hate)* and *invejar (envy)* are examples of lexical units specifically used to express the respective feelings. The same meaning may be conveyed through other verbs associated to sentiment nouns. This study firstly identifies seven recurrent patterns of sentiment expression without sentiment verbs and then employs these patterns to identify sentiment nouns associated to them. This was performed in five steps. First, we identified recurrent lexical-syntactic patterns to express feelings using sentiment nouns instead of sentiment verbs. Second, we used the patterns identified as search arguments to identify sentiment expression. Third, a human analysed the candidate lists resulting from step two, determining whether the noun collocated at the right of each pattern was or not a sentiment noun. Fourth, we analysed the validated candidates and assigned them some features. Fifth, we combined the patterns of step one with the sentiment nouns identified in step three and searched the combinations in the web. Analysis of the patterns showed that combining sentiment nouns with the seven patterns may be useful to automatically identify sentiment expression and additionally know who is feeling and who or what is causing the feeling.

7 APPLICATION 2: EMPIRICAL MACHINE TRANSLATION

Throughout the previous chapters, we have demonstrated at several points that MWEs are a source of errors for machine translation (MT) systems and for human non-native speakers of a language. As Manning and Schütze (1999, p. 184) point out, "a nice way to test whether a combination is a collocation [MWE] is to translate it into another language. If we cannot translate the combination word by word, then there is evidence that we are dealing with a collocation". In Section 2.3.1, we argue that the fact that MWEs cannot be translated word-for-word is a consequence of their limited compositionality. Adequate solutions for the variable syntactic/semantic fixedness of MWEs are not easy to find, especially in the context of empirical phrase-based MT systems. However, for high quality MT, it is important to detect MWEs, to disambiguate them semantically and to treat them appropriately in order to avoid generating unnatural translations or losing information in the process.

The automatic translation of MWEs can generate unnatural and sometimes funny translations, as exemplified in Table 6.1 and in Table 1.1. In addition to these cases, where a MWE in the source language is translated as another MWE in the target language, MWEs may imply lexical and grammatical asymmetries between languages. In other words, an expression in the source language can be expressed as a single word in the target language, and vice versa. This particular case is the focus of our experiments in this chapter. Concretely, we will deal with *phrasal verbs* (PVs), so abundant in English, but absent in other languages like Portuguese, where the particle may be omitted (e.g., *clean up* as *limpar*, literally *clean*). However, as PVs are often semantically non-compositional, their contribution may involve a more complex translation to another language with the target verb being unrelated to the source verb and possibly the inclusion of additional material (e.g., *they made out* as *eles se beijaram*, literally *they themselves kissed*).

In the following experiments, we adopted as experimental context the Moses system, a phrase-based empirical MT toolkit. When it comes to complex linguistic phenomena like MWEs, traditional expert systems have, since decades, much more sophisticated mechanisms to deal with them and would be the natural choice for our experiments in this chapter. However, empirical systems are a very popular MT paradigm that has received much emphasis in the last years. Moreover, there are many open source and freely available tools to create a competitive empirical system from scratch quite quickly. In short, as discussed by Stymne (2011) phrase-based empirical MT is "a very successful approach, and has received much research focus. Other approaches [...] have the drawback of being more complex [and] can still gain from preprocessing." Nonetheless, in the future we would like to test the techniques developed for empirical MT in the context of expert systems, and therefore a reasonable option would be the open-source Apertium system (Forcada 2009).

Our experiments investigate how PVs affect the output of a standard English–Portuguese empirical MT system. Our goal is to explore possible ways for integrating them into the system in order to improve translation quality. We perform an in-depth automatic and manual evaluation. Our analysis shows how the linguistic, semantic and distributional characteristics of PVs affect the results obtained, and which solutions better handle these. We show that current empirical MT technology cannot deal with PVs and that further efforts need to be made in MT evaluation for taking them into account.

The figures reported here are not final, but correspond to the results of ongoing experiments that we are conducting in order to evaluate the feasibility of the integration between the `mwetoolkit` and MT systems. From this perspective, we were required to simplify many aspects of the experiments. In the ideal scenario, the results of automatic MWE acquisition are plugged directly into the translation model, generating improved translations for the correctly identified MWEs. There are two important differences between this ideal scenario and the one presented in our experiments:

– We use a manually built dictionary of PVs in English, instead of using the results of the `mwetoolkit`. In theory we could have used the `mwetoolkit`, but as its results will undoubtedly contain noise, we decided not to propagate this noise through the translation pipeline. We wanted to be 100% sure of the correct identification of MWEs, so that we could more easily identify the points in the translation model where the translation errors were generated. This allowed us to have more control over the process in order to, in the future, integrate automatically acquired MWEs directly into the MT system.
– The MWEs inserted into the MT system are identified only on the source side, using monolingual matching. But, intuitively, the use of bilingual identification would be more helpful and would generate more significant improvements. While the acquisition of bilingual MWEs has been the focus of some related work (see Section 3.2.2), this is far from being a solved problem and the quality of results is still below our expectations. When it comes to the acquisition of asymmetric constructions, to the best of our knowledge there are no published results. Nonetheless, we use a simplification similar to that used for the previous issue. That is, we use an existing manually constructed bilingual lexicon that, in the future, can be replaced by automatically acquired MWEs (assuming that the techniques for bilingual MWE acquisition evolve).

This chapter starts with a brief introduction of empirical methods used to train empirical MT systems (Section 7.1). Then, we discuss some existing techniques used in expert and empirical MT systems for dealing with MWEs (Section 7.2). Finally, we present one of the most important contributions of the present thesis, that is, the results of ongoing experiments on the integration of phrasal verbs into a baseline MT system (Section 7.3).

7.1 A brief introduction to empirical MT

Empirical MT includes what is often called *statistical MT (SMT)* (Lopez 2008, Koehn 2010), although the term *empirical* is more adequate to distinguish between this paradigm and expert systems. Conversely, *expert MT* includes *rule-based* or *transfer-based* MT, even if any system, expert or empirical, will probably contain both transfer rules *and* statistics at some point of processing. What distinguishes these two paradigms is not the translation model itself, but the way the model is built. We are going to employ the standard acronym, SMT, to make our text more accessible, even though we advocate for

the use of the term *empirical MT* instead, which includes systems traditionally referred to as *example-based MT*. We will also use the standard acronym PB-SMT to refer to phrase-based statistical MT systems.

Knight and Koehn (2003) describe SMT as translating a sentence from a source language S into a target language T with an intermediary "broken" target language B. The traditional path in Vauquois' triangle (Vauquois 1968) is as follows: at each step, there are rules allowing the analysis of S words in terms of their syntax, semantics and a corresponding language-independent representation; then progressively generating semantics, syntax and words in T. The "broken" language B corresponds to a knowledge-poor (empiric) translation model in which rules going directly from S words to T words are learned by probabilistic algorithms. This corresponds to a "shortcut" in the translation triangle, as shown by Knight and Koehn (2003).

This paradigm was inaugurated by seminal papers by the IBM research group (Brown et al. 1993). Given that a large amount of parallel text is available (Koehn 2005), one can overcome the limitations of using barely no linguistic information. Many concepts from Section 3.1.3 are used in these systems, such as the noisy channel model and the markovian n-gram language models. Och (2005) emphasises that SMT is a classical decision problem of finding the best translation for a sentence in a very large search space and relying on an approximative model. Lopez (2008) views SMT as a machine learning problem, like in traditional artificial intelligence, with extra difficulties related to the complexity of language.

One advantage of SMT systems is that their models are independent of languages, and a new language pair may be added to the MT system with little effort. However, in order to allow this straightforward adaptation, one needs a very large volume of parallel data to train the model on, and this is not readily available for every language pair. Even though SMT seems to be the current trend in MT, the approach seems to reach its limitations when it comes to domain adaptation, traceability of errors, integration of external lexical, syntactic and semantic knowledge. The experiments reported in Section 7.3 represent a step toward the integration of external lexical resources containing MWEs into SMT systems.

The next sections will overview the main aspects of SMT. The construction of a SMT system is viewed as a composition of three tasks: preprocessing (Section 7.1.1), model learning (Section 7.1.2) and decoding (Section 7.1.3). We also dedicate some lines to the evaluation of MT systems (Section 7.1.4). The present section is based on the textbook of Koehn (2010) and on the freely available survey of Lopez (2008).

7.1.1 Preprocessing a parallel corpus

A parallel corpus is a set of texts in two or more languages, in which the documents are the translations of each others. Examples of parallel corpora found in everyday life include film subtitles translated into several languages, the multilingual instructions manual of your new hair drier, phrase books and restaurant menus for tourists. In SMT, most large parallel corpora come from international political institutions such as the United Nations, the World International Patent Organisation or from the transcriptions of multilingual parliaments such as those of Canada (Hansard corpus) or the European Union (Europarl corpus).

Once we have gathered a set of parallel documents, it is necessary to align them, both at the sentence level and at the word level. There are many algorithms and tools for performing sentence alignment of a parallel corpus. Anchors like numbers, question

marks, dates and proper names can be used to find equivalent fragments in both languages. One of the most popular algorithms for sentence alignment is based on a statistical model of sentence length, assuming that equivalent sentences have roughly the same length (Gale and Church 1993). Generally, sentence pairs in which the difference in length is too large are discarded, as well as very long sentences (say longer than 40 words).

After sentence alignment, it is necessary to tokenise the text so that words are represented coherently throughout the text (see Section 3.1.1). Then, it is usual to lowercase the corpus in order to avoid double representation of the same word depending on its position (at the beginning or in the middle of the sentence). However, as discussed in Section 3.1.1, one should apply lowercasing with parsimony, especially on domain-specific corpora.

Generally, after these four steps (sentence alignment, cleaning, tokenisation and lowercasing), the corpus can be word-aligned. In our experiments, all these four steps were performed. For tokenisation and lowercasing, we did not use simple regular expressions, instead we used the TreeTagger. This avoided, for instance, to lowercase words that should be kept in original capitalisation and allowed an informed decision about tokenisation.

7.1.2 Learning a translation model

A parallel corpus aligned on the sentence level does not contain links between the individual source and target words. Most current SMT systems rely on some word alignment software such as GIZA++ to align the words in the parallel corpus (Och and Ney 2000; 2004; 2003). In the latter, word alignment is modelled using probabilities. Given one source word s_i in a sentence, there is a probability $p(t_j|s_i)$ that it is translated as any of the words t_j on the target side, and each of these probabilities is a parameter of the model. Parameter estimation can be solved using the expectation-maximisation algorithm (Dempster et al. 1977), that tries to maximise the model parameters based on the probability of seen data and on the model of unseen data.

The word alignment algorithm initialises the parameters uniformly, assuming that each pair of words s_i, t_j in a sentence pair are connected with the same probability. In a first iteration, word pairs that co-occur in the same sentence pairs will have their probabilities increased. For example, considering that the English word *doctor* occurs often on the source side of sentences containing the French word *médecin* on the target side, the probability $p(\textit{médecin}|\textit{doctor})$ increases while the probabilities linking the word *doctor* to other words decreases. Thus, with a succession of expectation and maximisation steps, the algorithm will strengthen the links between words that co-occur frequently, eliminating weak links whose probabilities fall below a threshold. More sophisticated word translation models include probabilities for words to be inserted (fertility), removed (null word) and reordered. This progression of translation models of increasing complexity, trained using the expectation-maximisation algorithm, is referred to as the IBM models.

One of the problems of the IBM models is that, while one-to-many alignments are possible, many-to-one alignments cannot be represented. Phrase-based SMT (PB-SMT) models emerged as an attempt to solve this problem, better taking into account the local context of words. One of the most popular software for training an PB-SMT system is the open-source Moses toolkit [1] (Koehn et al. 2007). Moses uses a word-aligned corpus as input, from which it learns a translation model composed of several different components that are combined using a log-linear feature model (Knight and Koehn 2003, Koehn et al.

1. http://www.statmt.org/moses/

Source s	**Target** t	$p(t\|s)$	$lex(t\|s)$	$p(s\|t)$	$lex(s\|t)$
a baby being born blind	uma criança cega	1	0.0106327	1	0.026239
a backward step .	de uma regressão .	1	0.0280532	0.5	0.002579
a backward step .	uma regressão .	1	0.0280532	0.5	0.027814
a backward step	de uma regressão	1	0.0287083	0.5	0.002676
a backward step	uma regressão	1	0.0287083	0.5	0.028855
a bad foundation for	uma má base para	1	0.0009332	1	0.004316
a bad foundation	uma má base	1	0.0036263	1	0.018618
a bad	uma má	1	0.1378	1	0.049648

Table 7.1: Example of phrase table containing bi-phrases with English source (s), Portuguese target (t), phrase translation probabilities ($p(t|s)$ and $p(s|t)$) and lexical translation probabilities ($lex(t|s)$ and $lex(s|t)$).

2003).

The main component of the Moses translation model is the phrase table, that is, a table containing sentence fragments (the "phrases") in the source language and the corresponding sentence fragment or phrase in the target language, as the example shown in Table 7.1.[2] Each bilingual phrase (also called bi-phrase) has several associated probabilities that are integrated into the log-linear model as features. In order to create the phrase table, Moses generates word alignments by running GIZA++ in both directions (source $\rightarrow$ target and target $\rightarrow$ source) and then calculates their intersection. Afterwards, some heuristics (which must be tuned according to the task) are used to grow the alignments and cover all the words. The word alignment induced phrases are extracted by grouping word pairs that maximise the total translation probability. This grouping must respect the following constraint: if two phrases are aligned, all words present in these phrases must also be aligned. The probabilities for each bi-phrase are simply their relative frequencies weighted according to the relative frequencies of their individual words (lexical probabilities).

Besides the features represented in the phrase table, Moses creates a generation model and a reordering model from the parallel data. Additionally, a target n-gram language model is required, and can be built using a software like the SRILM toolkit (Stolcke 2002). The features coming from these different models (translation, generation, reordering and target language) are joined using a log-linear combination in which each feature f_k has a weight λ_k. Hence, it is necessary to estimate the optimal value for these weights. At this point, Moses uses a minimum error rate training (MERT) algorithm with a modified version of the NIST measure (Shinozaki and Ostendorf 2008). The optimisation of the λ_k weights is called *tuning* in the PB-SMT jargon, and requires a held-out *tuning set* of about one to two thousand parallel sentences.

7.1.3 Decoding

The translation of a new sentence is called *decoding* because of the origins of MT, as an analogy with cryptography (Weaver 1955). It is the decoder that is responsible for the actual translation of an unseen source sentence. This consists of choosing the bi-phrases that cover the source sentence and maximise the joint translation probabilities considering

2. The term *phrase* is used here to denote any sequence of words, in opposition to its standard use in linguistic to denote a well formed linguistic constituent.

all features f_k weighted by the w_k coefficients.

The decoding process is a search problem in a huge search space. In a phrase-based model, this space consists of all possible replacements of source phrases by target phrases until the source sentence is completely covered. Each step in this process, that is, each incomplete translation, is called a *hypothesis*. Each hypothesis has a probability that depends on the previous ones. When two hypotheses arrive by different paths at the same sequence of target words, they are unified. The best translation is the terminal hypothesis where all words are covered and the probability is maximised.

Knight (1999) proves that decoding is an NP-complex problem. He reduces the Hamiltonian cycle and the minimal coverage subset problems, both known NP-complete, to the decoding step of the simplest IBM model (model 1). Additionally, he shows that finding an optimal path in the possible translations graph is analogous to finding an optimal path in the travelling salesman problem, highlighting the importance of improvements in the resolution of such theoretical problems.

In such cases, as no exact solution can be calculated in reasonable time, approximative algorithms exist. In practice, decoding uses search heuristics such as A* and beam search (Tillmann and Ney 2003). The Moses decoder implements a beam search algorithm in the space of possible translations. This means that hypotheses are put into stacks and organised according to the number of source words covered. Only potentially good hypotheses will be further extended and the stacks are pruned on their size. Possible pruning techniques are histogram pruning and threshold pruning. Histogram pruning limits the size of each stack and allows the definition of a different threshold for each stack, depending on the number of covered words or on their position. Threshold pruning only keeps hypotheses whose current probability is not inferior to x times the best hypothesis already generated.

7.1.4 Evaluating

The evaluation of machine translation has been a very active research topic for many years, and still seems to be an open problem. Subjective evaluation relies on human judgements. There are many subjective measures, such as readability, fidelity, grammaticality and usability. Since the advent of large MT evaluation campaigns, two objective measures have been particularly popular: adequacy and fluency. Adequacy is the amount of meaning transferred from the source sentence to the target sentence, and fluency is the naturalness and grammaticality of the target sentence. Objective evaluation metrics can involve human-related factors such as average reading or post-editing times. Very often, when it comes to empirical MT systems, evaluation is performed automatically by comparing the automatic translation with a (set of) reference translation(s) proposed by human translators. Several measures exist for calculating the similarity between automatic and reference translations. Two of the most popular evaluation measures used in current SMT technology are BLEU and NIST.

The BLEU measure consists of two parts: a brevity factor *BP* that penalises too short translations, and a weighted n-gram difference with some constraints that avoid multiple counts (Papineni et al. 2002). When considering a single reference [3] sentence of length r

3. For multiple references, r is the length of the reference sentence that is closest to the length of the target sentence.

and a target automatically translated sentence of length t, BLEU is defined as:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} \lambda_n \log P_n\right)$$

The brevity penalty BP equals 1 if the target sentence is longer than the reference sentence, and $\exp(\frac{1-r}{t})$ else, strongly penalising very short translations while accepting slightly shorter ones. An equivalent formulation is $BP = \exp(min(0, \frac{1-r}{t}))$. The weight of every n-gram precision λ_n is the constant $1/N$. The P_n term, summed over a finite number of n ($N = 4$) is an n-gram proportion calculated as follows:

$$P_n = \frac{\sum_{i=1}^{t-n+1} c_{clip}(w_i^{i+n-1})}{\sum_{j=1}^{t-n+1} 1}$$

That is, we count how many of the n-grams of the target sentence appear in the reference ($c_{clip}(w_1^n)$) and divide it by the total number of n-grams in the target sentence. The value $c_{clip}(w_1^n)$ is clipped by the maximum count of w_1^n in any reference, to avoid that an n-gram is counted more than once. The interpretation of the BLEU measure as a probability is constrained to statistical significance considerations. (Och 2005) states that, for a corpus of 20K tokens, a difference of less than 1% in BLEU is not significant.

The NIST measure is a variation of BLEU where n-grams are weighted according to their informativeness. NIST does not iterate over sentences, but over the whole test corpus (Doddington 2002). It is calculated as:

$$NIST = BP \times \sum_{n=1}^{N} \left(\frac{\sum_{i=1}^{t-n+1} \text{info}(w_i^{i+n-1})}{\sum_{i=1}^{t-n+1} 1} \right)$$

The numerator sum runs over the n-grams co-occurring in the target and reference test sets and the denominator sum counts all n-grams in the target corpus. The informativeness of an n-gram is based on the conditional probability of its occurrence given that the previous $n-1$ words occurred. The brevity penalty of NIST is also slightly different, once r is the average length of all sentences in all references and t the length of the entire test result.

Even though automatic evaluation measures are very popular in SMT, in our experiments we will not rely on measures such as BLEU and NIST. MWEs are a complex phenomenon and their translation cannot be evaluated automatically. Instead, we will perform a careful manual evaluation of the targeted phenomenon, that is, the translation of phrasal verbs from English into Portuguese.

7.2 MWEs and SMT

In current MT systems, various practical solutions have been implemented. The expert MT system ITS-2 handles MWEs at two levels (Wehrli 1998). Contiguous compounds are dealt with during lexical analysis and treated as single words in subsequent steps. Idiomatic, non-fixed units are treated by the syntax analysis module, requiring a much

more sophisticated description. Once they are correctly identified, however, their transfer is executed in the same way as regular structures. Recently, the system implements a more sophisticated approach for non-fixed MWE identification in the syntactic analysis module (Wehrli et al. 2010). When evaluated on a data set of English/Italian→French translations, it improved the quality of 10% to 16% of the sentences.

The Jaen Japanese–English MT system was enriched with MWE rules by Haugereid and Bond (2011). Jaen is a semantic transfer MT system based on the HPSG parsers JACY and ERG. The authors use GIZA++ and Anymalign in order to generate phrase tables from parallel corpora, from which they automatically extract the new transfer rules. These rules are then filtered and, when added to the system, improve translation coverage (19.3% to 20.1) and translation quality (17.8% to 18.2%). Even though the improvements are quite modest, the authors argue that they can be further improved by learning even more rules.

An improvement of 33% in the French–Japanese translation of MWEs is obtained by Morin and Daille (2010). They implement a morphologically-based compositional method for backing-off when there is not enough data in a dictionary to translate a MWE. For example, *chronic fatigue syndrome* can be decomposed as *[chronic fatigue] [syndrome]*, *[chronic] [fatigue syndrome]* or *[chronic] [fatigue] [syndrome]*.

The translation of noun compounds from German and Spanish into English was addressed by Grefenstette (1999). He uses web counts to select translations for compositional noun compounds, and achieves an impressive accuracy of 0.86–0.87. Similarly, Tanaka and Baldwin (2003) compare two shallow translation methods for English–Japanese noun compounds. The first one is a static memory-based method where the compound needs to be present in the dictionary in order to be translated correctly. The second is a dynamic compositional method in which alternative translations are validated using corpus evidence. Their evaluation considers the compounds as test translation units (as opposed to traditional sentence-based evaluation). When they combine the two methods, they achieve 95% coverage and potentially high translation accuracy. This method is further refined by the use of a support vector machine model to rank all possible translations (Baldwin and Tanaka 2004). The model learns the translation scores based on several features coming from monolingual and bilingual dictionaries and corpora. Their method significantly outperforms previous methods and is particularly robust to low-frequency compounds.

The more-than-popular empirical MT system Moses represents MWEs as flat contiguous sequences of words (Koehn et al. 2007). Bilingual MWEs are bilingual sequences, called "bi-phrases", and have several probabilities but no linguistic information associated to them. Two complementary strategies have been adopted to add monolingual MWEs from WordNet into an English–Arabic Moses system (Carpuat and Diab 2010) . The first strategy is a static single-tokenisation that treats MWEs as word-with-spaces. The second strategy is dynamic, adding to the translation model a count for the number of MWEs in the source part of the bi-phrase. They found that both strategies result in improvement of translation quality, which suggests that Moses bi-phrases alone do not model all MWE information.

Another approach for minimizing data sparseness is the generation of monolingual paraphrases to augment the training corpus (Nakov 2008a). The basis for generating paraphrases that are nearly-equivalent semantically (e.g., *ban on beef import* for *beef import ban* and vice versa) are the parse trees. They are syntactically transformed by a set of heuristics, looking at noun compounds and related constructions. Using Moses'

ancestor, Pharaoh, on Spanish–English data, this technique generates an improvement equivalent to 33%-50% of that of doubling training data.

Automatic word alignment can be more challenging when translating from and to morphologically rich languages. In German and in Scandinavian languages, for instance, a compound is in fact a single token formed through concatenation of words and special infixes (*Hauptbahnhof* is the concatenation of *haupt (main), bahn (railway) and hof (station)*). Stymne (2011) develops a fine-grained typology for MT error analysis which includes concatenated definite and compound nouns. For definiteness, she makes the source text look more like the target text (or vice versa) during training, thus making the learning less prone to errors by using better word alignments. In Stymne (2009), she describes her approach to noun compounds, which she splits into their single word components prior to translation. Then, after translation, she applies some post-processing rules like the reordering or merging of the components.

Pal et al. (2010) explore the extension of a Moses English–Bengali system. Significant improvements are brought by applying preprocessing steps like single-tokenisation for named entities and compound verbs. However, larger improvements (4.59 absolute BLEU points) are observed when using a statistical model for the prior alignment of named entities, allowing for their adequate transliteration.

The domain adaptation of general-purpose MT systems can also be accomplished with the integration of multiword terms. Ren et al. (2009) adapt a Chinese–English standard Moses system using three simple techniques: appending the MWE lexicon to the corpus, appending it to the phrase table, and adding a binary feature to the translation model. They found significant BLEU improvements over the baseline, especially using the extra feature.

In translation memory systems such as Similis, the translation unit can be considered as a MWE as it is an intermediary between words and sentences. The correspondences of word sequences are automatically learned from the translation memory and expressed in a multi-layer architecture including surface forms, lemmas and parts of speech (Planas and Furuse 2000).

Hierarchical and tree-based translation systems like Joshua use tree rewriting rules in order to represent the correspondences between source and target structures (Li et al. 2009). However, it is difficult to implement special rules for MWEs and to distinguish them from rules that should be applied to ordinary word combinations. Promising results in the application of MWE resources such as lexicons and thesauri show that this is a recent and apparently growing topic in the MT community.

7.3 Integration into a PB-SMT system

This section starts with a discussion of the diversity of phrasal verbs in English and some related work (Section 7.3.1). Then we present the proposed methods for integrating PVs into a PB-SMT system (Section 7.3.2). We evaluate them on a baseline English–Portuguese PB-SMT system using automatic and manual evaluation of a test data set (Section 7.3.3). We finish with a discussion about the impact of our work on current MT technology, followed by conclusions and future work (Section 7.3.4).

7.3.1 Phrasal verbs in English

The translation of PVs is a challenging problem because they present a wide range of variability both in terms of syntax and semantics. Syntactically, they are combinations

of verbs with prepositions or adverbs, like the *verb-particle constructions* (VPCs) *put off* and *move on*, and the prepositional verbs *talk about* and *wait for*. The latter are syntactically rigid, usually selecting particular prepositions and requiring a complement after the preposition (Lohse et al. 2004). VPCs, however, can occur in different valency frames and in different word orders, in a joint (*make up NP*) or split configuration (*make NP up*). Moreover, as particles in English tend to be homographs with prepositions (*up, out, in*), a verb followed by a particle may be ambiguous between a VPC, a prepositional verb (e.g., *rely on*) and a verb followed by a prepositional phrase (e.g., *[eat up] [the chocolate]* and *[eat] [at the party]*). This affects how they are to be identified, interpreted, and, consequently, translated.

Even if "it is often said that phrasal verbs tend to be rather 'colloquial' or 'informal' and more appropriate to spoken English than written" (Sinclair 1989, p. iv), PVs are pervasive and appear often in all language registers. In our training corpus of speech transcriptions, for instance, around 17% of the sentences contained at least one detected PV.

PVs are challenging not only for computational systems but also for English language learners. According to the COLLINS-COBUILD dictionary (Sinclair 1989), which contains more than 3,000 PVs and more than 5,000 meanings, PVs are difficult for English learners because:

- they are often non-compositional, that is, even if a learner knows the meanings of *make* and of *out*, he/she cannot infer the meaning of *make out*;
- they have idiosyncratic grammatical behaviour with respect to object and adverb positioning;
- they often present strong collocational attachment to other elements (in other words, they are recursive MWEs);
- their number is constantly increasing, even though new PVs are not randomly coined but are mostly derived from productive patterns;
- they can often be replaced by a single-verb paraphrase, but sometimes the result may sound unnatural or pompous.

In terms of semantics, PVs can be described according to a three-way classification based on the predictability of their meaning from their parts (Bolinger 1971):

1. literal or compositional, like *take away*, *fight back* and *come out* $\rightarrow$ *leave*;
2. semi-idiomatic or aspectual, like *carry up*, *eat up*, *spread out* and *link up*;
3. idiomatic combinations, like *tell off* $\rightarrow$ *reprimand* and *go off* $\rightarrow$ *explode*.

Literal PVs are combinations in which both, the verb and the particle, keep the original meaning. However, there is overwhelming statistical evidence of their co-occurrence. In semi-idiomatic PVs, the meaning of the particle adds to the verb a sense of motion-through-location (*carry something up*) and of completion or result (*eat something up*). In other words, in semi-idiomatic PVs, the particle does not change the meaning of the verb, but is used to suggest that the action described by the verb is performed thoroughly, continuously or completely. Semi-productive patterns can be found in literal and semi-idiomatic combinations, for example, verbs of cleaning and the aspectual *up*. For idiomatic cases, on the other hand, it is not possible to straightforwardly determine their meanings by interpreting their components literally.

However, the borders between these classes are fuzzy. According to the COLLINS-COBUILD dictionary, "there is a general shading of meaning from one extreme to the other, but it is possible to point out four main types of combinations of verbs with particles" (Sinclair 1989). That is, PVs range from reasonably predictable constructions to

highly unpredictable ones, and in the middle there are reasonably predictable PVs reinforced by habitual collocation. In addition to Bolinger's three classes described above, the COLLINS-COBUILD dictionary includes a fourth type of PV: prepositional verbs. In prepositional verbs, even though the meaning of the expression is completely compositional, the verb is always used with a particular preposition or adverb and is normally not found without it, like *refer to* and *rely on*.

It may not be straightforward to model syntactic and semantic characteristics like these in SMT systems. However, it is important for a MT system to identify them and have an adequate treatment for them to avoid generating translations that sound unnatural or ungrammatical to native speakers, particularly for syntactically variable and idiomatic cases. The present work is one of the rare studies that look at PVs and address the question of their impact on MT systems, precisely because of the difficulties involved in such a study. For instance, using Google Translate, sentence 1 is translated into Portuguese as 1a, instead of the more natural and expected translation 2a, which is given when the VPC occurs in a joint configuration (English equivalents are provided in 1b and 2b):

1. *I will* **eat** *all the chocolate* **up**.
   (a) *Vou* **comer** *todo o chocolate* **para cima**.
   (b) *I will* **eat** *all the chocolate* **toward a higher position**.
2. *I will* **eat up** *all the chocolate.*
   (a) *Vou* ***comer*** *todo o chocolate.*
   (b) *I will* ***eat*** *all the chocolate.*

For the automatic identification of PVs, syntactic and semantic variability combined with association measures have resulted in an F-measure of 90.1% (Ramisch et al. 2008b). For PV tokens, an F-measure of 97.4% was obtained using syntactic and semantic information like the selectional preferences of the verb and of the PV (Kim and Baldwin 2010).

The compositionality of a PV may influence the performance of NLP tasks. In parsing, the identification of more idiomatic PVs, whose valency may differ from those of the simplex verb, is not problematic for a statistical parser like RASP. However, highly compositional cases (e.g., *call in*) may be less distinct syntactically from verb-prepositional phrase combinations, requiring additional information, such as selectional preferences (Kim and Baldwin 2010). Other methods for determining compositionality examine, for instance, the overlap between the synonym sets of the verb and the PV (McCarthy et al. 2003), or the extent to which the components of a PV contribute their simplex meanings to the interpretation of the PV (Bannard 2005).

7.3.2 Experimental setup

We have built a standard non-factored PB-SMT system using the open-source Moses toolkit, with parameters similar to those of the baseline system for the 2011 WMT campaign [4] (Callison-Burch et al. 2011). For training, we used a fragment of the English–Portuguese parallel Europarl v6 (EP) corpus. [5] The training data contains the first 200K sentences tokenised, lowercased and cleaned, resulting in 152,235 parallel sentences and around 3.1M tokens. The whole Portuguese EP, containing around 50M tokens, was used as training data for the 5-gram language model built with SRILM (Stolcke 2002).

4. `http://www.statmt.org/wmt11/baseline.html`
5. See Appendix D.

The controlled development and test sets were built using a random sample coming from two data sets of the Euromatrix project: the WMT 2008 test set [6], which contains 2,000 parallel sentences from the held-out portion of the EP corpus (not included in the training data), and the JRC-Acquis test set [7], with 4,107 sentences. Development set (500 sentences) and test set (1,000 sentences) contain 50% of sentences with at least one detected PV and 50% of PV-less cases. [8]

To annotate the PV tokens in the source corpora, we used a lexicon containing 2,168 PV types from the Phrasal Verb Demon [9], WordNet (Fellbaum 1998), and the English PVs dataset (Baldwin 2008). In the future, we intend to replace this lexicon by an automatically acquired one, using the `mwetoolkit`. We developed a joint and split PV detector using the jMWE library (Kulkarni and Finlayson 2011). It takes as input the corpora POS-tagged with the TreeTagger, and the English PV lexicon, searching for entries from the lexicon where verbs and particles are separated by at most 4 words, and ignoring cases where:

- *to* is the particle, due to the large number of mistagged infinitival cases;
- the verb is preceded by a determiner/possessive, as these are mistagged nouns;
- the verb and particle are split by the complementiser *that* or by another verb;
- a particle is shared by two PV candidates, in which case we consider only the PV with the rightmost verb, for example, ***take*** *this* ***depending on***;
- a verb is shared by two particles, in which case we consider only the PV with the leftmost particle, for example, *procedure* ***laid down in*** *Article*.

We also used a bilingual dictionary containing 3,224 English PVs and their equivalents in Portuguese, built from the Linguateca lexicon [10] and the Reverso dictionary [11] and manually validated by two Portuguese native speakers. As these only listed base forms, we generated inflected forms using RASP morphg [12] for English and the NILC dictionary [13] for Portuguese.

7.3.2.1 Integration strategies

We compare the following strategies for PV integration into the SMT system: TOK, PV?, PART, VERB and BILEX.

- TOK (or *single tokenisation*): before translation, the verb and the particle were detected and rearranged in a joint configuration (e.g., *call him up* into *call_up him*). Since it is represented as a single token with underscore, we expect to improve word alignment of the PV by handling it as a unit when preparing the SMT system.
- PV?: a binary feature is added to each bi-phrase indicating whether a source PV has been detected in it or not. This flag is subsequently used during decoding to inform the SMT system. [14]

6. `http://matrix.statmt.org/test_sets/test2008.tgz`
7. `http://matrix.statmt.org/test_sets/acquis.tgz`
8. PV-less sentences have been included to determine whether there are negative side-effects to the proposed approaches on sentences without PVs.
9. `http://www.phrasalverbdemon.com`
10. `http://linguateca.pt/Repositorio/RecursosLogos`
11. `http://dictionary.reverso.net`
12. `http://www.cogs.susx.ac.uk/lab/nlp/carroll/morph.html`
13. `http://www.nilc.icmc.usp.br/nilc`
14. This differs from the feature adopted by Carpuat and Diab (2010) in that the flag used here does not record the number of PVs per bi-phrase, but only whether the bi-phrase contained at least one PV.

– PART: to avoid marked prepositional verbs in the target sentence due to unusual selection of preposition (e.g., *work on* translated as *trabalhar ?sobre/em*, literally *work ?about/in*), the latter is replaced by the one most frequently used with the target verb. The new preposition $p*$ replaces the generated target preposition p — which occurs in position k in the translated sentence — according to the formula:

$$p* = \underset{p_i \in P}{\text{argmax}} \times \sum_{j=1}^{3} \frac{c(w_{k-j} \ldots p_i \ldots w_{k+j})}{3}$$

We retrieve Google [15] hit counts $c(\cdot)$ for all p_i in a set of possible prepositions P, averaging the context in a symmetric window around p_i (1 to 3 words).

– VERB: the tense of the verb (gerund or infinitive) is modified in the target sentence according to the tense detected on the source side, avoiding incorrect verbal inflections.
– BILEX (or *bilingual lexicon*): the phrase table of the baseline system is augmented with 179,133 new bilingual phrases generated from the 3,224 bilingual entries of the English–Portuguese PV lexicon and their possible inflections of source and target verbs. Due to the lack of estimates for translation and lexical probabilities, all translations scores were uniformly set to 1.

The pre-processing strategy TOK is uniformly applied to training, development and test sets. Post-processing strategies like PART and VERB are directly applied to the baseline translation of the test set. Strategies PV? and BILEX are also applied to the baseline translation model, but required re-tuning lambda weights with MERT.

7.3.3 Results

In this section, we first evaluate the identification of PVs in the source text (Section 7.3.3.1). Then, we analyse the baseline PB-SMT system and the PV integration strategies in terms of automatic measures (Section 7.3.3.2), in terms of manual annotation (Section 7.3.3.3), and in terms of the compositionality of the PVs (Section 7.3.3.4).

7.3.3.1 Detection of PV tokens

A set of 100 English sentences was manually annotated by two human judges with respect to correct detection of PVs: 50 PV and 50 PV-less sentences. Error analysis of incorrect detection indicated the following causes:

– for false positives:
    1. ambiguous prepositional phrase attachment or prepositional phrases as part of another expression (13%);
    2. POS tagger error (2%);
– and for false negatives:
    1. missing lexical entries in the PV lexicon (5%).

According to this manual evaluation, precision of the automatic PV detection is 68%, recall is 87% and F-measure is 76%. Most of the false positive instances were caused by ambiguity between particles and prepositions heading a prepositional phrase adjunct

15. We use Google counts to provide an additional source of information to that given by the LM and used for generating the translation.

	# PVs	# Verbs	# Part.
≥1	630	129	21
≥5	22	23	9

Table 7.2: Statistics of PVs in test corpus occurring more than once (≥1) and more than five times (≥5).

	PVs		No PVs		Total	
	bleu	nist	bleu	nist	bleu	nist
Baseline	0.258	6.54	0.270	6.35	0.262	6.79
TOK	0.250	6.44	0.268	6.30	0.256	6.70
PV?	0.256	6.55	0.273	6.38	0.262	6.81
PART	0.257	6.53	0.270	6.35	0.261	6.79
VERB	0.258	6.54	0.270	6.35	0.262	6.79
BILEX	0.253	6.47	0.270	6.35	0.259	6.75

Table 7.3: Translation of PVs in sentences — automatic evaluation. 500 sentences containing PVs, 500 sentences with no PVs and total 1,000 sentences.

(e.g., **ask** *you* **in** *all seriousness*). These results are expected, as we use only shallow information about sequences of words and parts of speech for detection. More detailed distributional and linguistic information (parsing) would be required for achieving a higher accuracy.

7.3.3.2 Automatic MT evaluation (BLEU/NIST)

The test set consists of a 1,000-sentence sample, half of them containing PVs. Statistics about PV occurrences in the test set are summarised in Table 7.2. Only few PVs and verbs occur more than 5 times. The most frequent combinations include *lay down, set up, carry out* and *originate in*.

To quantify the impact of the strategies for PV integration, we use as baseline the standard results obtained with Moses on this test set. For the automatic evaluation of the translation quality comparing the obtained Portuguese MT results with reference translations, we use BLEU and NIST (see Section 7.1.4). Table 7.3 presents the metric results for the baseline and for the integration strategies.

As the strategies transform the sentences locally around the PV, automatic measures vary slightly. Thus, it is impossible to determine whether these results are due to the strategies or to noise in the test sample. A consistent trend indicated by the metrics is that, according to BLEU, sentences without PVs get better n-gram precision than those with PVs, but, the informativeness of n-grams increases in sentences containing PVs, according to NIST. In addition, we can assume that the strategies have no negative impact on PV-less sentences.

BLEU and NIST metrics are not always an adequate way to evaluate MT quality, especially concerning a complex linguistic phenomenon like PVs. For the sake of comparability of the results with the state of the art, we included these results in our analysis. However, they are not the focus of these experiments, and no further conclusions can be

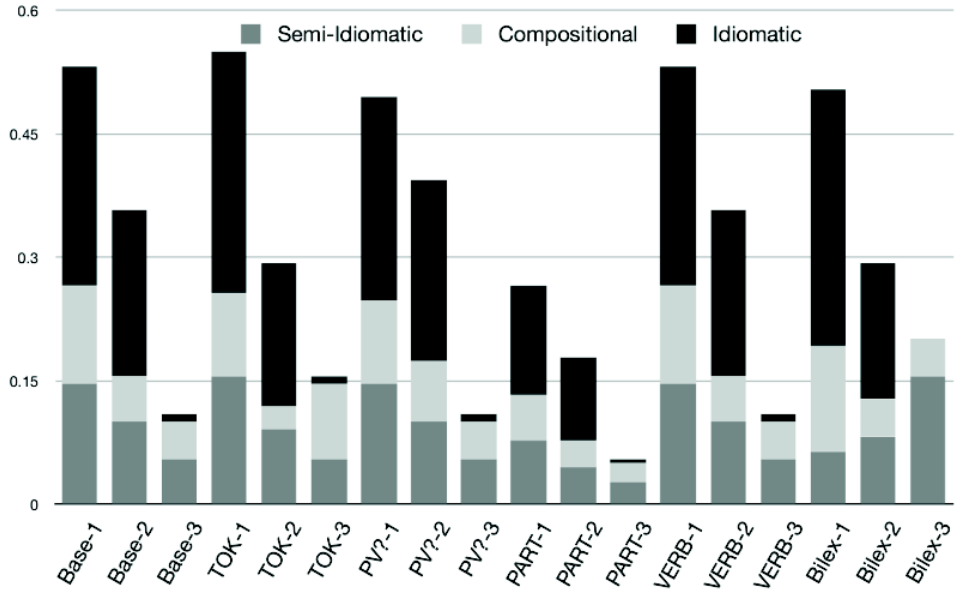


Figure 7.1: PV semantics and quality scores per system. Scores are (1) good, (2) acceptable and (3) incorrect translation.

drawn from them.

7.3.3.3 Human analysis of translations

Since we are dealing with complex phenomena, BLEU/NIST scores are necessarily complemented by a manual error analysis of the results obtained by the various strategies for a more in-depth analysis of the factors that affect translation quality. The quality of the translation was evaluated by a Portuguese native speaker with good knowledge of PVs. The evaluation considered a limited context of the phrases used in the translation of the verb and of the particle, other parts of the sentence being ignored. Possible scores were 1 for correct translation, 2 for acceptable translation, but where the verbal inflection or particle selection could be improved, and 3 for incorrect translation that modifies the meaning of the sentence; the results are summarised in Table 7.4. We report results of manual annotation only for correctly identified PVs. Here, our goal is to examine the translation of the identified PVs and not the identification task itself. Therefore, errors in the automatic identification of PVs do not impact evaluation as these cases were manually identified and left out of the test set for human annotation.

The strategy that improves the amount of correct translations in comparison with the baseline is TOK. It was useful in cases of split VPCs where the links between the particle

	% good (1)	% acceptable (2)	% incorrect (3)
Baseline	0.53	0.36	0.11
TOK	0.55	0.29	0.16
PV?	0.50	0.39	0.11
PART	0.53	0.36	0.11
VERB	0.53	0.36	0.11
BILEX	0.50	0.29	0.20

Table 7.4: Translation of PVs — human evaluation.

and the verb were not captured by the baseline, and individual translations for the words produced inadequate results (e.g., *take something away* as *seize something finish)*.

Although the BILEX system improved the translation of some difficult cases for the baseline, including one of inversion and stranding, the entries of the lexicon were not ranked by frequency or sense usage. As a consequence, the lexical choice was essentially performed by the language model and, in many cases, the translation involved an unusual sense of the word that changed the meaning of the sentence. However, in the phrase tables used by the other strategies, the phrases contain accurate probabilities according to training data. The availability of estimators for the translation probabilities for the bilingual lexicon entries would allow a fairer comparison and possibly improved results.

The PART post-processing strategy was expected to improve results in cases where the verb selects a very specific (group of) particle(s) in the target language, that may not correspond to those used in the source. However, for this test corpus, it did not seem to alter the results of the baseline, producing variations that were as acceptable as those produced by the baseline.

7.3.3.4 Human analysis of compositionality

To investigate if there is a correlation between translation quality and the semantics of the source PVs, they were further annotated according to 3 classes: compositional (22%), semi-idiomatic (30%) and idiomatic combinations (48%). Figure 7.1 shows the performance of each system per translation quality score in terms of the semantics of the PVs.

The strategies that produced best results for idiomatic cases (Table 7.5), which account for almost half of the PVs in the test set, were TOK and BILEX. For these cases, good quality translation depends on PVs being treated as a unit, and a word by word literal translation of such opaque cases would produce incorrect results. Therefore, the availability of dedicated wide-coverage resources is a significant factor for performance on idiomatic PVs. The use of a bilingual lexicon, in particular, resulted in no idiomatic cases incorrectly translated (BILEX-3 in Figure 7.1), and for most of the acceptable translations the correct verb was used. This was also the best approach for compositional PVs, accounting for more successful translations for these two semantic types than the baseline.

For the compositional cases, TOK resulted in a decrease in performance. Some of the problems were caused by a usage of the PV different from the one appropriate for the sentence (e.g., *call for* treated as part of *call for papers*, due to its high frequency). Such problems may arise from changes in the frequency counts from treating PVs as single tokens. Although the PV? system tends to translate PVs as units, it is softer than the TOK system which *always* translates them as such, producing fewer incorrect compositional translations than the latter (TOK-3 vs PV?-3 in Figure 7.1).

For semi-idiomatic PVs, single tokenisation also improves over the baseline. In contrast, the use of a bilingual lexicon reduced significantly the quality of the translation. This may be a consequence of the lack of frequency information for the entries in the bilingual lexicon and customisation to the domain (e.g., *set out* is a frequent PV, translated as *aim* instead of *define*). Indeed, for most systems, incorrect translations were correlated with lower average frequencies (41.07 in average for score 3 vs 144.3 for score 1), but this was not found for BILEX (130.45 in average for 3 vs 141.67 for 1).

	% good (1)	% acceptable (2)	% incorrect (3)
Baseline	0.56	0.42	0.02
TOK	0.63	0.35	0.02
PV?	0.54	0.44	0.02
PART	0.56	0.42	0.02
VERB	0.58	0.40	0.02
BILEX	0.63	0.37	0

Table 7.5: Translation of idiomatic PVs — human evaluation.

7.3.4 Discussion

We presented the results of ongoing experiments, in which we performed an in-depth analysis of the effects of PV handling in a SMT system. Related work on SMT has looked at other MWE types, like named entities and compound words, but there is very little work on verbal expressions. Most of the models proposed so far for the integration of MWEs into SMT systems only deal with MWEs that are exclusively sequences of contiguous words on both the source and the target side. More sophisticated, next-generation translation methods need to acknowledge the significant role that MWEs play in language. Therefore, they need to be able to translate not only 100% compositional and 100% rigid sequences, but also expressions which, like PVs, have a variable degrees of syntactic flexibility and, as a consequence, of compositionality.

In our experiments, common problems with the translation across the systems involved the following cases:

- verbal inflection mismatches including cases of gender, number and person, as Portuguese can have 52 different forms, not including verb clitics, which can significantly increase this number (e.g., *encontrá-la-ei*, literally *meet her I will*);
- the particle is not the one commonly required by the target verb;
- the source verb is translated as a target noun (e.g., *conclude* as *conclusão*, literally *conclusion*);
- the preposition in the target language should have been omitted, or is marked by being stranded at the end of the sentence.

The comparison of these heuristics indicates that they provide complementary strengths, which seem to be linked to compositionality and frequency. Here is one example of translation where the baseline has got it wrong, but one of the systems improved:

- *. . . the rural population represents 38% of the total population and* **accounts for** *4.9% of the gdp.*
- *. . . a população rural representa 38% da população total e* **contas de** *4,9% do pib,* (baseline)
- *. . . a população rural representa 38% da população total ,* **sendo responsável por** *4,9% do pib,* (TOK)

In this case, TOK has been the only strategy to get the PV translation right, with the baseline interpreting the verb as a noun (*accounts* → *contas*). A system that can detect PVs and identify their token semantics could adopt a targeted treatment whereby compositional cases would be treated by the use of a bilingual lexicon, semi-idiomatic cases by single tokenisation pre-processing and idiomatic cases would be dealt with by both.

7.4 Summary

As a second evaluation of the `mwetoolkit`, we performed experiments on the translation of English phrasal verbs (PVs) like *give up* and *get by [a name]* into Portuguese, using an empirical MT system. The translation of PVs is a challenge because they present a wide syntactic and semantic variability. PVs are very frequent and appear often in English, occurring in about 17% of the sentences of our corpus. Modelling the complex syntactic and semantic behaviour of PVs using the flat contiguous word sequences of current empirical MT systems is not straightforward. Nonetheless, it is important to identify them and have an adequate treatment for them to avoid generating translations that sound unnatural or ungrammatical.

The representation and integration of MWEs into machine translation systems has been the focus of considerable research. The ITS-2 MT system processes MWEs at two levels: during lexical analysis for contiguous compounds, and during syntactic analysis for collocations (Wehrli 1998, Wehrli et al. 2010). Carpuat and Diab (2010) adopt two complementary strategies for integrating MWEs: a static strategy of single-tokenisation that treats MWEs as word-with-spaces and a dynamic strategy that adds a count for the number of MWEs in the source phrase. Morin and Daille (2010) obtained an improvement of 33% in the French–Japanese translation of MWEs with a morphologically-based compositional method for backing-off when there is not enough data in a dictionary to translate a MWE. For translating from and to morphologically rich languages like German, where a compound is in fact a single token formed through concatenation, Stymne (2011) splits the compound into its single word components prior to translation and then applies some post-processing, like the reordering or merging of the components, after translation. Another approach for minimizing data sparseness is adopted by Nakov (2008a), who generates monolingual paraphrases to augment the training corpus.

In our experiments, a standard non factored phrase-based SMT system was built by training a Moses system with standard parameters on the English–Portuguese Europarl v6 corpus. Phrasal verbs were automatically identified using the jMWE tool and a dictionary of PVs. We compared the five strategies for the integration of automatically identified phrasal verbs in the system. The test set consists of a 1,000-sentence sample, half of them containing PVs. The most frequent constructions include *lay down, set up, carry out* and *originate in*.

Since we are dealing with a complex linguistic phenomenon, none of our conclusions could be drawn solely from automatic measures like BLEU and NIST, without careful error analysis through human evaluation of translation outputs. Common problems with the translation across the systems involved verbal inflection mismatches, wrong particle/preposition selection, translation of a verb as a noun and spurious prepositions being added to the target verb. The preliminary results of human evaluation performed on a test set of 100 sentences showed that, while some translations improve with the integration strategies, others are degraded. No absolute improvement was observed, but we believe that this is due to the fact that our evaluation needs to consider more fine-grained classes of phrasal verbs instead of mixing them all in the same test set. Additionally, we would need to annotate more data in order to obtain more representative results.

We discovered that there is a correlation between the quality of the translations given by each strategy and the compositionality of the PVs. The strategies that produced best results for idiomatic cases were TOK and BILEX. For the compositional cases, TOK resulted in a decrease in performance. Although the PV? strategy tends to translate PVs as units, it is softer than TOK, producing fewer incorrect compositional translations. The compar-

ison of these heuristics indicates that they provide complementary strengths, which seem to be linked to compositionality and frequency. These hypotheses motivate us to continue our investigation in order to obtain a deeper understanding the impact of each integration strategy on each step of the SMT system.

8 CONCLUSIONS

This chapter summarises the work presented in previous chapters and describes the current and future directions of our research. It is organised in three sections: thesis achievements (Section 8.1), ongoing experiments (Section 8.2) and future perspectives (Section 8.3).

8.1 Thesis achievements

We started our work by presenting its motivations, trying to answer three questions: what are MWEs, why do they matter and what happens if we ignore them? Through many examples, we illustrated the vagueness of the concept of MWEs concerning a large number of constructions in everyday language, like idioms, phrasal verbs and noun compounds. Due to the ubiquitous nature of MWEs, NLP applications dealing with real text should provide adequate MWE treatment, otherwise they will fail in generating high-quality natural output.

We have presented some theoretical approaches to MWEs, such as constructionism and meaning-text theory. There are many definitions for the term multiword expressions, but we chose to adopt a generic one that considers MWEs as word combinations that, at some point of linguistic processing, must be treated as a unit. This allowed us to discuss some important characteristics of MWEs such as arbitrariness, heterogeneity, recurrence and limited semantic/syntactic variability. Additionally, a MWE taxonomy can be useful when evaluating the acquisition, and we suggested a new one based on the morphosyntactic role of the MWE in a sentence and the difficulty to deal with it using computational methods.

MWE acquisition methods often use a common set of linguistic and statistical tools such as analysis software, word frequency distributions, *n*-gram language models and association measures. Therefore, we provided a brief overview of these foundational concepts before reviewing related work in MWE acquisition. Other tasks concerning MWE treatment, namely interpretation, disambiguation, representation and applications have also been illustrated. An important contribution of this thesis is this broad and deep review of the state of the art. This constitutes a significant step toward the consolidation of MWEs as a field in NLP.

In Section 1.2, we have described three main goals for our work, which we recall here:

1. To develop techniques for automatic MWE acquisition from corpora.
2. To evaluate them extrinsically by measuring their usefulness in NLP applications.
3. To investigate their acquisition and integration in multilingual contexts.

At the current state of research, it is safe to state that goal number one can be considered as achieved. A variety of combinations of languages, domains and types of MWE were investigated, and this analysis provided foundational knowledge about the behaviour of MWEs in texts. Languages whose writing systems have no word separators (such as Chinese, Japanese, Korean, Thai, Laotian and Khmer) have not been experimented with, but `mwetoolkit` can handle them as any other language, once texts are preprocessed by one of the numerous word segmenters available. The resulting software tool, the `mwetoolkit` is freely available. [1]

The evaluation of MWE acquisition being an open problem, we have proposed a theoretical framework which hopefully will shed some light on a possible structure for describing the problem. As for the extrinsic evaluation goals, we have demonstrated the usefulness of our methodology in the development of three different lexical resources. In addition, there are other applications of the `mwetoolkit` that were not included in the thesis for lack of space (Villavicencio et al. 2012, Granada et al. 2012). Therefore, our second goal can also be considered as achieved.

Finally, concerning the third goal, we provided preliminary results on the integration of MWEs into an empirical MT system. This is still work in progress, and further experiments, improvements and extensions are planned as future work, as described in Section 8.2.

8.2 Ongoing experiments

Ongoing experiments follow two parallel directions. First, we are trying to obtain better results and deeper understanding of the results obtained in Chapter 7, about the integration of MWEs into an empirical MT system. Second, we are actively participating in a project that aims at bringing together ontologies and lexical resources for mutual improvements in multilingual contexts.

8.2.1 MWEs and MT

The integration of phrasal verbs into a SMT system, described in Chapter 7, is a very hard problem due to the variability of these constructions. Our experiments have showed that, while a standard SMT system does get some of the phrasal verbs right (mostly joint compositional and semi-idiomatic phrasal verbs), it makes mistakes when the verb and the particle have an idiomatic interpretation and when they are split by intervening material.

Further research on a combination of integration techniques can potentially bring a unified solution to this problem, in particular in relation to adaptation and ranking of a bilingual lexicon to the domain of the corpus. Since we are dealing with a complex linguistic phenomenon, none of these conclusions could be drawn solely from automatic measures like BLEU and NIST, without careful error analysis through human evaluation of translation outputs.

Inserting entries directly into the phrase table is just one possible way of integrating a bilingual lexicon into the SMT system. Alternatively, one could estimate translation probabilities for them as done by Bouamor et al. (2011), [2] or use them to guide word alignment, to post-process the translation output based on the lexicon, or to append them

1. We intend to continue its maintenance, support and development in the future, as it is very important to improve its usability based on user feedback.

2. However, phrasal verbs are productive and it is not possible to preview all possible variations as in the case of other MWEs like compounds.

to the training corpus as artificial sentences. Inclusion strategies that do not force the use of the phrases from the lexicon through maximum probabilities would probably better handle compositional phrasal verbs than the current approach.

There is also room for improvement in phrasal verb detection. Better precision could probably be obtained with a deeper processing to capture longer dependencies between syntactically variable candidates, such as suggested in Seretan (2008), for instance, for general collocations, or in Baldwin (2005a) for phrasal verbs. Potentially, syntax information can provide additional features for the translation model. Also, idioms containing phrasal verbs like *put in place* or *put in order* are not treated by the current approach.

We showed that the strategies proposed here perform differently according to the compositionality of the phrasal verb. Therefore, corpus-based detection of compositionality in phrasal verbs (McCarthy et al. 2003, Bannard et al. 2003, Baldwin et al. 2003) could also help in generating more precise translations. For future work, we plan to investigate further crosslinguistic asymmetries and equivalences between languages. Our long term goal is to integrate MWE treatment into SMT systems in order to achieve high quality translation through the combination of statistical and linguistic information.

All of these hypotheses are being currently validated in a new set of experiments that features notably a new evaluation data set, constituted by careful profiling of the behaviour of phrasal verbs in our test corpus and their syntactic and distributional characteristics. We expect to publish the new results in an upcoming conference or as a journal paper.

8.2.2 CAMELEON project

One of the outcomes of the present thesis is the CAMELEON project, funded by CAPES-COFECUB grant 707-11. [3] The goal of this project is to investigate, propose, experiment, apply and validate automatic and collaborative techniques for the development of lexical and ontological resources that can be useful in the context of multilingual applications, particularly for French, Portuguese and English. The integration of automatic and collaborative methods has several advantages because they are somehow complementary. On the one hand, collaborative methods could use automatically generated data as a starting point, thus saving time and effort when creating a new instance (for a new language/domain/language pair). On the other hand, data-driven methods produce noisy results that should be later filtered by human experts. The use of collaborative platforms seems the most natural environment for post-editing automatically extracted lexical and ontological resources. Therefore, we would like to investigate the feasibility of using a system for collaborative management of lexical resources in order to filter and validate automatically acquired MWEs.

In the CAMELEON project, we have ongoing experiments and some first published results in related problems concerning the automatic acquisition of lexical information. We have built a comparable corpus in Portugues, English and French representing a sample of language in the conference organisation domain (Granada et al. 2012). Our goal is to use this corpus to support ontology-related tasks, such as multilingual ontology matching, extension, automatic ontology learning and population.

In parallel we are currently investigating the feasibility of an approach for the construction of lexical resources in Portuguese based on the serious lexical game JeuxDeMots [4] (Mangeot and Ramisch 2012). There are many potential applications for this resource, such as semantic role labelling and word sense disambiguation. There is also an

3. `http://cameleon.imag.fr`
4. `http://jeuxdemots.imag.fr/por`

interesting and open research problem concerning the multilingual alignment of lexical networks.

These experiments are inserted in the context of automatic lexical acquisition and constitute a continuation and extension of the work presented in this thesis. They are an interesting experimental set-up for future research in which MWEs play an important role.

8.3 Perspectives

One of the possibilities for future work comes from the fact that, for the moment, we were not able to develop further the acquisition of bilingual MWEs. In spite of some promising preliminary results (de Medeiros Caseli et al. 2010, Ramisch et al. 2010a, Villavicencio et al. 2010), we chose to focus our research on application-based evaluation instead of focusing on bilingual acquisition. There are many ideas in the drawer waiting to be put in practice. For instance, we would like to explore MWE acquisition from comparable corpora and from the web as a corpus. We would also like to investigate active learning or incremental methods to obtain cross-lingual correspondences for two monolingual MWE lists acquired independently from monolingual corpora. Related to our experiment with MT of asymmetric constructions, we would like to investigate techniques that explore cross-lingual asymmetries for bilingual acquisition. For instance, given that German compounds are concatenated together as single words, is it possible to detect their multiword counterparts in other languages automatically?

Some types of MWEs require more sophisticated, semantic information, in order to be correctly identified. This is the case, for instance, of phrasal verbs and idiomatic expressions. We have developed an integrated and stable experimental framework for MWE acquisition and evaluation, and we would like to extend it by implementing and developing new methods for the automatic interpretation and disambiguation of MWE semantics. Similarly, we believe that fine-grained syntactical information, such as syntactico-semantic valency frames, can help obtain more precise acquisition results. The drawback of using this kind of information is that the method becomes quite language-dependent. However, distributional methods inspired on their semantic counterpart could be a good trade-off between linguistic precision and generality.

We have tested the integration of MWEs into an empirical MT system. However, we argued that expert MT systems would be a more natural choice, and that they were not used simply because they it is not easy to obtain access to the source code and translation model of most expert systems. However, we still would like to investigate the integration of MWEs, not only of phrasal verbs but also of other types, into different MT paradigms. A possible solution would be to use the open source transfer-based system Apertium (Forcada 2009). The challenges in this case would be the adequate lexical representation of MWEs and the disambiguation between compositional and idiomatic occurrences.

Finally, we would like to apply extrinsic evaluation on other NLP applications. In the schedule of the CAMELEON project, there is a task planned for integrating automatically acquired MWEs into an information retrieval system. In Section 1.1, we listed several other applications that could benefit from MWE treatment. Therefore, there is much room for future research in this direction.

In spite of a large amount of work in the area, the treatment of MWEs in NLP applications is still an open problem, and a very challenging one! This is not surprising, given that the complex and heterogeneous nature of MWEs has been demonstrated by numer-

ous linguistic studies. At the beginning of the 2000's, Schone and Jurafsky (2001) asked whether the identification of MWEs was a solved problem, and the answer that this paper gave was: "no, it is not." More recent specialised publications show evidences that this is still the case. For instance, the preface of recent journal special issues on MWEs (Villavicencio et al. 2005b, Rayson et al. 2010b) and of the proceedings of the MWE workshops (Laporte et al. 2010, Kordoni et al. 2011b) list several challenges in MWE treatment such as multilingualism, lexical representation and application-oriented evaluation.

One of the main contributions of the present thesis is that it represents a significant step toward the full integration of automatically extracted MWEs into real-life NLP applications. However, given the complexity of the phenomenon, there is a constant need for improvements and it seems unlikely that, in the near future, a unified push-button solution will be proposed. Therefore, our long-term goal can be summarised as extending and improving the work presented here. If, on the one hand, a significant first step has been taken, on the other hand, there is still a long road ahead.

REFERENCES

Abreu, D. T. B. (2011). A semântica de construções com verbos-suporte e o paradigma Framenet. Master's thesis, São Leopoldo, RS, Brazil. 143 p.

Acosta, O., Villavicencio, A., and Moreira, V. (2011). Identification and treatment of multiword expressions applied to information retrieval. In Kordoni et al. (2011a), pages 101–109.

Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., and Urizar, R. (2004). Representation and treatment of multiword expressions in Basque. In Tanaka, T., Villavicencio, A., Bond, F., and Korhonen, A., editors, *Proc. of the ACL Workshop on MWEs: Integrating Processing (MWE 2004)*, pages 48–55, Barcelona, Spain. ACL.

Alsina, A., Bresnan, J., and Sells, P., editors (1997). *Complex Predicates*. CSLI Publications, Stanford, CA, USA. 514 p.

Anastasiadi-Symeonidi, A. (1986). *Neology in Modern Greek (in Greek)*. PhD thesis, Aristotle University of Thessaloniki.

Anastasiou, D., Hashimoto, C., Nakov, P., and Kim, S. N., editors (2009). *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)*, Suntec, Singapore. ACL. 70 p.

Apresian, J., Boguslavsky, I., Iomdin, L., and Tsinman, L. (2003). Lexical functions as a tool of ETAP-3. In *Proc. of the First MTT Conference (MTT 2003)*.

Araujo, V. D., Ramisch, C., and Villavicencio, A. (2011). Fast and flexible MWE candidate generation with the mwetoolkit. In Kordoni et al. (2011a), pages 134–136.

Athayde, M. F. (2001). *Construções com verbo-suporte (Funktionsverbgefüge) do português e do alemão*. Number 1 in Cadernos do CIEG Centro Interuniversitário de Estudos Germanísticos. Universidade de Coimbra, Coimbra, Portugal.

Atkins, S. (2010). The DANTE database: Its contribution to English lexical research, and in particular to complementing the FrameNet data. In de Schryver, G.-M., editor, *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*, Kampala, Uganda. Menha Publishers.

Atkins, S., Fillmore, C., and Johnson, C. R. (2003). Lexicographic relevance: Selecting information from corpus evidence. *International Journal of Lexicography*, 16(3):251–280.

Attia, M., Toral, A., Tounsi, L., Pecina, P., and van Genabith, J. (2010). Automatic extraction of Arabic multiword expressions. In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 18–26, Beijing, China. ACL.

Baayen, R. H. (2001). *Word Frequency Distributions*, volume 18 of *Text, Speech and Language Technology*. Springer.

Bai, M.-H., You, J.-M., Chen, K.-J., and Chang, J. S. (2009). Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In *Proc. of the 2009 EMNLP (EMNLP 2009)*, pages 478–486, Suntec, Singapore. ACL.

Baldwin, T. (2005a). Bootstrapping deep lexical resources: Resources for courses. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 67–76, Ann Arbor, Michigan. Association for Computational Linguistics.

Baldwin, T. (2005b). Deep lexical acquisition of verb-particle constructions. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):398–414.

Baldwin, T. (2008). A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 1–2, Marrakech, Morocco.

Baldwin, T. (2011). MWEs and topic modelling: Enhancing machine learning with linguistics. In Kordoni et al. (2011a), page 1.

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In Bond, F., Korhonen, A., McCarthy, D., and Villavicencio, A., editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 89–96, Sapporo, Japan. ACL.

Baldwin, T. and Tanaka, T. (2004). Translation by machine of complex nominals: Getting it right. In Tanaka, T., Villavicencio, A., Bond, F., and Korhonen, A., editors, *Proc. of the ACL Workshop on MWEs: Integrating Processing (MWE 2004)*, pages 24–31, Barcelona, Spain. ACL.

Baldwin, T. and Villavicencio, A. (2002). Extracting the unextractable: A case study on verb-particles. In Roth, D. and van den Bosch, A., editors, *Proc. of the Sixth CoNLL (CoNLL 2002)*, pages 98–104, Taipei, Taiwan. ACL.

Banerjee, S. and Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381, Mexico City, Mexico.

Bannard, C. (2005). Learning about the meaning of verb-particle constructions from corpora. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):467–478.

Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In Grégoire, N., Evert, S., and Kim, S. N., editors, *Proc. of the ACL Workshop on A Broader Perspective on MWEs (MWE 2007)*, pages 1–8, Prague, Czech Republic. ACL.

Bannard, C., Baldwin, T., and Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. In Bond, F., Korhonen, A., McCarthy, D., and Villavicencio, A., editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 65–72, Sapporo, Japan. ACL.

Baptista, J., Correia, A., and Fernandes, G. (2004). Frozen sentences of Portuguese: Formal descriptions for NLP. In Tanaka, T., Villavicencio, A., Bond, F., and Korhonen, A., editors, *Proc. of the ACL Workshop on MWEs: Integrating Processing (MWE 2004)*, pages 72–79, Barcelona, Spain. ACL.

Barreiro, A. and Cabral, L. M. (2009). ReEscreve: a translator-friendly multi-purpose paraphrasing software tool. In *Proceedings of the Workshop Beyond Translation Memories: New Tools for Translators, The Twelfth Machine Translation Summit*, pages 1–8, Ottawa, Canada.

Basili, R., Pazienza, M. T., and Velardi., P. (1994). A "not-so-shallow" parser for collocational analysis. In *Proc. of the 15th COLING (COLING 1994)*, pages 447–453, Kyoto, Japan.

Bergsma, S., Lin, D., and Goebel, R. (2009). Web-scale *N*-gram models for lexical disambiguation. In *IJCAI*, pages 1507–1512, Pasadena, California.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education Ltd, Harlow, Essex, 1st edition. 1204 p.

Bick, E. (2000). *The parsing system Palavras*. Aarhus University Press. 411 p.

Boitet, C., Bey, Y., Tomokio, M., Cao, W., and Blanchon, H. (2006). IWSLT-06: experiments with commercial MT systems and lessons from subjective evaluations. In *International Workshop on Spoken Language Translation (IWSLT 2006)*, Kyoto, Japan.

Bolinger, D. (1971). *The phrasal verb in English*. Harvard UP, Harvard, USA. 187 p.

Bond, F., Korhonen, A., McCarthy, D., and Villavicencio, A., editors (2003). *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, Sapporo, Japan. ACL. 104 p.

Bonin, F., Dell'Orletta, F., Montemagni, S., and Venturi, G. (2010a). A contrastive approach to multi-word extraction from domain-specific corpora. In *Proc. of the Seventh LREC (LREC 2010)*, Malta. ELRA.

Bonin, F., Dell'Orletta, F., Venturi, G., and Montemagni, S. (2010b). Contrastive filtering of domain-specific multi-word terms from different types of corpora. In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 76–79, Beijing, China. ACL.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2011). Improved statistical machine translation using multiword expressions. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011)*, pages 15–20, Barcelona, Spain.

Boulaknadel, S., Daille, B., and Aboutajdine, D. (2008). A multi-word term extraction program for Arabic language. In *Proc. of the Sixth LREC (LREC 2008)*, pages 1485–1488, Marrakech, Morocco. ELRA.

Bouma, G. and Moirón, B. V. (2001). Corpus-based acquisition of collocational prepositional phrases. In *Proc. of the Twelfth Conf. of CLIN (CLIN 2001)*, pages 23–37, Twente, Netherlands. CLIN.

Briscoe, T., Carroll, J., and Watson, R. (2006). The second release of the RASP system. In Curran, J., editor, *Proc. of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 77–80, Sidney, Australia. ACL.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Comp. Ling.*, 19(2):263–311.

Bu, F., Zhu, X., and Li, M. (2010). Measuring the non-compositionality of multiword expressions. In Huang, C.-R. and Jurafsky, D., editors, *Proc. of the 23rd COLING (COLING 2010)*, pages 116–124, Beijing, China. The Coling 2010 Organizing Committee.

Burnard, L. (2007). User Reference Guide for the British National Corpus. Technical report, Oxford University Computing Services.

Butnariu, C., Kim, S. N., Nakov, P., Séaghdha, D. O., Szpakowicz, S., and Veale, T. (2010). Semeval-2 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In Erk, K. and Strapparava, C., editors, *Proc. of the 5th SemEval (SemEval 2010)*, pages 39–44, Uppsala, Sweden. ACL.

Butt, M. (2003). The light verb jungle. In *Proceedings of the Workshop on Multi-Verb Constructions*, pages 243–246, Trondheim, Norway.

Cabré, M. T. (1992). *La terminologia. La teoria, els mètodes, les aplicacions*. Empúries, Barcelona, Spain. 527 p.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. F., editors (2011). *Proc.of the Sixth StatMT (WMT 2011)*, Ediburgh, Scotland. ACL.

Calzolari, N. and Bindi, R. (1990). Acquisition of lexical information from a large textual Italian corpus. In *Proc. of the 13th COLING (COLING 1990)*, pages 54–59, Helsinki, Finland.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., Macleod, C., and Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *Proc. of the Third LREC (LREC 2002)*, pages 1934–1940, Las Palmas, Canary Islands, Spain. ELRA.

Carpuat, M. and Diab, M. (2010). Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proc. of HLT: The 2010 Annual Conf. of the NAACL (NAACL 2003)*, pages 242–245, Los Angeles, California. ACL.

Carvalho, P., Sarmento, L., Teixeira, J., and Silva, M. J. (2011). Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proc. of the 49th ACL: HLT (ACL HLT 2011)*, pages 564–568, Portland, OR, USA. ACL.

Català, D. and Baptista, J. (2007). Spanish adverbial frozen expressions. In Grégoire, N., Evert, S., and Kim, S. N., editors, *Proc. of the ACL Workshop on A Broader Perspective on MWEs (MWE 2007)*, pages 33–40, Prague, Czech Republic. ACL.

Chakraborty, T. and Bandyopadhyay, S. (2010). Identification of reduplication in Bengali corpus and their semantic analysis: A rule-based approach. In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 72–75, Beijing, China. ACL.

Chakraborty, T., Das, D., and Bandyopadhyay, S. (2011). Semantic clustering: an attempt to identify multiword expressions in Bengali. In Kordoni et al. (2011a), pages 8–13.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Comp. Speech & Lang.*, 13(4):359–394.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *RIAO'88*, pages 609–624.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *COMPLEX 1994*, pages 23–32, Budapest, Hungary.

Church, K. (2011). How many multiword expressions do people know? In Kordoni et al. (2011a), pages 137–144.

Church, K. and Hanks, P. (1990). Word association norms mutual information, and lexicography. *Comp. Ling.*, 16(1):22–29.

Conejo, C. R. (2008). O verbo-suporte fazer na língua portuguesa: um exercício de análise de base funcionalista. Master's thesis, PPG de Letras, Universidade Estadual de Maringá, Maringá, PR, Brazil.

Constant, M. and Sigogne, A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. In Kordoni et al. (2011a), pages 49–56.

Cook, P., Fazly, A., and Stevenson, S. (2007). Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In Grégoire, N., Evert, S., and Kim, S. N., editors, *Proc. of the ACL Workshop on A Broader Perspective on MWEs (MWE 2007)*, pages 41–48, Prague, Czech Republic. ACL.

Cook, P., Fazly, A., and Stevenson, S. (2008). The VNC-tokens dataset. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 19–22, Marrakech, Morocco.

Cook, P. and Stevenson, S. (2006). Classifying particle semantics in English verb-particle constructions. In Moirón, B. V., Villavicencio, A., McCarthy, D., Evert, S., and Stevenson, S., editors, *Proc. of the COLING/ACL Workshop on MWEs: Identifying and Exploiting Underlying Properties (MWE 2006)*, pages 45–53, Sidney, Australia. ACL.

da Silva, B. C. D. (2010). Brazilian Portuguese wordnet: A computational linguistic exercise of encoding bilingual relational lexicons. *International Journal of Computational Linguistics and Applications*, 1(1-2):137–150.

da Silva, J. F., Dias, G., Guilloré, S., and Lopes, J. G. P. (1999). Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, EPIA 1999, pages 113–132, London, UK. Springer.

Dagan, I. and Church, K. (1994). Termight: Identifying and translating technical terminology. In *Proc. of the 4th ANLP Conf. (ANLP 1994)*, pages 34–40, Stuttgart, Germany. ACL.

Daille, B. (2003). Conceptual structuring through term variations. In Bond, F., Korhonen, A., McCarthy, D., and Villavicencio, A., editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 9–16, Sapporo, Japan. ACL.

Daille, B., Dufour-Kowalski, S., and Morin, E. (2004). French-English multi-word term alignment based on lexical context analysis. In *Proc. of the Fourth LREC (LREC 2004)*, pages 919–922, Lisbon, Portugal. ELRA.

Danlos, L. and Samvelian, P. (1992). Translation of the predicative element of a sentence: category switching, aspect and diathesis. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 21–34, Montréal, Canada.

Das, D., Pal, S., Mondal, T., Chakraborty, T., and Bandyopadhyay, S. (2010). Automatic extraction of complex predicates in Bengali. In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 36–44, Beijing, China. ACL.

de Cruys, T. V. and Moirón, B. V. (2007). Semantics-based multiword expression extraction. In Grégoire, N., Evert, S., and Kim, S. N., editors, *Proc. of the ACL Workshop on A Broader Perspective on MWEs (MWE 2007)*, pages 25–32, Prague, Czech Republic. ACL.

de Medeiros Caseli, H., Ramisch, C., das Graças Volpe Nunes, M., and Villavicencio, A. (2010). Alignment-based extraction of multiword expressions. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2):59–77.

de Medeiros Caseli, H., Villavicencio, A., Machado, A., and Finatto, M. J. (2009). Statistically-driven alignment-based multiword expression identification for technical domains. In Anastasiou, D., Hashimoto, C., Nakov, P., and Kim, S. N., editors, *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)*, pages 1–8, Suntec, Singapore. ACL.

Déjean, H., Gaussier, É., and Sadat, F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proc. of the 19th COLING (COLING 2002)*, Taipei, Taiwan.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the RSS. Series B*, 39(1):1–38.

Devereux, B. and Costello, F. (2007). Learning to interpret novel noun-noun compounds: evidence from a category learning experiment. pages 89–96, Prague, Czech Republic. ACL.

Dias, G. (2003). Multiword unit hybrid extraction. In Bond, F., Korhonen, A., McCarthy, D., and Villavicencio, A., editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 41–48, Sapporo, Japan. ACL.

Doddington, G. (2002). Automatic evaluation of machine translation quality using *n*-gram co-occurrence statistics. In *Proc. of the Second HLT Conf. (HLT 2002)*, pages 128–132, San Diego, CA, USA. Morgan Kaufmann Publishers.

Doucet, A. and Ahonen-Myka, H. (2004). Non-contiguous word sequences for information retrieval. In Tanaka, T., Villavicencio, A., Bond, F., and Korhonen, A., editors, *Proc. of the ACL Workshop on MWEs: Integrating Processing (MWE 2004)*, pages 88–95, Barcelona, Spain. ACL.

Dras, M. (1995). Automatic identification of support verbs: A step towards a definition of semantic weight. In *Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence*, pages 451–458, Canberra, Australia. World Scientific Press.

Duan, J., Lu, R., Wu, W., Hu, Y., and Tian, Y. (2006). A bio-inspired approach for multiword expression extraction. In Curran, J., editor, *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 176–182, Sidney, Australia. ACL.

Duarte, I., Gonçalves, A., Miguel, M., Mendes, A., Hendrickx, I., Oliveira, F., Cunha, L. F., Silva, F., and Silvano, P. (2010). Light verbs features in European Portuguese. In *Proceedings of the Interdisciplinary Workshop on Verbs: The Identification and Representation of Verb Features (Verb 2010)*, Pisa, Italy.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Comp. Ling.*, 19(1):61–74.

Duran, M. S. and Ramisch, C. (2011). How do you feel? investigating lexical-syntactic patterns in sentiment expression. In *Proceedings of Corpus Linguistics 2011: Discourse and Corpus Linguistics Conference*, Birmingham, UK.

Duran, M. S., Ramisch, C., Aluísio, S. M., and Villavicencio, A. (2011). Identifying and analyzing Brazilian Portuguese complex predicates. In Kordoni et al. (2011a), pages 74–82.

Esuli, A. and Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proc. of the Sixth LREC (LREC 2006)*, pages 417–422, Genoa, Italy. ELRA.

Eugenio, B. D. and Glass, M. (2004). The kappa statistic: A second look. *Comp. Ling.*, 30(1):95–101.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany. 353 p.

Evert, S. and Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):450–466.

Fazly, A., Cook, P., and Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Comp. Ling.*, 35(1):61–103.

Fazly, A. and Stevenson, S. (2006). Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proc. of the 11th Conf. of the EACL (EACL 2006)*, Trento, Italy. ACL.

Fazly, A. and Stevenson, S. (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In Grégoire, N., Evert, S., and Kim, S. N., editors, *Proc. of the ACL Workshop on A Broader Perspective on MWEs (MWE 2007)*, pages 9–16, Prague, Czech Republic. ACL.

Fazly, A., Stevenson, S., and North, R. (2007). Automatically learning semantic knowledge about multiword predicates. *Lang. Res. & Eval.*, 41(1):61–89.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. MIT Press. 423 p.

Fillmore, C. J., Kay, P., and O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64:501–538.

Finlayson, M. and Kulkarni, N. (2011). Detecting multi-word expressions improves word sense disambiguation. In Kordoni et al. (2011a), pages 20–24.

Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. Oxford UP, Oxford, UK. 233 p.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Forcada, M. L. (2009). Apertium: traducció automàtica de codi obert per a les llengües romàniques. *Linguamática*, 1(1):13–23.

Fotopoulou, A. (1993). *Une classification des phrases à compléments figés en grec moderne : étude morphosyntaxique des phrases figées*. PhD thesis, Université Paris VIII. 248 p.

Fotopoulou, A. (1997). *L'ordre des mots dans les phrases figées à un complément libre en grec moderne*, pages 37–48. Saint-Cloud. INALF.

Fotopoulou, A., Giannopoulos, G., Zourari, M., and Mini, M. (2008). Automatic recognition and extraction of multiword nominal expressions from corpora (in Greek). In *Proceedings of the 29th Annual Meeting, Department of Linguistics*, Aristotle University of Thessaloniki, Greece.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multiword terms: the C-value/NC-value method. *Int. J. on Digital Libraries*, 3(2):115–130.

Fritzinger, F., Weller, M., and Heid, U. (2010). A survey of idiomatic preposition-noun-verb triples on token level. In *Proc. of the Seventh LREC (LREC 2010)*, pages 2908–2914, Malta. ELRA.

Gale, W. A. and Church, K. (1993). A program for aligning sentences in bilingual corpora. *Comp. Ling.*, 19(1):75–102.

Gil, A. and Dias, G. (2003). Using masks, suffix array-based data structures and multidimensional arrays to compute positional *n*-gram statistics from corpora. In Bond, F., Korhonen, A., McCarthy, D., and Villavicencio, A., editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 25–32, Sapporo, Japan. ACL.

Gill, A. J., French, R. M., Gergle, D., and Oberlander, J. (2008). The language of emotion in short blog texts. San Diego, CA, USA. ACM.

Girju, R., Moldovan, D., Tatu, M., and Antohe, D. (2005). On the semantics of noun compounds. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):479–496.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D. (2009). Classification of semantic relations between nominals. *Lang. Res. & Eval. Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond*, 43(2):105–121.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Graliński, F., Savary, A., Czerepowicka, M., and Makowiecki, F. (2010). Computational lexicography of multi-word units: How efficient can it be? In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 1–9, Beijing, China. ACL.

Granada, R., Lopes, L., Ramisch, C., Trojahn, C., Vieira, R., and Villavicencio, A. (2012). A comparable corpus based on aligned multilingual ontologies. In *Proceedings of the ACL 2012 First Workshop on Multilingual Modeling (MM 2012)*, Jeju, Republic of Korea. Association for Computational Linguistics.

Green, S., de Marneffe, M.-C., Bauer, J., and Manning, C. D. (2011). Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In Barzilay, R. and Johnson, M., editors, *Proc. of the 2011 EMNLP (EMNLP 2011)*, pages 725–735, Edinburgh, Scotland, UK. ACL.

Grefenstette, G. (1999). The World Wide Web as a resource for example-based machine translation tasks. In *Proc. of the Twenty-First Translating and the Computer*, London, UK. ASLIB.

Grégoire, N. (2007). Design and implementation of a lexicon of Dutch multiword expressions. In Grégoire, N., Evert, S., and Kim, S. N., editors, *Proc. of the ACL Workshop on A Broader Perspective on MWEs (MWE 2007)*, pages 17–24, Prague, Czech Republic. ACL.

Grégoire, N. (2010). DuELME: a Dutch electronic lexicon of multiword expressions. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2):23–39.

Grégoire, N., Evert, S., and Kim, S. N., editors (2007). *Proc. of the ACL Workshop on A Broader Perspective on MWEs (MWE 2007)*, Prague, Czech Republic. ACL. 80 p.

Grégoire, N., Evert, S., and Krenn, B., editors (2008). *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, Marrakech, Morocco. 57 p.

Gurrutxaga, A. and Alegria, I. (2011). Automatic extraction of NV expressions in Basque: Basic issues on cooccurrence techniques. In Kordoni et al. (2011a), pages 2–7.

Haugereid, P. and Bond, F. (2011). Extracting transfer rules for multiword expressions from parallel corpora. In Kordoni et al. (2011a), pages 92–100.

Hautli, A. and Sulger, S. (2011). Extracting and classifying Urdu multiword expressions. In *Proc. of the ACL 2011 SRW*, pages 24–29, Portland, OR, USA. ACL.

Hazelbeck, G. and Saito, H. (2010). A hybrid approach for functional expression identification in a Japanese reading assistant. In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 80–83, Beijing, China. ACL.

Heid, U. and Weller, M. (2008). Tools for collocation extraction: Preferences for active vs. passive. In *Proc. of the Sixth LREC (LREC 2008)*, pages 1266–1272, Marrakech, Morocco. ELRA.

Hendrickx, I., Mendes, A., Pereira, S., Gonçalves, A., and Duarte, I. (2010). Complex predicates annotation in a corpus of Portuguese. In *Proceedings of the ACL 2010 Fourth Linguistic Annotation Workshop*, pages 100–108, Uppsala, Sweden.

Hoang, H. H., Kim, S. N., and Kan, M.-Y. (2009). A re-examination of lexical association measures. In Anastasiou, D., Hashimoto, C., Nakov, P., and Kim, S. N., editors, *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)*, pages 31–39, Suntec, Singapore. ACL.

Hogan, D., Foster, J., and van Genabith, J. (2011). Decreasing lexical data sparsity in statistical syntactic parsing - experiments with named entities. In Kordoni et al. (2011a), pages 14–19.

Huang, C.-R., Kilgarriff, A., Wu, Y., Chiu, C., Smith, S., Rychly, P., Bai, M.-H., and Chen, K.-J. (2005). Chinese sketch engine and the extraction of grammatical collocations. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 48–55, Jeju Island, South Korea.

Hwang, J. D., Bhatia, A., Bonial, C., Mansouri, A., Vaidya, A., Zhou, Y., Xue, N., and Palmer, M. (2010). Propbank annotation of multilingual light verb constructions. In *Proceedings of the ACL 2010 Fourth Linguistic Annotation Workshop*, pages 82–90, Uppsala, Sweden.

Ikehara, S., Tokuhisa, M., and Murakami, J. (2008). Non-compositional language model and pattern dictionary development for Japanese compound and complex sentences. In Scott, D. and Uszkoreit, H., editors, *Proc. of the 22nd COLING (COLING 2008)*, pages 353–360, Manchester, UK. The Coling 2008 Organizing Committee.

Izumi, T., Imamura, K., Kikui, G., and Sato, S. (2010). Standardizing complex functional expressions in Japanese predicates: Applying theoretically-based paraphrasing rules. In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 63–71, Beijing, China. ACL.

Jackendoff, R. (1997). Twistin' the night away. *Language*, 73:534–559.

Jespersen, O. (1965). *A Modern English Grammar on Historical Principles*. George Allen and Unwin Ltd., London, UK. 400 p.

Joshi, A. (2010). Multi-word expressions as discourse relation markers (DRMs). In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, page 89, Beijing, China. ACL.

Joyce, T. and Srdanović, I. (2008). Comparing lexical relationships observed within Japanese collocation data and Japanese word association norms. In Zock, M. and Huang, C.-R., editors, *Proc. of the COLING 2008 COGALEX workshop (COGALEX 2008)*, pages 1–8, Manchester, UK. The Coling 2008 Organizing Committee.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing*. Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition. 1024 p.

Justeson, J. S. and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Nat. Lang. Eng.*, 1(1):9–27.

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Comp. Ling. Special Issue on the Web as Corpus*, 29(3):459–484.

Kilgarriff, A. (2007). Googleology is bad science. *Comp. Ling.*, 33(1):147–151.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Comp. Ling. Special Issue on the Web as Corpus*, 29(3):333–347.

Kim, J.-D., Ohta, T., Teteisi, Y., and Tsujii, J. (2006). GENIA ontology. Technical report, Tsujii Laboratory, University of Tokyo.

Kim, S., Yang, Z., Song, M., and Ahn, J.-H. (1999). Retrieving collocations from Korean text. In Fung, J. Z. P., editor, *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP 1999)*, pages 71–81, College Park, MD, USA.

Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. In *Proc. of the 20th COLING (COLING 2004)*, pages 1367–1373, Geneva, Switzerland. ICCL.

Kim, S. N. and Baldwin, T. (2008). Standardised evaluation of English noun compound interpretation. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 39–42, Marrakech, Morocco.

Kim, S. N. and Baldwin, T. (2010). How to pick out token instances of English verb-particle constructions. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2):97–113.

Kim, S. N. and Kan, M.-Y. (2009). Re-examining automatic keyphrase extraction approaches in scientific articles. In Anastasiou, D., Hashimoto, C., Nakov, P., and Kim, S. N., editors, *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)*, pages 9–16, Suntec, Singapore. ACL.

Kim, S. N. and Nakov, P. (2011). Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In Barzilay, R. and Johnson, M., editors, *Proc. of the 2011 EMNLP (EMNLP 2011)*, pages 648–658, Edinburgh, Scotland, UK. ACL.

Kneser, R. and Ney, H. (1995). Improved backing-off for *M*-gram language modeling. In *Proc.of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1995)*, volume 1, pages 181–184.

Knight, K. (1999). Decoding complexity in word-replacement translation models. *Comp. Ling.*, 25(4):607–615.

Knight, K. and Koehn, P. (2003). What's new in statistical machine translation. In *Proc. of the 2003 Conf. of the NAACL on HLT (NAACL 2003)*, page 5, Edmonton, Canada. ACL.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proc. of the Tenth MT Summit(MT Summit 2005)*, pages 79–86, Phuket, Thailand. AAMT.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge UP, Cambridge, UK. 488 p.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proc. of the 45th ACL (ACL 2007)*, pages 177–180, Prague, Czech Republic. ACL.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proc. of the 2003 Conf. of the NAACL on HLT (NAACL 2003)*, pages 48–54, Edmonton, Canada. ACL.

Kordoni, V., Ramisch, C., and Villavicencio, A., editors (2011a). *Proc. of the ACL Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, Portland, OR, USA. ACL.

Kordoni, V., Ramisch, C., and Villavicencio, A., editors (2011b). *Proc.of the ACL Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, Portland, OR, USA. ACL. 144 p.

Kordoni, V., Ramisch, C., and Villavicencio, A., editors (2013). *ACM Trans. Speech and Lang. Process. Special Issue on multiword expressions (TSLP) — to appear*. ACM, New York, NY, USA.

Korkontzelos, I. and Manandhar, S. (2010). Can recognising multiword expressions improve shallow parsing? In *Proc. of HLT: The 2010 Annual Conf. of the NAACL (NAACL 2003)*, pages 636–644, Los Angeles, California. ACL.

Krenn, B. (2008). Description of evaluation resource – German PP-verb data. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 7–10, Marrakech, Morocco.

Krieger, M. and Finatto, M. J. B. (2004). *Introdução à Terminologia: teoria & prática*. Editora Contexto, São Paulo, SP, Brazil. 223 p.

Kulkarni, N. and Finlayson, M. (2011). jMWE: A Java toolkit for detecting multi-word expressions. In Kordoni et al. (2011a), pages 122–124.

Langer, S. (2004). A linguistic test battery for support verb constructions. *Special issue of Linguisticae Investigationes*, 27(2):171–184.

Langer, S. (2005). A formal specification of support verb constructions. In Langer, S. and Schnorbusch, D., editors, *Semantik im Lexikon*, pages 179–202, Tübingen, Germany. Gunter Naar Verlag.

Lapata, M. (2002). The disambiguation of nominalizations. *Comp. Ling.*, 28(3):357–388.

Lapata, M. and Keller, F. (2005). Web-based models for natural language processing. *ACM Trans. Speech and Lang. Process. (TSLP)*, 2(1):1–31.

Laporte, É., Nakamura, T., and Voyatzi, S. (2008). A French corpus annotated for multiword nouns. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 27–30, Marrakech, Morocco.

Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors (2010). *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, Beijing, China. ACL. 89 p.

Laporte, É. and Voyatzi, S. (2008). An electronic dictionary of French multiword adverbs. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 31–34, Marrakech, Morocco.

Lavagnino, E. and Park, J. (2010). Conceptual structure of automatically extracted multiword terms from domain specific corpora: a case study for Italian. In Zock, M. and Rapp, R., editors, *Proc. of the 2nd COGALEX workshop (COGALEX 2010)*, pages 48–55, Beijing, China. The Coling 2010 Organizing Committee.

Lee, J. (2011). Two types of Korean light verb constructions in a typed feature structure grammar. In Kordoni et al. (2011a), pages 40–48.

Lee, L., Aw, A., Zhang, M., and Li, H. (2010). EM-based hybrid model for bilingual terminology extraction from comparable corpora. In Huang, C.-R. and Jurafsky, D., editors, *Proc. of the 23rd COLING (COLING 2010) — Posters*, pages 639–646, Beijing, China. The Coling 2010 Organizing Committee.

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W. N. G., Weese, J., and Zaidan, O. F. (2009). Joshua: an open source toolkit for parsing-based machine translation. In *Proc.of the Fourth StatMT (WMT 2009)*, pages 135–139, Athens, Greece. ACL.

Linardaki, E., Ramisch, C., Villavicencio, A., and Fotopoulou, A. (2010). Towards the construction of language resources for Greek multiword expressions: Extraction and evaluation. In Piperidis, S., Slavcheva, M., and Vertan, C., editors, *Proc. of the LREC Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*, pages 31–40, Valetta, Malta. May.

Lohse, B., Hawkins, J. A., and Wasow, T. (2004). Domain minimization in English verb-particle constructions. *Language*, 80(2):238–261.

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49.

López, J. M., Gil, R., García, R., Cearreta, I., and Garay, N. (2008). Towards an ontology for describing emotions. In *Emerging Technologies and Information Systems for the Knowledge Society*, volume 5288 of *LNCS*, pages 96–104, Athens, Greece. Springer.

Manber, U. and Myers, G. (1990). Suffix arrays: a new method for on-line string searches. In *SODA '90: Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*, pages 319–327, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Mangeot, M. and Chalvin, A. (2006). Dictionary building with the jibiki platform: the GDEF case. In *Proc. of the Sixth LREC (LREC 2006)*, pages 1666–1669, Genoa, Italy. ELRA.

Mangeot, M. and Ramisch, C. (2012). A serious lexical game for building a Portuguese lexical-semantic network. In *Proceedings of the ACL 2012 3rd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources and their Applications to NLP*, Jeju, Republic of Korea. Association for Computational Linguistics.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge, USA. 620 p.

Marchello-Nizia, C. (1996). A diachronic survey of support verbs: the case of old French. *Langages*, 30(121):91–98.

Martens, S. (2010). Varro: An algorithm and toolkit for regular structure discovery in treebanks. In Huang, C.-R. and Jurafsky, D., editors, *Proc. of the 23rd COLING (COLING 2010) — Posters*, pages 810–818, Beijing, China. The Coling 2010 Organizing Committee.

Martens, S. and Vandeghinste, V. (2010). An efficient, generic approach to extracting multi-word expressions from dependency trees. In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 84–87, Beijing, China. ACL.

Mathieu, Y. Y. (2005). Annotation of emotions and feelings in texts. In *Affective Computing and Intelligent Interaction*, volume 3784 of *LNCS*, pages 350–357, Beijing, China. Springer.

McCarthy, D., Keller, B., and Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. In Bond, F., Korhonen, A., McCarthy, D., and Villavicencio, A., editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 73–80, Sapporo, Japan. ACL.

McCarthy, D., Venkatapathy, S., and Joshi, A. (2007). Detecting compositionality of verb-object combinations using selectional preferences. In Eisner, J., editor, *Proc. of the 2007 Joint Conference on EMNLP and Computational NLL (EMNLP-CoNLL 2007)*, pages 369–379, Prague, Czech Republic. ACL.

Melamed, I. D. (1997). Automatic discovery of non-compositional compounds in parallel data. In *Proc. of the 2nd EMNLP (EMNLP-2)*, pages 97–108, Brown University, RI, USA. ACL.

Mel'čuk, I., Arbatchewsky-Jumarie, N., Elnitsky, L., Iordanskaja, L., and Lessard, A. (1984). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I*. Les presses de l'Université de Montréal, Montréal, Canada. 172 p.

Mel'čuk, I. and Polguère, A. (1987). A formal lexicon in the meaning-text theory or (how to do lexica with words). *Comp. Ling.*, 13(3-4):261–275.

Mel'čuk, I., Arbatchewsky-Jumarie, N., Clas, A., Mantha, S., and Polguère, A. (1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV*. Les presses de l'Université de Montréal, Montréal, Canada. 347 p.

Mel'čuk, I., Arbatchewsky-Jumarie, N., Dagenais, L., Elnitsky, L., Iordanskaja, L., Lefebvre, M.-N., and Mantha, S. (1988). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II*. Les presses de l'Université de Montréal, Montréal, Canada. 332 p.

Mel'čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., and Mantha, S. (1992). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III*. Les presses de l'Université de Montréal, Montréal, Canada. 323 p.

Mel'čuk, I., Clas, A., and Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Editions Duculot, Louvain la Neuve, Belgium. 256 p.

Messiant, C., Poibeau, T., and Korhonen, A. (2008). Lexschem: a large subcategorization lexicon for French verbs. In *Proc. of the Sixth LREC (LREC 2008)*, pages 533–538, Marrakech, Morocco. ELRA.

Michou, A. and Seretan, V. (2009). A tool for multi-word expression extraction in modern Greek using syntactic parsing. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 45–48, Athens, Greece. ACL.

Mini, M. and Fotopoulou, A. (2009). Typology of multiword verbal expressions in modern Greek dictionaries: limits and differences (in Greek). Proceedings of the 18th International Symposium of Theoretical & Applied Linguistics, School of English, pages 491–503, Aristotle University of Thessaloniki, Greece.

Moirón, B. V., Villavicencio, A., McCarthy, D., Evert, S., and Stevenson, S., editors (2006). *Proc. of the COLING/ACL Workshop on MWEs: Identifying and Exploiting Underlying Properties (MWE 2006)*, Sidney, Australia. ACL. 61 p.

Morin, E. and Daille, B. (2010). Compositionality and lexical alignment of multi-word terms. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2):79–95.

Moustaki, A. (1995). *Les expressions figées ε′ιμαι/être Prép C W en grec moderne*. PhD thesis, Université Paris VIII. 476 p.

Mukerjee, A., Soni, A., and Raina, A. M. (2006). Detecting complex predicates in Hindi using POS projection across parallel corpora. In Moirón, B. V., Villavicencio, A., McCarthy, D., Evert, S., and Stevenson, S., editors, *Proc. of the COLING/ACL Workshop on MWEs: Identifying and Exploiting Underlying Properties (MWE 2006)*, pages 28–35, Sidney, Australia. ACL.

Nakov, P. (2007). *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. PhD thesis, EECS Department, University of California, Berkeley, CA, USA. 392 p.

Nakov, P. (2008a). Improved statistical machine translation using monolingual paraphrases. In Ghallab, M., Spyropoulos, C. D., Fakotakis, N., and Avouris, N. M., editors, *Proc. of the 18th ECAI (ECAI 2008)*, volume 178 of *Frontiers in Artificial Intelligence and Applications*, pages 338–342, Patras, Greece. IOS Press.

Nakov, P. (2008b). Paraphrasing verbs for noun compound interpretation. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 46–49, Marrakech, Morocco.

Nakov, P. and Hearst, M. A. (2005). Search engine statistics beyond the *n*-gram: Application to noun compound bracketing. In Dagan, I. and Gildea, D., editors, *Proc. of the Ninth CoNLL (CoNLL-2005)*, pages 17–24, University of Michigan, MI, USA. ACL.

Nakov, P. and Hearst, M. A. (2008). Solving relational similarity problems using the web as a corpus. In *Proc. of the 46th ACL: HLT (ACL-08: HLT)*, pages 452–460, Columbus, OH, USA. ACL.

Nematzadeh, A., Fazly, A., and Stevenson, S. (2012). Child acquisition of multiword verbs: A computational investigation — to appear. In Poibeau, T., Villavicencio, A., Korhonen, A., and Alishahi, A., editors, *Cognitive Aspects of Computational Language Acquisition*.

Neves, M. H. M. (1996). Estudo das construções com verbos-suporte em português. In Koch, I. G. V., editor, *Gramática do português falado VI: Desenvolvimentos*, pages 201–231, Campinas, SP, Brazil. Unicamp FAPESP.

Newman, M. E. J. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351.

Nicholson, J. and Baldwin, T. (2006). Interpretation of compound nominalisations using corpus and web statistics. In Moirón, B. V., Villavicencio, A., McCarthy, D., Evert, S., and Stevenson, S., editors, *Proc. of the COLING/ACL Workshop on MWEs: Identifying and Exploiting Underlying Properties (MWE 2006)*, pages 54–61, Sidney, Australia. ACL.

Nicholson, J. and Baldwin, T. (2008). Interpreting compound nominalisations. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 43–45, Marrakech, Morocco.

North, R. (2005). Computational measures of the acceptability of light verb constructions. Master's thesis, University of Toronto, Toronto, ON, Canada. 100 p.

Och, F. J. (2005). Statistical machine translation: Foundations and recent advances. In *Proc. of the Tenth MT Summit(MT Summit 2005)*, Phuket, Thailand. AAMT.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proc. of the 38th ACL (ACL 2000)*, pages 440–447, Hong Kong, China. ACL.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comp. Ling.*, 29(1):19–51.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Comp. Ling.*, 30(4):417–449.

Ohta, T., Tateishi, Y., and Kim, J.-D. (2002). The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proc. of the Second HLT Conf. (HLT 2002)*, pages 82–86, San Diego, CA, USA. Morgan Kaufmann Publishers.

Pal, S., Naskar, S. K., Pecina, P., Bandyopadhyay, S., and Way, A. (2010). Handling named entities and compound verbs in phrase-based statistical machine translation. In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 45–53, Beijing, China. ACL.

Pang, B. and Lee, L. (2008). *Opinion Mining and Sentiment Analysis*, volume 2 of *Foundations and Trends in Information Retrieval*. 135 p.

Papageorgiou, H., Prokopidis, P., Giouli, V., and Piperidis, S. (2000). A unified POS tagging architecture and its application to Greek. In *Proc. of the Second LREC (LREC 2000)*, pages 1455–1462, Athens, Greece. ELRA.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evalution of machine translation. In *Proc. of the 40th ACL (ACL 2002)*, pages 311–318, Philadelphia, PA, USA. ACL.

Pearce, D. (2001). Synonymy in collocation extraction. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop)*, pages 41–46.

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. In *Proc. of the Third LREC (LREC 2002)*, pages 1530–1536, Las Palmas, Canary Islands, Spain. ELRA.

Pecina, P. (2005). An extensive empirical study of collocation extraction methods. In *Proc. of the ACL 2005 SRW*, pages 13–18, Ann Arbor, MI, USA. ACL.

Pecina, P. (2008a). A machine learning approach to multiword expression extraction. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 54–57, Marrakech, Morocco.

Pecina, P. (2008b). Reference data for Czech collocation extraction. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 11–14, Marrakech, Morocco.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2):137–158.

Pedersen, T. (1996). Fishing for exactness. In *Proc. of the South-Central SAS Users Group Conference (SCSUG-96)*, pages 188–200, Austin, TX, USA.

Pedersen, T., Banerjee, S., McInnes, B., Kohli, S., Joshi, M., and Liu, Y. (2011). The *n*-gram statistics package (text::NSP) : A flexible tool for identifying *n*-grams, collocations, and word associations. In Kordoni et al. (2011a), pages 131–133.

Piao, S. S. L., Rayson, P., Archer, D., Wilson, A., and McEnery, T. (2003). Extracting multiword expressions with a semantic tagger. In Bond, F., Korhonen, A., McCarthy, D., and Villavicencio, A., editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 49–56, Sapporo, Japan. ACL.

Piao, S. S. L., Sun, G., Rayson, P., and Yuan, Q. (2006). Automatic extraction of Chinese multiword expressions with a statistical tool. In Rayson, P., Sharoff, S., and Adolphs, S., editors, *Proc. of the EACL Workshop on MWEsin Multilingual Context (EACL-MWE 2006)*, Trento, Italy. ACL.

Planas, E. and Furuse, O. (2000). Multi-level similar segment matching algorithm for translation memories and example-based machine translation. In *Proc. of the 18th COLING (COLING 2000)*, Saarbrücken, Germany.

Preiss, J., Briscoe, T., and Korhonen, A. (2007). A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proc. of the 45th ACL (ACL 2007)*, pages 912–919, Prague, Czech Republic. ACL.

Ramisch, C. (2009). Multiword terminology extraction for domain-specific documents. Master's thesis, École Nationale Supérieure d'Informatique et de Mathématiques Appliquées, Grenoble, France. 79 p.

Ramisch, C. (2012a). A generic framework for multiword expressions treatment: from acquisition to applications. In *Proc. of the ACL 2012 SRW*, Jeju, Republic of Korea. ACL.

Ramisch, C. (2012b). Une plate-forme générique et ouverte pour le traitement des expressions polylexicales. In Molina Mejia, J. M. and Schwab, D., editors, *Actes de 14e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2012)*, Grenoble, France.

Ramisch, C., Araujo, V. D., and Villavicencio, A. (2012). A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proc. of the ACL 2012 SRW*, Jeju, Republic of Korea. ACL.

Ramisch, C., de Medeiros Caseli, H., Villavicencio, A., Machado, A., and Finatto, M. J. (2010a). A hybrid approach for multiword expression identification. In *Proc. of the 9th PROPOR (PROPOR 2010)*, volume 6001 of *LNCS (LNAI)*, pages 65–74, Porto Alegre, RS, Brazil. Springer.

Ramisch, C., Schreiner, P., Idiart, M., and Villavicencio, A. (2008a). An evaluation of methods for the extraction of multiword expressions. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 50–53, Marrakech, Morocco.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010b). Multiword expressions in the wild? the mwetoolkit comes in handy. In Liu, Y. and Liu, T., editors, *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China. The Coling 2010 Organizing Committee.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010c). mwetoolkit: a framework for multiword expression identification. In *Proc. of the Seventh LREC (LREC 2010)*, pages 662–669, Malta. ELRA.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010d). Web-based and combined language models: a case study on noun compound identification. In Huang, C.-R. and Jurafsky, D., editors, *Proc. of the 23rd COLING (COLING 2010) — Posters*, pages 1041–1049, Beijing, China. The Coling 2010 Organizing Committee.

Ramisch, C., Villavicencio, A., Moura, L., and Idiart, M. (2008b). Picking them up and figuring them out: Verb-particle constructions, noise and idiomaticity. In Clark, A. and Toutanova, K., editors, *Proc. of the Twelfth CoNLL (CoNLL 2008)*, pages 49–56, Manchester, UK. The Coling 2008 Organizing Committee.

Ranchhod, E. (1999). Construções com nomes predicativos na crónica geral de espanha de 1344. In Faria, I. H., editor, *Lindley Cintra. Homenagem ao Homem, ao Mestre e ao Cidadão*, pages 667–682, Lisbon, Portugal. Cosmos.

Rapp, R. (2008). The computation of associative responses to multiword stimuli. In Zock, M. and Huang, C.-R., editors, *Proc. of the COLING 2008 COGALEX workshop (COGALEX 2008)*, pages 102–109, Manchester, UK. The Coling 2008 Organizing Committee.

Rayson, P., Piao, S., Sharoff, S., Evert, S., and Moirón, B. V., editors (2010a). *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, volume 44. Springer.

Rayson, P., Piao, S., Sharoff, S., Evert, S., and Moirón, B. V. (2010b). Multiword expressions: hard going or plain sailing? *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2):1–5.

Rayson, P., Sharoff, S., and Adolphs, S., editors (2006). *Proc. of the EACL Workshop on MWEsin Multilingual Context (EACL-MWE 2006)*, Trento, Italy. ACL. 79 p.

Ren, Z., Lü, Y., Cao, J., Liu, Q., and Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. In Anastasiou, D., Hashimoto, C., Nakov, P., and Kim, S. N., editors, *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)*, pages 47–54, Suntec, Singapore. ACL.

Rio-Torto, G. (2006). O Léxico: semântica e gramática das unidades lexicais. In *Estudos sobre léxico e gramática*, pages 11–34, Coimbra, Portugal. CIEG/FLUL.

Rychlý, P. and Smrz, P. (2004). Manatee, bonito and word sketches for Czech. In *Proceedings of the Second International Conference on Corpus Linguisitcs*, pages 124–131, Saint-Petersburg, Russia.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd CICLing (CICLing-2002)*, volume 2276/2010 of *LNCS*, pages 1–15, Mexico City, Mexico. Springer.

Salkoff, M. (1990). Automatic translation of support verb constructions. In *Proc. of the 13th COLING (COLING 1990)*, pages 243–246, Helsinki, Finland.

SanJuan, E., Dowdall, J., Ibekwe-SanJuan, F., and Rinaldi, F. (2005). A symbolic approach to automatic multiword term structuring. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):524–542.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Schone, P. and Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In Lee, L. and Harman, D., editors, *Proc. of the 2001 EMNLP (EMNLP 2001)*, pages 100–108, Pittsburgh, PA USA. ACL.

Schuler, W. and Joshi, A. (2011). Tree-rewriting models of multi-word expressions. In Kordoni et al. (2011a), pages 25–30.

Seretan, V. (2008). *Collocation extraction based on syntactic parsing*. PhD thesis, University of Geneva, Geneva, Switzerland. 249 p.

Seretan, V. and Wehrli, E. (2006). Multilingual collocation extraction: Issues and solutions. In Witt, A., Sérasset, G., Armstrong, S., Breen, J., Heid, U., and Sasaki, F., editors, *Proc. of the ACL Workshop on Multilingual Language Resources and Interoperability*, pages 40–49, Sydney, Australia. ACL.

Seretan, V. and Wehrli, E. (2009). Multilingual collocation extraction with a syntactic parser. *Lang. Res. & Eval. Special Issue on Multilingual Language Resources and Interoperability*, 43(1):71–85.

Seretan, V. and Wehrli, E. (2011). Fipscoview: On-line visualisation of collocations extracted from multilingual parallel corpora. In Kordoni et al. (2011a), pages 125–127.

Shimohata, S., Sugio, T., and Nagata, J. (1997). Retrieving collocations by co-occurrences and word order constraints. In *Proc. of the 35th ACL and 8th Conf. of the EACL (ACL-EACL 1997)*, pages 476–481, Madrid, Spain. ACL.

Shinozaki, T. and Ostendorf, M. (2008). Cross-validation and aggregated EM training for robust parameter estimation. *Comp. Speech & Lang.*, 22(2):185–195.

Silva, H. M. F. (2009). Verbos-suporte ou expressões cristalizadas? *Soletras*, 9(17):175–182.

Silva, J. and Lopes, G. (1999). A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Proceedings of the Sixth Meeting on Mathematics of Language (MOL6)*, pages 369–381, Orlando, FL, USA.

Silva, J. and Lopes, G. (2010). Towards automatic building of document keywords. In Huang, C.-R. and Jurafsky, D., editors, *Proc. of the 23rd COLING (COLING 2010) — Posters*, pages 1149–1157, Beijing, China. The Coling 2010 Organizing Committee.

Silva, M. J., Carvalho, P., Sarmento, L., Oliveira, E., and Magalhães, P. (2009). The design of OPTIMISM, an opinion mining system for Portuguese politics. In *Proc.of the Fourteenth Portuguese Conference on Artificial Intelligence (EPIA 2006)*, pages 565–576, Aveiro, Portugal.

Sinclair, J., editor (1989). *Collins COBUILD Dictionary of Phrasal Verbs*. Collins COBUILD, London, UK. 512 p.

Sinha, R. M. K. (2009). Mining complex predicates in Hindi using a parallel Hindi-English corpus. In Anastasiou, D., Hashimoto, C., Nakov, P., and Kim, S. N., editors, *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)*, pages 40–46, Suntec, Singapore. ACL.

Sinha, R. M. K. (2011). Stepwise mining of multi-word expressions in Hindi. In Kordoni et al. (2011a), pages 110–115.

Smadja, F. A. (1993). Retrieving collocations from text: Xtract. *Comp. Ling.*, 19(1):143–177.

Spina, S. (2010). The dictionary of Italian collocations: Design and integration in an online learning environment. In *Proc. of the Seventh LREC (LREC 2010)*, pages 3202–3208, Malta. ELRA.

Steedman, M. (2008). On becoming a discipline. *Comp. Ling.*, 34(1):137–144.

Stevenson, S., Fazly, A., and North, R. (2004). Statistical measures of the semi-productivity of light verb constructions. In Tanaka, T., Villavicencio, A., Bond, F., and Korhonen, A., editors, *Proc. of the ACL Workshop on MWEs: Integrating Processing (MWE 2004)*, pages 1–8, Barcelona, Spain. ACL.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In Hansen, J. H. L. and Pellom, B., editors, *Proc. of the Seventh ICSLP, Third INTERSPEECH Event (ICSLP 2001 – INTERSPEECH 2002)*, pages 901–904, Denver, CO, USA. ISCA.

Stymne, S. (2009). A comparison of merging strategies for translation of German compounds. In *Proc. of the Student Research Workshop at EACL 2009*, pages 61–69.

Stymne, S. (2011). Pre- and postprocessing for statistical machine translation into Germanic languages. In *Proc. of the ACL 2011 SRW*, pages 12–17, Portland, OR, USA. ACL.

Tanaka, T. and Baldwin, T. (2003). Noun-noun compound machine translation A feasibility study on shallow processing. In Bond, F., Korhonen, A., McCarthy, D., and Villavicencio, A., editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 17–24, Sapporo, Japan. ACL.

Tanaka, T., Villavicencio, A., Bond, F., and Korhonen, A., editors (2004). *Proc. of the ACL Workshop on MWEs: Integrating Processing (MWE 2004)*, Barcelona, Spain. ACL. 103 p.

Teufel, S. and Grefenstette, G. (1995). Corpus-based method for automatic identification of support verbs for nominalizations. In *Proc. of the 7th Conf. of the EACL (EACL 1995)*, pages 98–103, Dublin, Ireland. ACL.

Tillmann, C. and Ney, H. (2003). Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Comp. Ling.*, 29(1):97–133.

Uchiyama, K., Baldwin, T., and Ishizaki, S. (2005). Disambiguating Japanese compound verbs. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):497–512.

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In *IFIP Congress (2)*, pages 1114–1122.

Venkatapathy, S. and Joshi, A. K. (2006). Using information about multi-word expressions for the word-alignment task. In Moirón, B. V., Villavicencio, A., McCarthy, D., Evert, S., and Stevenson, S., editors, *Proc. of the COLING/ACL Workshop on MWEs: Identifying and Exploiting Underlying Properties (MWE 2006)*, pages 20–27, Sidney, Australia. ACL.

Venkatsubramanyan, S. and Perez-Carballo, J. (2004). Multiword expression filtering for building knowledge. In Tanaka, T., Villavicencio, A., Bond, F., and Korhonen, A., editors, *Proc. of the ACL Workshop on MWEs: Integrating Processing (MWE 2004)*, pages 40–47, Barcelona, Spain. ACL.

Villavicencio, A., Bond, F., Korhonen, A., and McCarthy, D., editors (2005a). *Comp. Speech & Lang. Special issue on MWEs*, volume 19. Elsevier.

Villavicencio, A., Bond, F., Korhonen, A., and McCarthy, D. (2005b). Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):365–377.

Villavicencio, A., Idiart, M., Ramisch, C., Araujo, V. D., Yankama, B., and Berwick, R. (2012). Get out but don't fall down: verb-particle constructions in child language. pages 43–50, Avignon, France. ACL.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., and Ramisch, C. (2007). Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In Eisner, J., editor, *Proc. of the 2007 Joint Conference on EMNLP and Computational NLL (EMNLP-CoNLL 2007)*, pages 1034–1043, Prague, Czech Republic. ACL.

Villavicencio, A., Ramisch, C., Machado, A., de Medeiros Caseli, H., and Finatto, M. J. (2010). Identificação de expressões multipalavra em domínios específicos. *Linguamática*, 2(1):15–33.

Vincze, V., T., I. N., and Berend, G. (2011). Detecting noun compounds and light verb constructions: a contrastive study. In Kordoni et al. (2011a), pages 116–121.

Weaver, W. (1955). Translation. In Locke, W. N. and Booth, A. D., editors, *Machine Translation of Languages: Fourteen Essays*, pages 15–23. MIT Press.

Wehrli, E. (1998). Translating idioms. In *Proc. of the 36th ACL and 17th COLING (ACL-COLING 1998)*, pages 1388–1392, Montreal, Quebec, Canada. ACL.

Wehrli, E., Seretan, V., and Nerima, L. (2010). Sentence analysis and collocation identification. In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 27–35, Beijing, China. ACL.

Weller, M. and Heid, U. (2010). Extraction of German multiword expressions from parsed corpora using context features. In *Proc. of the Seventh LREC (LREC 2010)*, pages 3195–3201, Malta. ELRA.

Wermter, J. and Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge) – A qualitative evaluation of association measures for collocation and term extraction. In *Proc. of the 21st COLING and 44th ACL (COLING/ACL 2006)*, pages 785–792, Sidney, Australia. ACL.

Xu, Y., Goebel, R., Ringlstetter, C., and Kondrak, G. (2010). Application of the tightness continuum measure to Chinese information retrieval. In Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 54–62, Beijing, China. ACL.

Yamamoto, M. and Church, K. (2001). Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Comp. Ling.*, 27(1):1–30.

Yarowsky, D. (2001). One sense per collocation. In *Proc. of the First International Conference on HLT Research (HLT 2001)*, pages 266–271, San Diego, CA, USA. Morgan Kaufmann Publishers.

Zaninello, A. and Nissim, M. (2010). Creation of lexical resources for a characterisation of multiword expressions in Italian. In *Proc. of the Seventh LREC (LREC 2010)*, pages 654–661, Malta. ELRA.

Zarrieß, S. and Kuhn, J. (2009). Exploiting translational correspondences for pattern-independent MWE identification. In Anastasiou, D., Hashimoto, C., Nakov, P., and Kim, S. N., editors, *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)*, pages 23–30, Suntec, Singapore. ACL.

Zhang, Y. and Kordoni, V. (2006). Automated deep lexical acquisition for robust open texts processing. In *Proc. of the Sixth LREC (LREC 2006)*, pages 275–280, Genoa, Italy. ELRA.

Zhang, Y., Kordoni, V., Villavicencio, A., and Idiart, M. (2006). Automated multiword expression prediction for grammar engineering. In Moirón, B. V., Villavicencio, A., McCarthy, D., Evert, S., and Stevenson, S., editors, *Proc. of the COLING/ACL Workshop on MWEs: Identifying and Exploiting Underlying Properties (MWE 2006)*, pages 36–44, Sidney, Australia. ACL.

APPENDIX A RÉSUMÉ ÉTENDU

UNE PLATE-FORME GÉNÉRIQUE ET OUVERTE POUR LE TRAITEMENT DES EXPRESSIONS POLYLEXICALES: DE L'ACQUISITION AUX APPLICATIONS

A.1 Introduction

Les expressions polylexicales (EPL, en anglais MWE ou *multiword expression*) sont un problème ouvert et difficile dans le traitement automatique de la langue naturelle, en raison de leur nature complexe. Savoir ce qui peut être considéré comme une EPL est une question polémique. Dit simplement, les EPL sont des combinaisons conventionnelles et récurrentes de mots dans le langage courant (Firth 1957). Probablement les exemples les plus prototypiques d'EPL sont les expressions idiomatiques comme *esprit ouvert*, *casser sa pipe*, *tête de linotte*, *se faire avoir*, *sens dessus dessous*, et *donner un coup de main*. En plus des expressions idiomatiques, d'autres constructions peuvent être considérées comme des EPL. D'autres exemples d'EPL incluent les noms communs (par exemple, *machine à laver*, *messagerie vocale*, *talon aiguille*) et les expressions verbales (par exemple, *avoir du sens*, *tirer avantage*, *prendre une douche*, et *se rendre compte*).

Les locuteurs natifs s'en rendent rarement compte, mais le langage courant est riche en expressions figées comme *bonne journée*, *c'est moi*, *tant pis* et *au revoir*. Il est souvent supposé que le lexique d'un locuteur natif contient autant d'EPL que de mots simples (Jackendoff 1997). Ainsi, tout système informatique traitant le langage humain doit considérer les EPL. Dans de nombreuses applications de TAL, lorsque les mots qui composent une EPL sont traités comme des unités indépendantes, cela peut engendrer des problèmes. Un système de traduction automatique, par exemple, doit identifier les EPL afin d'éviter les traductions littérales.

L'intégration des EPL dans les systèmes de TAL traditionnels peut être compliquée, parce que les EPL se trouvent dans une zone floue entre le lexique et la syntaxe d'une langue. Alors que les EPL sont omniprésentes, la qualité et le nombre de langues couvertes par les outils et les ressources contenant des EPL sont faibles. Il existe donc un besoin croissant de développement, de consolidation et d'évaluation des techniques pour l'acquisition automatique des EPL à partir des corpus.

Cette thèse aborde le problème du traitement des EPL dans les applications de TAL, allant de leur acquisition automatique dans des textes bruts jusqu'à à leur intégration dans deux applications réelles : la lexicographie assistée par ordinateur et la traduction automatique empirique. Nous avons développé un modèle conceptuel pour le processus de traitement des EPL, ainsi qu'une plate-forme logicielle concrète qui valide la méthodologie proposée. Nous avons évalué ce modèle de façon approfondie et systématique. On peut résumer les objectifs de cette thèse comme suit :

1. Développer des techniques génériques et portables pour l'acquisition automatique des EPL à partir des corpus.
2. Évaluer ces techniques extrinsèquement, c'est-à-dire, en mesurant leur utilité dans des applications réelles TAL.
3. Étudier ces techniques dans des contextes bilingues et multilingues, en analysant comment les différents paramètres du contexte d'acquisition influencent la qualité des EPL acquises automatiquement.

A.2 Définitions et caractéristiques

L'étude des EPL est pratiquement aussi ancienne que la propre linguistique. Lorsque l'on essaye de distinguer les phénomènes lexicaux des phénomènes syntaxiques, on s'aperçoit vite que certains d'entre eux, et en particulier les EPL, se situent entre ces deux niveaux. Par conséquent, il existe des limites à l'approche structurelle de la langue à la Chomsky et Tesnière. L'un des articles fondamentaux de la *grammaire à constructions* est l'œuvre de Fillmore et al. (1988). Ils illustrent et discutent en détail les faiblesses de cette approche atomistique et idéalisée de la grammaire. Dans la grammaire à constructions, les idiomes font partie du noyau de la grammaire : une langue peut être entièrement décrite par ses idiomes et leurs propriétés. Ces idiomes correspondent à ce que nous appelons ici des EPL. Une autre théorie linguistique qui confère beaucoup d'importance aux EPL est la *théorie sens-texte* (TST). Les EPL sont présentes en deux points du modèle de la TST : comme *phrasèmes* et comme fonctions lexico-sémantiques dans la *zone de combinatoire lexicale*. Un résumé des EPL dans différentes théories linguistiques est présenté dans Seretan (2008, p. 20–27).

Les EPL sont difficiles à définir, car il n'y a même pas de consensus sur la définition du mot *mot*. La notion d'EPL est originaire de la célèbre citation de Firth « dîtes moi qui vous fréquentez, je vous dirai quel mot vous êtes ». Il affirmait que les « collocations d'un mot donné sont des affirmations sur la place habituelle et usuelle de ce mot » (Firth 1957, p. 181). Smadja (1993) définit une collocation comme étant une « combinaison arbitraire et récurrente de mots ». Pour Choueka (1988), une collocation est « une unité syntaxique et sémantique dont le sens exact ou la connotation ne peuvent pas être dérivés directement et sans ambiguïté du sens ou de la connotation de ses composantes ». Pour Fillmore et al. (1988, p. 504), « une expression idiomatique ou construction est quelque chose qu'un utilisateur de la langue ne peut pas connaître même s'il connaît tout le reste dans cette langue ». Sag et al. (2002) généralisent cette même propriété pour définir les EPL comme des « interprétations idiosyncrasiques qui dépassent la limite du mot (ou les espaces) ».

Toutes ces définitions sont valides dans un contexte expérimental donné. Néanmoins, la définition d'EPL adoptée influencera les résultats d'évaluation, car elle sera utilisée pour écrire les instructions aux annotateurs et pour choisir des références de comparaison. Par conséquent, dans la présente thèse, nous adaptons la définition de Calzolari et al. (2002). Pour nous, les EPL sont « [...] différents phénomènes liés [...]. De façon générale, chacun de ces phénomènes peut être décrit comme une [combinaison] de mots à voir comme une unité à un certain niveau d'analyse linguistique. » Cette définition générique et volontairement vague peut être restreinte selon les besoins des applications. Par exemple, pour un système de traduction automatique, une EPL est toute combinaison de mots qui, quand elle n'est pas traduite comme une unité, génère des traductions peu naturelles ou erronées. Le niveau d'analyse où la combinaison doit être traitée comme une unité varie selon le type d'application et d'expression.

La littérature décrit quelques propriétés communes à toutes les EPL : le caractère arbitraire, l'institutionnalisation, la variabilité sémantique limitée (non-compositionnalité, non-substituabilité, pas de traduction mot-à-mot, spécificité à un domaine), variabilité syntaxique limitée (extra-grammaticalité, lexicalisation), et l'hétérogénéité. Ceux-ci ne sont pas des valeurs binaires du type oui/non, mais des valeurs dans un continuum allant des combinaisons de mots totalement flexibles et ordinaires à des expressions totalement prototypiques et/ou figées.

Il existe plusieurs typologies pour classer les EPL selon les différents points de vue de chaque théorie grammaticale. Dans ce travail, nous proposons une typologie qui repose premièrement sur le rôle morphosyntaxique de l'expression dans une phrase, et deuxièmement sur sa difficulté à être traitée en utilisant des méthodes informatiques. La première typologie classifie les EPL comme expressions nominales, verbales et adverbiales/adjectivales. Les expressions nominales couvrent les noms composés (*roulette russe*), les noms propres (*Porto Alegre*) et les termes polylexicaux (*domaine de liaison à l'ADN*). Les expressions verbales comprennent les verbes à particule (*faire avec*) et les constructions à verbe support (*prendre une douche*). Les expressions adverbiales et adjectivales comprennent des expressions tels que *à poil* en français, *upside down* en anglais et *sem mais nem menos* en portugais. En plus de ces types, nous définissons trois autres types orthogonaux, en rapport avec les méthodes informatiques utilisées pour traiter les EPL : (i) les expressions figées telles que *en somme*, (ii) les expressions idiomatiques comme *coup de foudre*, *laisser à désirer* et *sur la même longueur d'onde*, et (iii) les « vraies » collocations, correspondant aux expressions parfaitement compositionnelles mais qui apparaissent trop souvent ensemble pour n'être que pur hasard. Cette typologie est assez simple mais assez rigoureuse pour décrire les EPL abordées dans nos expériences.

A.3 État de l'art en traitement des EPL

Avant d'entrer dans la discussion sur la vaste littérature en traitement des EPL, rappelons brièvement quelques notions élémentaires. Un *corpus* est tout simplement un corps de textes utilisés dans des études empiriques de la langue (Manning and Schütze 1999, p. 6). L'*analyse linguistique* est le processus qui engendre des représentations plus abstraites à partir du texte brut dans les corpus. Elle peut être vue comme une série d'étapes qui transforment une représentation d'un niveau plus concret vers le prochain niveau plus abstrait : séparation de phrases, séparation de mots, étiquetage morphosyntaxique, et analyse de dépendances.

L'hypothèse statistique qui guide l'acquisition automatique d'EPL est que les mots qui composent une expression vont apparaître ensemble plus souvent que s'ils étaient combinés aléatoirement. Cette hypothèse se concrétise dans la conception des mesures d'association lexicale dans l'acquisition à partir des corpus. Il existe un grand nombre de mesures d'association disponibles dans ce contexte (Evert 2004, Seretan 2008, Pecina 2008b). Pour un n-gramme arbitraire w_1^n, nous calculons sa probabilité par l'estimation du maximum de vraisemblance comme étant $p(w_1^n) = \frac{c(w_1)\times c(w_2)\times\ldots\times c(w_n)}{N^n}$. Quand nous multiplions cette estimation par le nombre total de n-grammes dans le corpus N, nous obtenons une estimation du nombre d'occurrences du n-gramme $E(w_1^n) = \frac{c(w_1)\times c(w_2)\times\ldots\times c(w_n)}{N^{n-1}}$. Les mesures d'association sont généralement fondées sur la différence entre le nombre d'occurrences estimé $E(w_1^n)$ et le nombre d'occurrences observé $c(w_1^n)$, par exemples :

$$\texttt{t-score} = \frac{c(w_1^n) - E(w_1^n)}{\sqrt{c(w_1^n)}}, \quad \texttt{pmi} = \log_2 \frac{c(w_1^n)}{E(w_1^n)}, \quad \texttt{dice} = \frac{n \times c(w_1^n)}{\sum_{i=1}^{n} c(w_i)}.$$

Pour le cas particulier des 2-grammes, il existe des mesures d'association théoriquement plus solides fondées sur les tableaux de contingence. Des exemples de telles mesures sont donnés ci-dessous, où $w_i \in \{w_1, \neg w_1\}$ et $w_j \in \{w_2, \neg w_2\}$:

$$\chi^2 \quad = \sum_{w_i,w_j} \frac{[c(w_i w_j) - E(w_i w_j)]^2}{E(w_i w_j)}, \quad \texttt{ll} \quad = 2 \times \sum_{w_i,w_j} c(w_i w_j) \times \log \frac{c(w_i w_j)}{E(w_i w_j)}.$$

A.3.1 Acquisition d'EPL

Le terme *acquisition d'EPL* comprend leur identification (en contexte) et leur extraction (hors contexte). L'acquisition d'EPL est généralement vue comme un processus à deux étapes.

1. **Extraction de candidates** : une des approches les plus populaires est l'utilisation des séquences d'étiquettes morphosyntaxiques, surtout en acquisition de termes (Justeson and Katz 1995, Daille 2003), mais aussi de noms composés (Vincze et al. 2011) et d'expressions verbales (Baldwin 2005b). Si un analyseur syntaxique est disponible, les motifs syntaxiques peuvent être plus efficaces que les séquences d'étiquettes morphosyntaxiques, surtout lors de l'extraction d'EPL non-figées (Seretan and Wehrli 2009, Seretan 2008). Des grammaires de substitution d'arbres (Green et al. 2011) et des régularités structurelles dans les arbres d'analyse (Martens and Vandeghinste 2010) peuvent aussi être utilisées pour apprendre des modèles syntaxiques d'EPL à partir des corpus. L'algorithme LocalMaxs réalise une extraction fondée sur la maximisation d'une mesure d'association appliquée à des paires de mots adjacents (Silva and Lopes 1999). Un algorithme de correspondance de chaînes de caractères inspiré de la biologie informatique a été proposé pour extraire des séquences qui apparaissent de façon récurrente à travers le corpus (Duan et al. 2006).
2. **Filtrage de candidates** : quelques procédures simples pour le filtrage sont l'utilisation de listes de mots interdits et de seuils de nombres d'occurrences. Des mesures d'association sont souvent employées pour classer les candidates, de façon à ce que seulement les candidates dont la valeur d'association est en dessus d'un certain seuil soient conservées (Evert and Krenn 2005, Pecina 2005). Les coefficients optimaux des mesures d'association et des autres attributs des candidates peuvent être obtenus à l'aide de méthodes d'apprentissage supervisé (Ramisch et al. 2008b, Pecina 2008b).

Quelques outils disponibles gratuitement peuvent être utilisés pour l'acquisition des EPL dans des contextes monolingues : LocalMaxs, [1] Text : :NSP, [2] UCS, [3] jMWE, [4] et Varro. [5] Il existe aussi plusieurs services web disponibles gratuitement, ainsi que de nombreux outils téléchargeables et systèmes commercialisés pour l'extraction automatique de termes à partir des corpus spécialisés.

En ce qui concerne l'acquisition bilingue, les alignements lexicaux peuvent en eux-même fournir des listes d'EPL candidates (de Medeiros Caseli et al. 2010). Bai et al. (2009) présentent un algorithme capable de trouver des traductions pour une EPL donnée dans un corpus parallèle. La découverte automatique d'EPL non compositionnelles a

1. http://hlt.di.fct.unl.pt/luis/multiwords/
2. http://search.cpan.org/dist/Text-NSP
3. http://www.collocations.de/software.html
4. projects.csail.mit.edu/jmwe
5. http://sourceforge.net/projects/varro/

été explorée par Melamed (1997). La paire de langues hindi-anglais présente une grande variation dans l'ordre des mots, et il a été démontré que des attributs fondés sur la compositionnalité les EPL peuvent aider à réduire le taux d'erreur d'alignement lexical (Venkatapathy and Joshi 2006). Zarrieß and Kuhn (2009) utilisent un corpus parallèle aligné avec GIZA++ et analysé syntaxiquement pour extraire des paires du type verbe-objet à partir d'un corpus allemand-anglais. Daille et al. (2004) ont extrait des termes polylexicaux à partir de corpus comparables en français et en anglais, et ensuite ils ont utilisé les distances entre les vecteurs de contexte de ces termes pour obtenir des correspondances entre les langues.

A.3.2 Autres tâches dans le traitement des EPL

Il existe un nombre considérable de travaux publiés qui abordent d'autres tâches dans le traitement des EPL, résumés ci-dessous.

- **Interprétation** : L'*interprétation syntaxique* des noms composés a été explorée par Nicholson and Baldwin (2006), qui distinguent trois types de relations syntaxiques dans des composés du type nom–nom : sujet, objet direct et objet prépositionnel. Des noms composés avec trois mots ou plus doivent être interprétés de façon à découvrir leur hiérarchie de constituants. Nakov and Hearst (2005) comparent deux modèles, le modèle d'adjacences et le modèle de dépendances. Ils utilisent les comptages issus d'un mécanisme de recherche web pour estimer les probabilités de paraphrases générées par des heuristiques au niveau superficiel. Nakov and Hearst (2008) effectuent une *interprétation sémantique* non-supervisée des noms composés. Ils génèrent un grand nombre de paraphrases avec des verbes correspondant à chacune des classes sémantiques, et ensuite ils obtiennent leurs nombres d'occurrences dans le web. Kim and Nakov (2011) utilisent une combinaison de ré-échantillonnage et de comptage dans le web, avec des paraphrases fondées sur les arbres syntaxiques, pour obtenir des meilleurs résultats en interprétation sémantique. Cook and Stevenson (2006) utilisent des machines à vecteurs de support pour classifier les sens de la particule *up* dans des verbes à particules en anglais. Bannard (2005) quantifie la compositionnalité des verbes à particules par rapport à chacune de ses parties. Un travail similaire a été effectué par McCarthy et al. (2003), qui proposent plusieurs mesures fondées sur un thesaurus construit automatiquement pour estimer l'idiomaticité des verbes à particules.
- **Désambigüisation** : La désambigüisation des EPL est similaire à leur interprétation, sauf que les EPL sont considérées dans leur contexte d'occurrence. Nicholson and Baldwin (2008) ont créé un ensemble de données pour la désambigüisation des composés du type nom-nom, où un grand nombre de phrases a été manuellement annoté. Girju et al. (2005) étudient des méthodes pour leur désambigüisation à travers l'application de plusieurs techniques d'apprentissage supervisé. Fritzinger et al. (2010) ont analysé manuellement un grand nombre de constructions ambigües en allemand du type préposition–nom–verbe. Ils ont attribué une de ces trois classes à chaque construction : littérale, compositionnelle ou inconnue. Les verbes légers en japonais ont été étudiés par Uchiyama et al. (2005), qui proposent deux méthodes de désambigüisation : une approche statistique et une méthode par règles. Cook et al. (2007) explorent l'idiomaticité des paires du type verbe–nom, où le nom est l'objet direct du verbe et peut avoir une interprétation idiomatique (*faire la tête*) ou littérale (*faire un gâteau*). Fazly and Stevenson (2007) proposent une classification plus fine pour les constructions à verbe léger, avec une stratégie d apprentissage

supervisé et une séparation en quatre classes sémantiques.

– **Représentation** : La représentation lexicale des EPL est un problème qui, depuis longtemps, intrigue les lexicographes dans la compilation des ressources lexicales. Sag et al. (2002) ont proposé deux approches : mots-à-espaces et approche compositionnelle. Cependant, entre ces deux bouts du spectre de compositionnalité, il existe d'autres possibilités explorées dans la littérature. Laporte and Voyatzi (2008) décrivent un dictionnaire d'expressions adverbiales du français et leurs motifs morphosyntaxiques correspondants dans le formalisme lexique–grammaire. Graliński et al. (2010) comparent de manière qualitative et quantitative deux représentations structurées, POLENG et Multiflex, pour les EPL en polonais. Grégoire (2007; 2010) utilise une méthode de classes d'équivalence pour grouper des expressions similaires selon leurs caractéristiques syntaxiques. Izumi et al. (2010) suggèrent une méthode par règles capable de normaliser des expressions fonctionnelles en japonais, optimisant ainsi leur représentation. Schuler and Joshi (2011) proposent une description d'EPL à travers des grammaires de ré-écriture d'arbres.
– **Applications** : Dans quelques applications de TAL, des résultats concrets concernant les EPL ont été obtenus. Par exemple, en analyse syntaxique, Constant and Sigogne (2011) montrent des résultats prometteurs pour l'étiquetage morphosyntaxique du français. Korkontzelos and Manandhar (2010) obtiennent des améliorations considérables de qualité quand ils enrichissent un analyseur superficiel avec des entrées polylexicales. Zhang and Kordoni (2006) et Villavicencio et al. (2007) obtiennent une amélioration significative de couverture dans un analyseur du type HPSG en anglais lors de l'insertion d'EPL dans le lexique. Wehrli et al. (2010) démontrent que les EPL ne sont pas des « épines dans le pied » mais des informations qui aident à réduire les ambigüités syntaxiques. Un autre exemple d'application réussie des EPL est la récupération d'informations. Acosta et al. (2011) unissent les mots des EPL avant de réaliser l'indexation du corpus, obtenant ainsi une amélioration de la précision moyenne. Xu et al. (2010) proposent une nouvelle mesure de cohésion des séquences de quatre caractères en chinois, et obtiennent également des améliorations en termes de précision moyenne sur un ensemble de test.

A.4 Évaluation de l'acquisition d'EPL

Le problème d'évaluation de l'acquisition d'EPL est complexe parce que les résultats dépendent de plusieurs paramètres du contexte d'acquisition, de sorte que les résultats obtenus dans un contexte donné sont difficiles à généraliser. Dans la littérature, nous pouvons trouver plusieurs styles d'évaluation : analyser des listes triées des premières k EPL retournées (da Silva et al. 1999), annoter manuellement ces premières k EPL (Seretan 2008), mesurer la précision et le rappel par rapport à un dictionnaire (Ramisch 2009), comparer la qualité des mesures d'association à travers leur précision moyenne (Evert and Krenn 2005), comparer plusieurs approches (Pearce 2002, Ramisch et al. 2008a), et mesurer l'impact des EPL acquises dans des applications de TAL (Finlayson and Kulkarni 2011, Xu et al. 2010, Carpuat and Diab 2010). Afin de fournir un cadre d'évaluation plus structuré, nous proposons une nouvelle typologie pour classifier le *contexte d'évaluation*.

1. **Selon l'objectif de l'acquisition :**
   – **Intrinsèque**. Les résultats considèrent l'évaluation des EPL en elles-mêmes, directement, en tant que produit final d'un processus. L'évaluation intrinsèque est fortement dépendante de l'application cible et de la cohérence des instructions

d'annotation, mais elle donne tout de même une estimation de qualité utile des EPL acquises.
- **Extrinsèque**. L'évaluation extrinsèque consiste à intégrer les EPL dans une application de TAL extérieure et vérifier si elles améliorent la qualité du résultat produit par l'application. Éventuellement, il peut être plus facile d'estimer la qualité du résultat pour une tâche de TAL concrète que pour une liste d'EPL dont on ne connait pas l'application. Cette évaluation peut être très concluante pour démontrer si les EPL acquises sont utiles.

2. **Selon la nature des mesures :**
   - **Quantitative**. Cela consiste à utiliser des mesures objectives telles que la précision, le rappel, la F-mesure et la précision moyenne. Alors que de nombreux articles calculent uniquement la précision sur les premières k EPL retournées, il faut aussi évaluer le rappel, car la quantité de (nouvelles) EPL découvertes est un facteur aussi important que leur qualité.
   - **Qualitative**. Le but est d'obtenir une compréhension approfondie des erreurs commises par la méthode d'acquisition. Cela consiste à observer les motifs récurrents en analysant les listes résultantes en termes de comportement syntaxique, distribution de fréquences, contexte, etc. Les analyses quantitative et qualitative sont complémentaires, et sont souvent effectuées de façon simultanée et/ou itérative.
3. **Selon les ressources disponibles :**
   - **Annotation manuelle**. Un groupe de locuteurs natifs et/ou d'experts parcourra la liste d'EPL, jugeant pour chaque combinaison proposée s'il s'agit d'une vraie EPL. Cette annotation peut demander beaucoup de temps selon la disponibilité des annotateurs, et est souvent effectuée sur un échantillon de la sortie.
   - **Annotation automatique**. Dans l'annotation automatique, nous considérons qu'il existe un dictionnaire complet ou, au moins, avec une très bonne couverture, des expressions cibles. Ainsi, les candidates qui apparaissent dans le dictionnaire sont des vraies positives (des EPL authentiques/intéressante), tandis que les autres sont des fausses EPL.
4. **Selon le type d'EPL :**
   - **Fondée sur les types**. Certaines expressions non ambiguës, comme les noms composés, les termes techniques et les constructions à verbe support, peuvent être annotées hors contexte. Il existe plusieurs lexiques disponibles qui peuvent être employés comme référence standard dans l'annotation fondée sur les types. Si une telle ressource n'existe pas, l'annotation doit être effectuée manuellement.
   - **Fondée sur les occurrences**. Cette annotation doit être effectuée quand les EPL cibles sont ambiguës, comme les verbes à particule et les expressions idiomatiques. Hors contexte, il est impossible de dire si les mots doivent être traités comme une unité ou indépendamment. Dans ce type d'annotation, les juges humains annotent une phrase entière, en opposition à une EPL candidate isolée du contexte.

Si nous modélisons le résultat de l'acquisition d'EPL sous forme d'une liste C de candidates triées selon un score numérique donné, la précision P(C) du système est la proportion de candidates jugées comme des vraies EPL dans l'ensemble de candidates retournées, $\mathrm{P}(C) = \frac{|\text{EPL dans } C|}{|C|}$. La précision indique la quantité de travail nécessaire pour transformer la liste brute d'EPL acquises automatiquement dans une liste définitive va-

lidée par un spécialiste. Cependant, la précision ne tient pas compte des vraies EPL qui n'ont pas été trouvées quand elles le devraient. Par conséquent, il est essentiel de calculer le rappel $\mathrm{R}(C) = \frac{|\text{EPLs dans } C|}{|\text{Total d'EPL à acquérir}|}$. En dépit de son importance, R(C) est rarement calculé car il est difficile d'estimer le nombre total d'EPL qui devraient être acquises par un système.

Il existe deux styles d'annotation : automatique et manuelle. Dans l'annotation automatique, il y a un *standard de référence*, c'est-à-dire, un lexique contenant la liste complète des EPL qui doivent être trouvés. Dans l'annotation automatique, P(C) et R(C) sont sous-estimés car ils supposent que les candidates absentes du standard de référence sont des fausses EPL. En dépit de cette simplification, l'annotation automatique est souvent utilisée, principalement parce qu'elle est rapide et peu onéreuse. L'annotation manuelle est rarement réalisée sur toute la liste d'EPL retournées, mais plutôt sur un échantillon. Si la liste est classée, les k premières candidates peuvent être annotées, mais cela introduit un biais en faveur des combinaisons très fréquentes alors que l'échantillon devrait inclure des candidates de tous les intervalles de nombres d'occurrences. Il est important de bien concevoir les instructions d'évaluation données aux annotateurs, qui sont un groupe de locuteurs natifs ou, si les EPL cibles sont complexes, des experts linguistes. Il est recommandé de laisser une certaine marge de manœuvre aux annotateurs, par exemple, en préférant les catégories à plusieurs valeurs ou les échelles numériques aux décisions binaires. Le score kappa de Fleiss est souvent utilisé pour estimer l'accord inter-annotateur, même si son interprétation est controversée. Les annotations manuelle et automatique sont complémentaires. Il est possible d'utiliser l'annotation mixte, par exemple, annoter manuellement les entrées absentes du standard de référence.

Le *contexte d'acquisition* est l'ensemble de paramètres qui peuvent influencer les résultats de l'évaluation. Nous affirmons que les résultats d'une évaluation effectuée dans un contexte d'acquisition donné sont difficiles à généraliser car ils dépendent d'un nombre trop important de paramètres.

Certains paramètres du contexte d'acquisition dépendent des caractéristiques des EPL, comme :

- **Type**. Différents types d'EPL exigent des évaluations différentes. Par exemple, les séquences d'étiquettes morphosyntaxiques sont souvent employés pour l'acquisition des noms composés, mais ne génèrent pas de bons résultats avec les expressions verbales (Villavicencio et al. 2012).
- **Langue**. Non seulement les EPL mais aussi les ressources de TAL ne sont pas équivalentes dans toutes les langues. L'utilisation d'un analyseur syntaxique pour l'acquisition de collocations, comme dans Seretan (2008), par exemple, est impossible pour les langues peu dotées, pour lesquelles une telle ressource n'existe pas, nécessitant des solutions alternatives fondées sur une analyse superficielle.
- **Domaine**. Le domaine de l'expression doit être pris en compte lors de l'évaluation. Par exemple, les listes de séquences d'étiquettes morphosyntaxiques proposées par Justeson and Katz (1995) ne donnent pas de bons résultats quand elles sont directement appliquées à un corpus du domaine biomédical (Ramisch 2009).

Certains paramètres du contexte d'acquisition dépendent des caractéristiques des corpus, comme :

- **Taille**. Des grands corpus contiennent plus de données, donc intuitivement une méthode sera capable de récupérer plus de candidates, ce qui augmente son rappel. De même, des méthodes statistiques peuvent être sensibles à des données creuses, et des échantillons plus grands permettent d'avoir des mesures plus précises.

– **Nature**. Les résultats d'évaluation dépendent du domaine et du genre de textes. Par exemple, les expériences montrent que, dans l'extraction de noms composé spécialisés, l'utilisation du web comme corpus n'est pas recommandée (Ramisch et al. 2010d).
– **Niveau d'analyse**. Les méthodes d'acquisition varient entre les méthodes peu profondes et pauvres en connaissances (da Silva et al. 1999) et les méthodes profondes dépendant d'un formalisme syntaxique spécifique (Seretan 2008). Il n'est pas toujours vrai qu'une analyse plus profonde donne de meilleurs résultats (Baldwin 2005b).

L'évaluation de l'acquisition d'EPL demeure un problème ouvert. Si d'une part les mesures telles que la précision et le rappel lors de l'annotation automatique supposent l'existence d'un standard de référence complet, d'autre part l'annotation manuelle est souvent très couteuse et donne plus d'importance à la précision qu'au nombre de nouvelles EPL acquises. Certains articles décrivent des évaluations comparatives (Schone and Jurafsky 2001, Pecina 2005, Ramisch et al. 2008a) et récemment un certain nombre de travaux publiés réalise des évaluations extrinsèques dans des applications de TAL telle que la récupération d'informations (Doucet and Ahonen-Myka 2004, Xu et al. 2010, Acosta et al. 2011), la désambigüisation lexicale (Finlayson and Kulkarni 2011), la traduction automatique (Carpuat and Diab 2010, Pal et al. 2010) et l'apprentissage d'ontologies (Venkatsubramanyan and Perez-Carballo 2004).

A.5 Une plate-forme pour l'acquisition d'EPL

Nous introduisons une nouvelle plate-forme appelée `mwetoolkit`, qui intègre de multiples techniques et couvre l'ensemble du pipeline d'acquisition d'EPL. Le fonctionnement de la plate-forme est détaillé dans le schéma de la figure 5.1. Davantage de détails sont fournis sur le site web de l'outil et dans des publications précédentes (Ramisch et al. 2010b;c). Le fonctionnement de la plate-forme est résumé ci-dessous :

1. Avant de traiter un corpus monolingue brut, il est possible de le prétraiter, si les outils de prétraitement sont disponibles pour la langue cible, en l'enrichissant avec des étiquettes morphosyntaxiques, des lemmes et de la syntaxe de dépendances.
2. Ensuite, on décrit les EPL cibles en définissant des motifs multiniveaux qui reposent sur des connaissances linguistiques expertes, sur l'intuition, sur l'observation empirique et/ou sur des exemples, dans un formalisme similaire aux expressions régulières.
3. L'application de ces motifs sur un corpus indexé génère une liste d'EPL candidates.
4. Pour le filtrage, une multitude de méthodes est disponible, allant de simples seuils de nombres d'occurrences à des listes de mots interdits et des mesures d'association sophistiquées.
5. Enfin, les candidates filtrées sont soit directement injectées dans une application de TAL, soit validées manuellement avant l'application. Une autre utilisation pour les candidates validées est la création d'un modèle d'apprentissage automatique, qui peut être appliqué sur des nouveaux corpus afin d'identifier et d'extraire automatiquement des EPL en fonction des caractéristiques de celles déjà acquises.

À ce jour, il n'y a pas de consensus sur une méthode optimale d'acquisition d'EPL. Il n'est donc pas possible de déterminer si il existe une méthode unique pour toutes les EPL, ou alors s'il faudrait chercher une combinaison de méthodes ou un sous-ensemble de

méthodes qui fonctionne mieux pour un type d'EPL en particulier. La contribution principale de la plate-forme et de l'outil proposés est l'intégration systématique des processus et des tâches requises pour l'acquisition qui proportionne une vue globale de la chaîne de traitement d'EPL. Un de ses avantages réside dans le fait qu'ils modélisent le processus d'acquisition par des tâches modulaires, étant ainsi hautement personnalisables et permet un paramétrage détaillé. Le `mwetoolkit` peut être utilisé pour accélérer le travail de lexicographes et terminographes, ainsi que pour aider à l'adaptation des applications de TAL à d'autres langues et à d'autres domaines. La méthodologie employée dans le toolkit est indépendante de la langue car elle n'est pas fondée sur des connaissances symboliques ou sur des dictionnaires existants. De plus, les techniques développées ne dépendent pas d'une longueur fixe d'expressions candidates (par exemple, les paires de mots) ni sur l'hypothèse de contiguïté. Grâce à cette souplesse, cette méthodologie peut être facilement appliquée à un grand nombre de langues, de types d'EPL et de domaines, ne dépendant pas d'un formalisme donné ou d'un outil. En somme, le `mwetoolkit` permet aux utilisateurs de réaliser une acquisition d'EPL systématique avec des étapes intermédiaires consistantes et avec des modules et des arguments bien définis.

Nous avons comparé le `mwetoolkit` avec trois autres outils disponibles gratuitement, téléchargeables et documentés : l'implémentation de référence du LocalMaxs, (`LocMax`), le *N-gram statistics package* (`NSP`) et la boîte à outils `UCS`. Nous avons exploré l'acquisition des expressions verbales et nominales en anglais (`en`) et des expressions nominales en français (`fr`). Les EPL acquises ont été évaluées automatiquement à travers la comparaison avec des standards de référence.

La qualité des candidates extraites du corpus de taille moyenne varie selon les types d'EPL et les langues, comme le montre la figure 5.5. Pour les EPL nominales, les approches ont des résultats similaires, avec un rappel élevé et une faible précision. Pour les expressions verbales, le `LocMax` a la plus haute précision (environ 70%), mais un faible rappel, tandis que les autres approches ont des valeurs plus équilibrées de P et de R, autour de 20%. Les techniques diffèrent en termes de stratégie d'extraction : (i) le `mwetoolkit` et le `NSP` permettent de définir des filtres linguistiques tandis que le `LocMax` permet d'appliquer de filtres externes (*grep/sed*) uniquement après l'acquisition, (ii) il n'ya pas de filtrage préliminaire dans le `mwetoolkit` ni dans le `NSP` car ils renvoient toutes les candidates correspondant aux motifs, alors que le `LocMax` filtre les candidates a priori en fonction du critère de maximum local et (iii) le `LocMax` extrait uniquement les candidates contigües tandis que les autres outils permettent l'extraction de candidates non contigües. L'évaluation des candidates nominales `fr` selon la taille du corpus est montrée dans le tableau 5.4. Pour toutes les approches, la précision diminue lorsque la taille du corpus augmente, tandis que le rappel augmente pour toutes les approches sauf pour le `LocMax`.

Le tableau 5.6 présente les résultats de l'évaluation des mesures d'association. La mesure `glue` du `LocMax` présente la meilleure précision moyenne parmi toutes les mesures testées, ce qui suggère que le critère de maximum local est un bon indicateur d'EPL et que `glue` est une mesure efficace pour générer des résultats très précis. Pour le `mwetoolkit`, la meilleure mesure testée a été `dice`, tandis que les autres mesures ne sont pas systématiquement meilleures que la ligne de base. La mesure de Poisson-Stirling (`Poisson`) a obtenu des valeurs de précision moyenne assez bonnes, mais les deux autres mesures testées pour le `NSP` ont été en dessous de la ligne de base dans certaines configurations. Finalement, toutes les mesures appliquées par l'`UCS` ont obtenu des performances supérieures à celles de la ligne de base et, pour les EPL nominales, sont

comparables à la meilleure mesure d'association.

Des aspects tels que le degré de variabilité de l'EPL et la performance de calcul influencent sur la décision de la (ou des) mesure(s) d'association adoptée(s). Par exemple, la mesure `dice` peut être facilement appliquée à tous les *n*-grammes, tandis que des mesures plus sophistiquées comme `Poisson` ne sont définies que pour les 2-grammes et sont parfois lourdes à calculer. L'`UCS` n'extrait pas les candidates à partir du corpus, mais prend en entrée une liste de 2-grammes. Le `NSP` étend une partie des mesures d'association disponibles aux 3- et 4-grammes, et le `mwetoolkit` et le `LocMax` n'ont aucune contrainte sur la longueur du *n*-gramme. Le `LocMax` extrait uniquement des EPL contigües tandis que le `mwetoolkit` et le `NSP` permettent l'extraction de mots non adjacents. Seulement le `mwetoolkit` intègre des filtres linguistiques sur les lemmes, les étiquettes morpho-syntaxiques et la syntaxe. Ceci peut être simulé en utilisant des outils externes (*grep/sed*) sur la sortie des autres systèmes.

Le `mwetoolkit` est une première étape importante vers un traitement d'EPL robuste et fiable dans les applications de TAL. Il est également un logiciel de base, disponible gratuitement, doté d'outils puissants et d'une documentation actualisée et cohérente. Ces dernières sont des caractéristiques essentielles pour l'extension et la maintenance de tout logiciel.

A.6 Application 1 : lexicographie

Dans le contexte de la lexicographie assistée par ordinateur, nous avons effectué une première évaluation quantitative et qualitative de la plate-forme proposée pour l'acquisition d'EPL. Pour cela, nous avons compté sur la participation de collègues linguistes et lexicographes expérimentés dans la création de ressources lexicales en portugais et en grec. Les ensembles de données créés sont gratuitement disponibles. [6]

Pour le grec, il existe une vaste littérature portant sur les propriétés linguistiques des EPL, mais les approches informatiques sont encore limitées (Fotopoulou et al. 2008). Dans nos expériences, nous avons utilisé le `mwetoolkit` pour extraire de la partie grecque du corpus Europarl, étiquetée morpho-syntaxiquement, des noms composés (NC) correspondant aux motifs suivants : adjectif-nom, nom-nom, nom-déterminant-nom, nom-préposition-nom, préposition-nom-nom, nom-adjectif-nom et nom-conjonction-nom. Les candidates ont été comptées dans deux corpus et classées par quatre mesures d'association. Les premières 150 candidates selon chaque mesure d'association ont été évaluées par trois locuteurs natifs. Ainsi, chaque annotateur a jugé environ 1 200 candidates. Finalement, les annotations ont été combinées, entraînant la création d'un lexique avec 815 EPL nominales en grec.

Avec ces annotations, nous avons analysé la contribution exacte des différentes mesures d'association dans la liste des EPL retournées. La mesure qui a produit le meilleur résultat a été `dice`, qui a eu une performance significativement meilleure que les autres mesures. La mesure `t-score` a été la deuxième meilleure, mais étonnamment sa performance est très similaire à celle des comptages bruts des EPL, suggérant ainsi que les mesures d'association sophistiquées ne sont pas nécessaires quand des corpus suffisamment grands sont disponibles. En ce qui concerne l'utilisation du web comme un corpus, cela a de nombreux avantages par rapport aux corpus traditionnels, les plus marquants étant son accessibilité et sa disponibilité. Cependant, dans nos expériences, les résultats

6. `http://multiword.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets`

obtenus avec les comptages provenant du web n'ont pas apporté d'amélioration considérable. En somme, nos résultats indiquent que des méthodes automatiques peuvent en effet être utilisées pour étendre des ressources de TAL avec des EPL, améliorant ainsi la qualité des systèmes de TAL en grec.

L'objectif du travail avec les prédicats complexes en portugais était de réaliser une analyse qualitative de ces constructions. Nous avons généré deux ressources lexicales ciblées sur deux applications : CP-SRL est destiné à l'annotation d'étiquettes de rôle sémantique, et CP-SENT est destiné à l'analyse de sentiments. Pour créer ces deux ressources, nous avons étiqueté morpho-syntaxiquement le corpus PLN-BR-Full et ensuite nous avons extrait des séquences de mots correspondant à des motifs morphosyntaxiques spécifiques avec le `mwetoolkit`.

L'annotation des étiquettes de rôles sémantiques dépend de l'identification correcte du prédicat, avant d'identifier les arguments et affecter les étiquettes de rôles sémantiques. Pourtant, plusieurs prédicats ne sont pas constitués par un seul verbe : il s'agit de prédicats complexes qui ne sont pas toujours présents dans les lexiques informatisés. Pour créer le dictionnaire CP-SRL, nous avons utilisé des séquences d'étiquettes morphosyntaxiques plutôt qu'une liste limitée de verbes et de noms : verbe-[déterminant]-nom-préposition, verbe-préposition-nom, verbe-[préposition/déterminant]-adverbe et verbe-adjectif. Le processus d'extraction a engendré la création d'une liste avec 407 014 EPL candidates qui ont par la suite été filtrées avec des mesures d'association. Un annotateur humain expert a validé manuellement 12 545 candidates, dont 699 on été annotées comme des prédicats complexes compositionnels tandis que 74 ont été annotées comme des prédicats complexes idiomatiques. Les résultats incluent (mais ne se limitent pas à) des constructions à verbe support et à verbe léger. Nous avons observé les paires de paraphrases suivantes :

- V = V + N DÉVERBAL : *tratar* = *dar tratamento* (lit. *traiter* = *donner un traitement*) ;
- V DÉNOMINAL = V + N : *amedrontar* = *dar medo* (lit. *effrayer* = *donner de l'effroi*) ;
- V DÉADJECTIVAL = V + ADJ : *responsabilizar* = *tornar responsável* (lit. *responsabiliser* = *rendre responsable*).

Pour la création de CP-SENT, notre objectif était d'étudier la façon dont les sentiments sont exprimés en portugais brésilien. Des verbes de sentiment tels que *temer (craindre), odiar (haïr)* et *invejar (envier)* sont des exemples d'unités lexicales spécifiquement utilisées pour exprimer des sentiments. Le même sens peut être exprimé par d'autres verbes associés à des noms de sentiment. Cette étude identifie tout d'abord sept motifs récurrents d'expression de sentiments sans verbe de sentiment, et puis emploie ces motifs pour identifier des noms de sentiment associées. Ceci a été réalisé en cinq étapes.

Premièrement, nous avons identifié des motifs lexico-syntaxiques pour exprimer des sentiments en utilisant des noms de sentiment au lieu des verbes de sentiment. Deuxièmement, nous avons utilisé les motifs identifiés comme arguments de recherche pour identifier l'expression de sentiments dans le corpus. Troisièmement, un humain a analysé les listes de candidates résultantes de la deuxième étape, déterminant si le nom apparaissant à droite de chaque motif était ou non un nom de sentiment. Quatrièmement, nous avons analysé les candidates validées et nous avons annotés certains traits tels que la polarité et la source du sentiment. Cinquièmement, nous avons combiné les motifs de la première étape avec les noms de sentiment identifiés par la troisième étape, et nous avons recherché ces nouvelles combinaisons sur le web. L'analyse des motifs a montré que la combinaison de noms de sentiment avec les sept motifs peuvent être utiles pour identifier

automatiquement l'expression de sentiments et, de plus, aider à identifier la personne qui a un sentiment et ce qui est à l'origine du sentiment.

A.7 Application 2 : traduction automatique empirique

En guise de deuxième évaluation du `mwetoolkit`, nous avons effectué des expériences sur la traduction vers le portugais des verbes à particule en anglais comme *give up (renoncer)* et *get by [a name] (répondre au nom de)*, en utilisant un système de traduction automatique (TA) empirique. La traduction des verbes à particule est un défi, car ils présentent une grande variabilité syntaxique et sémantique. Les verbes à particule sont très fréquents en anglais, survenant dans environ 17% des phrases de notre corpus. La prise en compte du comportement syntaxique et sémantique complexe des verbes à particules dans des systèmes de TA empirique actuels, qui possèdent des lexiques plats fondés sur les séquences de mots adjacents, n'est pas simple. Néanmoins, il est important de les identifier et d'avoir un traitement adéquat pour eux afin d'éviter la génération de traductions qui sonnent peu naturelles ou agrammaticales.

La représentation et l'intégration des EPL dans les systèmes de traduction automatique a été l'objet de nombreuses recherches. Le système de TA ITS-2 traite les EPL à deux niveaux : lors de l'analyse lexicale pour les expressions contiguës, et lors de l'analyse syntaxique pour les collocations (Wehrli 1998, Wehrli et al. 2010). Carpuat and Diab (2010) adoptent deux stratégies complémentaires pour intégrer les EPL dans un système de TA empirique : une stratégie statique de tokenisation unique, qui traite les EPL comme des mots-à-espaces, et une stratégie dynamique qui rajoute le nombre d'EPL identifiées dans le segment source en tant qu'attribut du modèle de traduction. Morin and Daille (2010) obtiennent une amélioration de 33% dans la traduction des EPL en français–japonais avec une méthode morphologique compositionnelle pour faire le *backoff* lorsqu'il n'y a pas suffisamment de données dans un dictionnaire pour traduire une EPL. Pour la traduction de et vers des langues morphologiquement riches comme l'allemand, où un nom composé est en fait un mot unique formé par concaténation, Stymne (2011) divise le mot composé en ses composantes d'un seul mot avant la traduction, et applique ensuite des règles de post-traitement, comme le ré-ordonnancement ou la fusion des composantes, après la traduction. Une autre approche pour minimiser la dispersion des données est adopté par Nakov (2008a), qui génère des paraphrases monolingues pour augmenter le corpus d'apprentissage.

Dans nos expériences, un système de TA empirique fondé sur les segments et non factorisé a été construit à l'aide de la boîte à outils Moses avec des paramètres standard sur le corpus Europarl v6 en anglais–portugais. Les verbes à particule ont été automatiquement identifiés à l'aide de l'outil jMWE et d'un dictionnaire de verbes à particule. Nous avons comparé cinq stratégies pour l'intégration des verbes à particule automatiquement identifiés dans le système de TA. L'ensemble de test est constitué d'un échantillon de 1 000 phrases, dont la moitié contient des verbes à particules. Les constructions les plus fréquentes identifiées en anglais incluent *lay down, set up, carry out* and *originate in*.

Puisque nous étudions un phénomène linguistique complexe, aucune de nos conclusions n'auraient pu être tirées en vue uniquement de mesures automatiques comme BLEU et NIST, sans une analyse d'erreurs minutieuse par des annotateurs humains sur les résultats de la traduction automatique. Dans les traductions annotées, quelques problèmes récurrents et communs à toutes les stratégies testées incluent des erreurs d'accord de conjugaison verbale, des particules et/ou prépositions mal choisies, la traduction d'un verbe

comme un substantif et des prépositions parasites ajoutées au verbe cible. Les résultats préliminaires de l'évaluation humaine effectuée sur un ensemble de test de 100 phrases ont montré que, tandis que certaines traductions sont améliorées par les stratégies d'intégration, d'autres sont dégradées. Aucune amélioration absolue a été observée, mais nous pensons que cela est dû au fait que notre évaluation doit prendre en considération des classes plus fines de verbes à particule, au lieu de les mélanger dans le même jeu de test. De plus, nous aurions besoin d'annoter plus de données afin d'obtenir des résultats plus représentatifs.

Nous avons découvert qu'il y a une corrélation entre la qualité des traductions générées par chaque stratégie et la compositionnalité des verbes à particule. Les stratégies qui ont produit les meilleurs résultats pour le cas idiomatique sont TOK et BILEX. Pour le cas compositionnel, TOK a entraîné une diminution de qualité. Bien que la stratégie PV ? tende à traduire les verbes à particule comme des unités, elle est moins drastique que TOK, et produit ainsi moins de mauvaises traductions pour le cas compositionnel. La comparaison de ces heuristiques indique qu'elles fournissent des informations complémentaires, qui semblent être liées à la compositionnalité et au nombre d'occurrences du verbe à particule. Ces hypothèses nous motivent à continuer notre enquête afin d'obtenir une meilleure compréhension de l'impact de chaque stratégie d'intégration sur chaque étape du système de TA.

A.8 Conclusions et travaux futurs

Nous avons décrit les objectifs principaux de notre travail comme étant : (a) développer des techniques pour l'acquisition automatique des EPL à partir des corpus, (b) évaluer ces techniques extrinsèquement en mesurant leur utilité dans des applications de TAL, et (c) étudier l'acquisition et l'intégration des EPL dans des contextes multilingues. À ce jour, l'objectif (a) peut être considéré comme atteint, et le logiciel résultant, le `mwetoolkit`, est disponible gratuitement. L'évaluation de l'acquisition des EPL est un problème ouvert, et nous avons proposé une classification théorique afin d'aider à mieux structurer la description de ce problème. En ce qui concerne l'objectif (b), nous le considérons comme atteint car nous avons démontré l'utilité de notre plate-forme dans le développement de trois ressources lexicales différentes. D'autres applications où le `mwetoolkit` a été utilisé n'ont pas été discutées ici dans un souci de synthèse (Villavicencio et al. 2012, Granada et al. 2012). Finalement, pour l'objectif (c), nous avons présenté des résultats préliminaires sur l'intégration des EPL dans un système de TA empirique. Ces derniers sont issus d'un travail en cours pour lequel nous avons des expériences et des améliorions prévues prochainement.

Les expériences en cours et futures ont deux axes principaux :

– **Meilleure intégration des EPL dans les systèmes de TA** : L'intégration des verbes phrasaux dans les systèmes de TA est un problème difficile à résoudre en raison de la variabilité de ces constructions. Nos expériences ont montré que, tandis que le système de TA standard arrive à traduire certains verbes phrasaux correctement, il tend à commettre des erreurs quand la construction est idiomatique ou séparée. Nous avons l'intention d'étudier des alternatives pour insérer les entrées polylexicales dans le tableau de segments, par exemple à travers l'estimation des probabilités pour ces nouvelles entrées (Bouamor et al. 2011), les utiliser pour guider l'alignement lexical, post-traiter la sortie de la traduction, ou rajouter les EPL au corpus d'entraînement. Nous avons aussi l'intention d'améliorer la détection des

verbes phrasaux à travers l'emploi d'un analyseur plus profond qui peut capturer les dépendances de longue distance sur des expressions syntaxiquement variables (Seretan 2008, Baldwin 2005a). Potentiellement, l'information syntaxique peut fournir de nouveaux attributs au modèle de traduction. La détection de la compositionnalité des verbes phrasaux fondée sur les corpus (McCarthy et al. 2003, Bannard et al. 2003, Baldwin et al. 2003) pourrait aider à générer des traductions plus exactes. Nous voulons aussi étudier d'autres constructions polylexicales qui ont un impact sur les équivalences et asymétries entre les langues. Notre objectif à long terme est d'intégrer les EPL dans les systèmes de TA dans le but d'obtenir des traductions de haute qualité en combinant des informations statistiques et linguistiques.

– **Projet CAMELEON** : Un des résultats de cette thèse est le projet CAMELEON, financé par l'allocation 707-11 de l'appel CAPES-COFECUB. [7] Son objectif principal est d'étudier des techniques automatiques et collaboratives dans le développement de ressources ontologiques et lexicales pour les applications multilingues. Nous voulons étudier l'apport d'un système de gestion collaborative de ressources lexicales dans le filtrage et la validation les EPL acquises automatiquement. Nous avons également des expériences en cours et certains résultats préliminaires publiés dans l'acquisition automatique d'un corpus comparable représentant un échantillon du langage utilisé dans les conférences scientifiques en français, portugais et anglais (Granada et al. 2012). Simultanément, nous testons la faisabilité d'une approche fondée sur le jeu lexical sérieux JeuxDeMots pour la construction d'une ressource lexicale du portugais [8] (Mangeot and Ramisch 2012). Ces expériences constituent un environnement expérimental intéressant pour des recherches futures sur le rôle des EPL dans les ressources et applications créées dans ces trois langues.

En dépit d'un important effort de recherche dans ce domaine, le traitement d'EPL dans les applications de TAL est encore un problème ouvert et un grand défi à relever. Bien sûr, ceci n'est pas vraiment une surprise dans la mesure où la nature complexe et hétérogène des EPL a été démontrée par de nombreuses études linguistiques. Au début des années 2000, Schone and Jurafsky (2001) ont posé la question si l'identification automatique d'EPL était un problème résolu, et la réponse que cet article apporta à l'époque fut négative. De même, des publications spécialisées plus récentes montrent des indices que cela est encore vrai aujourd'hui. Par exemple, les préfaces des derniers numéros de revue consacrés aux EPL (Villavicencio et al. 2005b, Rayson et al. 2010b) et des annales de l'atelier MWE (Kordoni et al. 2011a) listent plusieurs défis dans le traitement des EPL tels que le multilinguisme, la représentation dans les lexiques et l'évaluation fondée sur les applications.

Une des contributions principales de ce travail réside dans le fait qu'il représente une étape vers l'intégration des EPL automatiquement extraites dans des applications de TAL réelles. Néanmoins, étant donné la complexité du problème, ce traitement doit être continuellement amélioré, car il semble peu probable que, dans un avenir proche, on puisse proposer une solution définitive et unifiée pour le traitement des EPL dans les applications de TAL. Ainsi, à long terme, notre objectif peut être résumé comme étant d'améliorer et étendre le travail présenté ici. Car si, d'une part, nous avons effectué un premier pas important, d'autre part la route qui reste à parcourir est encore longue.

7. http://cameleon.imag.fr
8. http://jeuxdemots.imag.fr/por

APPENDIX B RESUMO ESTENDIDO

UMA PLATAFORMA GENÉRICA E ABERTA PARA O TRATAMENTO DAS EXPRESSÕES POLILEXICAIS: DA AQUISIÇÃO ÀS APLICAÇÕES

B.1 Introdução

Devido à sua natureza complexa, as expressões polilexicais (EPLs, do inglês *multiword expressions*) constituem um grande desafio para o processamento de linguagem natural (PLN). A definição precisa dos fenômenos linguísticos que podem ser considerados como EPLs é uma questão polêmica. Simplificando os pormenores teóricos, EPLs são combinações de palavras habituais e recorrentes da linguagem do dia a dia (Firth 1957). Provavelmente os exemplos mais prototípicos de EPLs são as expressões idiomáticas como *mente aberta*, *quebrar um galho*, *lavar roupa suja*, *bater as botas*, *sem eira nem beira*, *cara de pau*, *amigo da onça* e *barra pesada*. Além das expressões idiomáticas, muitos outros tipos de construções podem ser considerados como EPLs. Outros exemplos de EPLs incluem substantivos compostos (por exemplo, *aspirador de pó*, *secretária eletrônica* e *sapato de salto alto*) e expressões verbais (por exemplo, *fazer sentido*, *tirar vantagem*, *tomar banho* e *dar-se conta*).

Falantes nativos de uma língua raramente se dão conta do número de expressões institucionalizadas que fazem parte do discurso coloquial, como *bom dia*, *nem te conto*, *até mais* e *depois de amanhã*. Pode-se assumir que o léxico de um falante nativo contém tantas entradas polilexicais quanto o número de entradas que correspondem às palavras simples (Jackendoff 1997). Portanto, qualquer sistema computacional que se proponha a processar a linguagem humana deve levar EPLs em consideração. Efetivamente, em diversas aplicações de PLN, quando as palavras que compõem uma EPL são tratadas como unidades separadas, o sistema pode ser induzido a produzir resultados errôneos. Por exemplo, um sistema de tradução automática deve detectar EPLs para evitar traduções literais.

Integrar EPLs em sistemas de PLN tradicionais é uma tarefa complicada pois as expressões polilexicais se encontram na fronteira entre o léxico e a sintaxe das linguagens. Consequentemente, os recursos linguístico-computacionais disponíveis para o tratamento das EPLs são limitados tanto em termos de qualidade quanto em termos de cobertura, contrastando com a natureza onipresente dessas expressões. Há, portanto, uma grande necessidade para se desenvolver, consolidar e avaliar técnicas para a aquisição automática de EPLs a partir de corpora textuais.

Esta tese descreve o tratamento dado às EPLs em aplicações de PLN, cobrindo desde sua aquisição automática a partir de textos brutos até sua integração em duas aplicações reais: lexicografia assistida por computador e tradução automática empírica. Desenvolveu-

se um modelo conceitual para o pipeline de tratamento de EPLs, assim como um pacote de software completo que valida a metodologia proposta. Esse modelo foi avaliado de maneira rigorosa e sistemática. Pode-se resumir os objetivos da presente tese da maneira seguinte:

1. Desenvolver técnicas portáveis e genéricas para a aquisição automática de EPLs a partir de corpora.
2. Realizar a avaliação extrínseca dessas técnicas, ou seja, quantificar sua utilidade em aplicações reais de PLN.
3. Investigar a aplicabilidade dessas técnicas em contextos bilíngues e multilíngues, estudando de que forma os diferentes parâmetros do contexto de aquisição influenciam a qualidade das EPLs adquiridas automaticamente.

B.2 Definições e características

O estudo das expressões polilexicais é quase tão antigo quanto a própria linguística. Quando tentamos classificar os diversos fenômenos linguísticos existentes entre léxicos e sintáticos, rapidamente descobrimos que alguns deles, e em particular as EPLs, encontram-se intuitivamente em algum ponto intermediário entre esses dois níveis. Portanto, as EPLs mostram que existem limitações na abordagem estrutural da língua de Chomsky e de Tesnière. Um dos artigos que deu origem à corrente linguística dita *construcionista* é o trabalho de Fillmore et al. (1988). Nesse artigo, os autores ilustram com detalhes alguns dos problemas inerentes à abordagem atomística e idealizada da gramática. Na gramática construcionista, os idiomas são um elemento central: uma língua pode ser completamente descrita por meio dos seus idiomas e das suas propriedades. Esses idiomas correspondem às EPLs no presente trabalho. Outra teoria linguística que coloca bastante ênfase nas EPLs é a *teoria sentido-texto* (MTT, do inglês *meaning-text theory*). EPLs ocorrem em dois pontos do modelo computacional da MTT: como *frasemas* e como funções léxico-semânticas na *zona de combinatória lexical*. Para uma visão sintética e ampla do tratamento dado às EPLs em diversas teorias linguísticas, recomenda-se a leitura de Seretan (2008, p. 20–27).

É difícil definir as EPLs, pois existe uma grande controvérsia em torno da definição da própria palavra *palavra*. A noção de EPL é originária da famosa frase de Firth "diga-me com quem andas e eu te direi que palavra és". Ele afirma que "colocações de uma dada palavra são declarações do lugar habitual e convencional daquela palavra" (Firth 1957, p. 181). Smadja (1993) considera colocações como "combinações de palavras arbitrárias e recorrentes". Para Choueka (1988), uma colocação é "uma unidade sintática e semântica cujo significado exato e não ambíguo não pode ser derivado diretamente do significado ou da conotação das suas componentes". Para Fillmore et al. (1988, p. 504), "uma expressão idiomática ou construção é algo que um usuário da língua não poderia saber mesmo que soubesse todo o restante daquela língua". Sag et al. (2002) generalizam essa mesma propriedade e definem EPLs de maneira vaga, como "interpretações idiossincráticas que atravessam as fronteiras (ou espaços) entre as palavras".

Todas essas definições são perfeitamente válidas em um determinado contexto experimental. No entanto, a definição de EPL adotada dentre as diversas possíveis afetará a avaliação dos resultados, pois a definição será usada para escrever as instruções aos anotadores humanos e/ou para selecionar listas de referência. Assim, nessa tese, adota-se a definição proposta por Calzolari et al. (2002). Para esse trabalho, EPLs são "[...] fenômenos diferentes, mas relacionados [...]. No nível mais alto de generalidade, cada

um desses fenômenos pode ser descrito como uma [combinação] de palavras que agem como uma unidade única em algum nível de análise linguística." Essa definição genérica e intencionalmente vaga pode ser delimitada de acordo com os requisitos da aplicação. Por exemplo, para um sistema de tradução automática, uma EPL é qualquer combinação de palavras que, caso não seja traduzida como uma unidade, irá gerar uma saída pouco natural ou errada. O nível no qual uma combinação de palavras deve ser tratada como uma unidade varia de acordo com o tipo de expressão e de sistema.

Na bibliografia, algumas propriedades gerais das EPLs são descritas: arbitrariedade, institucionalização, variabilidade semântica limitada (não composicionalidade, não substituibilidade, tradução não literal, especificidade em um domínio), variabilidade sintática limitada (extragramaticalidade, lexicalização), e heterogeneidade. Essas propriedades não são interruptores binários do tipo sim/não, mas seus valores variam em um *continuum* que vai desde as combinações de palavras completamente flexíveis e ordinárias até expressões totalmente fixas e/ou prototípicas.

Existem diversas tipologias para a classificação das EPLs, baseadas em pontos de vista das diferentes teorias gramaticais: construcionismo, teoria sentido-texto, engenharia de gramática e aquisição automática de EPLs. Neste trabalho, propõe-se uma tipologia baseada, em primeiro lugar, no papel morfossintático da expressão como um todo na frase e, em segundo lugar, na dificuldade para se lidar com a expressão usando métodos computacionais. A primeira tipologia classifica as EPLs como expressões nominais, expressões verbais e expressões adjetivais/adverbiais. Expressões nominais incluem substantivos compostos (*roleta russa*), nomes próprios (*Porto Alegre*) e termos polilexicais (*domínio de ligação ao DNA*). Expressões verbais incluem verbos frasais (*ir embora*) e construções com verbos leves (*tomar banho*). Expressões adverbiais e adjetivais incluem exemplos como *sem mais nem menos* em português, *upside down* em inglês e *à poil* em francês. Além desses três tipos, são definidos três tipos ortogonais que estão relacionados com os métodos computacionais usados para processar as EPLs: (i) expressões fixas como *no entanto*, (ii) expressões idiomáticas como *unha e carne*, *deixar a desejar* e *sem pé nem cabeça*, e (iii) colocações "verdadeiras", que correspondem a expressões completamente composicionais que coocorrem com mais frequência do que seria esperado por mero acaso. Essas tipologias são bastante simples, porém suficientemente rigorosas para descrever as EPLs tratadas pelos experimentos descritos a seguir.

B.3 Estado da arte em processamento de EPLs

Antes de aprofundar-se na vasta bibliografia existente em processamento de EPLs, é necessário revisar algumas noções elementares. Um *corpus* é simplesmente um corpo de textos usado em estudos empíricos da língua (Manning and Schütze 1999, p. 6). *Análise linguística* é o processo de criação de estruturas de representação mais abstratas a partir de textos brutos em corpora. A análise pode ser vista como uma cadeia de etapas que transformam representações mais concretas (o texto) em representações mais abstratas (árvores e grafos). Algumas dessas etapas são: separação de frases, tokenização, lematização, etiquetamento morfossintático e análise sintática de dependências.

Em aquisição automática de EPLs, assume-se a hipótese de que as palavras que compõem uma expressão coocorrerão com mais frequência do que se elas fossem combinadas aleatoriamente. Essa hipótese é aplicada na concepção de medidas de associação (MAs) para aquisição de EPLs a partir de corpora. Existem diversas MAs disponíveis para aquisição automática de EPLs (Evert 2004, Seretan 2008, Pecina 2008b). Para um *n*-grama

arbitrário w_1^n, estima-se a sua probabilidade sob máxima verossimilhança como $p(w_1^n) = \frac{c(w_1)\times c(w_2)\times\ldots\times c(w_n)}{N^n}$. Quando se pondera esse estimador pelo número total de n-gramas do corpus N, obtém-se o número de ocorrências esperado $E(w_1^n) = \frac{c(w_1)\times c(w_2)\times\ldots\times c(w_n)}{N^{n-1}}$. MAs são geralmente baseadas na diferença entre o número de ocorrências esperado $E(w_1^n)$ e o número de ocorrências observado $c(w_1^n)$, por exemplo:

$$\texttt{t-score} = \frac{c(w_1^n)-E(w_1^n)}{\sqrt{c(w_1^n)}}, \quad \texttt{pmi} = \log_2 \frac{c(w_1^n)}{E(w_1^n)}, \quad \texttt{dice} = \frac{n \times c(w_1^n)}{\sum_{i=1}^n c(w_i)}.$$

Para o caso especial de 2-gramas, existem também MAs mais robustas e mais teoricamente bem fundadas, baseadas em tabelas de contingência. Exemplos de tais medidas são mostrados abaixo, onde $w_i \in \{w_1, \neg w_1\}$ e $w_j \in \{w_2, \neg w_2\}$:

$$\chi^2 = \sum_{w_i,w_j} \frac{[c(w_iw_j)-E(w_iw_j)]^2}{E(w_iw_j)}, \quad \texttt{ll} = 2 \times \sum_{w_i,w_j} c(w_iw_j) \times \log \frac{c(w_iw_j)}{E(w_iw_j)}.$$

B.3.1 Aquisição de EPLs

O termo *aquisição de EPLs* inclui a sua *identificação* (em contexto) e sua *extração* (fora de contexto). A aquisição monolíngue de EPLs é usualmente vista como um processo em duas etapas:

1. **Extração de candidatos**. Um dos métodos mais populares para a extração de candidatos é a utilização de sequências de etiquetas morfossintáticas, especialmente em terminologia (Justeson and Katz 1995, Daille 2003), mas também em substantivos compostos (Vincze et al. 2011) e expressões verbais (Baldwin 2005b). Quando um analisador sintático existe para a língua alvo, padrões sintáticos podem ser muito mais precisos do que sequências de etiquetas morfossintáticas, especialmente na extração de EPLs flexíveis (Seretan and Wehrli 2009, Seretan 2008). Gramáticas de substituição de árvores (Green et al. 2011) e regularidades estruturais das árvores sintáticas (Martens and Vandeghinste 2010) também podem ser usadas a fim de aprender modelos sintáticos de EPLs a partir de corpora anotados. O algoritmo LocalMaxs realiza extração de candidatos usando o princípio de maximização local de uma MA aplicada a pares de palavras adjacentes (Silva and Lopes 1999). Outra proposta de extração consiste em aplicar um algoritmo de correspondência de cadeias de caracteres inspirado na biologia computacional, com o objetivo de encontrar sequências não contínuas de palavras que ocorrem de maneira recorrente no corpus (Duan et al. 2006).

2. **Filtragem de candidatos**. Para filtrar os candidatos, alguns métodos facilmente aplicáveis são as listas de palavras proibidas e os limiares de ocorrências. MAs também são amplamente empregadas para ordenar os candidatos, mantendo apenas aqueles cuja medida de associação se encontra acima de um determinado limiar (Evert and Krenn 2005, Pecina 2005). Métodos de aprendizado supervisionado podem ser usados para construir classificadores que otimizam os pesos dados a diferentes MAs e outros atributos dos candidatos (Ramisch et al. 2008b, Pecina 2008b).

Existem algumas ferramentas disponíveis gratuitamente e que podem ser usadas para aquisição monolíngue de EPLs, como LocalMaxs, [1] Text::NSP, [2] UCS, [3] jMWE, [4] e Varro. [5] Além disso, estão também disponíveis alguns serviços web gratuitos, assim como ferramentas gratuitas e comerciais para extração automática de terminologia a partir de textos especializados.

Quanto à aquisição bilíngue, alinhamentos lexicais automáticos podem ser diretamente utilizados como listas de candidatos a EPL (de Medeiros Caseli et al. 2010). Bai et al. (2009) descrevem um algoritmo capaz de minerar traduções para uma dada EPL em corpora paralelos alinhados. A descoberta automática de compostos não composicionais a partir de dados paralelos foi explorada por Melamed (1997). O par de línguas inglês-hindi apresenta uma grande variação de ordem das palavras, e Venkatapathy and Joshi (2006) demonstraram que atributos de composicionalidade baseados em EPLs podem ajudar a reduzir a taxa de erro de alinhamento.

Zarrieß and Kuhn (2009) usaram corpora analisados sintaticamente e alinhamentos lexicais gerados por GIZA++ para extrair pares do tipo verbo-objeto de um corpus paralelo em inglês-alemão. Daille et al. (2004) extraíram termos polilexicais de corpora comparáveis em inglês e em francês, e em seguida usaram as distâncias entre os vetores de contexto desses termos para obter correspondências entre as línguas.

B.3.2 Outras tarefas no processamento de EPLs

Existe um grande número de trabalhos publicados em que outras tarefas relacionadas ao tratamento de EPLs são abordadas. Alguns deles são discutidos abaixo.

– **Interpretação**: O problema de *interpretação sintática* de substantivos compostos é explorado por Nicholson and Baldwin (2006), que distinguem três tipos de relações sintáticas em substantivos compostos: sujeito, objeto direto e objeto indireto. Substantivos compostos de três ou mais palavras precisam de uma interpretação sintática da hierarquia dos seus componentes. Nakov and Hearst (2005) comparam dois modelos, baseados em adjacência e em dependência, usando um conjunto de heurísticas para gerar automaticamente paráfrases com as formas superficiais, e em seguida usando as contagens de um motor de busca para estimar as suas probabilidades. Nakov and Hearst (2008) mostram de que forma é possível realizar a *interpretação semântica* de substantivos compostos usando um grande conjunto de paráfrases que incluem verbos relacionados a suas respectivas classes semânticas, e então usando contagens da web para verificar sua validade. Kim and Nakov (2011) usam uma combinação de reamostragem por bootstrapping e contagens da web, usando paráfrases guiadas por árvores sintáticas e, assim, obtêm melhores resultados. Além de substantivos compostos, outros tipos de EPLs necessitam interpretação. Cook and Stevenson (2006) usam máquinas de vetores de suporte para classificar o significado da partícula *up* em inglês em verbos frasais. Bannard (2005) quantifica a contribuição em termos de significado de cada componente dos verbos frasais para a interpretação do todo. Um trabalho similar é apresentado por McCarthy et al. (2003), que propõem diversas medidas envolvendo um tesauro construído automaticamente a fim de estimar a idiomaticidade de verbos frasais.

1. http://hlt.di.fct.unl.pt/luis/multiwords/
2. http://search.cpan.org/dist/Text-NSP
3. http://www.collocations.de/software.html
4. projects.csail.mit.edu/jmwe
5. http://sourceforge.net/projects/varro/

– **Desambiguação**: A desambiguação de EPLs é análoga à interpretação, porém na desambiguação as EPLs são consideradas como parte de um contexto de ocorrência. Nicholson and Baldwin (2008) apresentam um conjunto de dados no qual um grande número de frases foi anotado manualmente com relação à classificação de compostos do tipo substantivo-substantivo. Girju et al. (2005) estudam métodos para a desambiguação desses compostos usando diversas técnicas de aprendizado supervisionado. Fritzinger et al. (2010) analisam manualmente diversas construções ambíguas do tipo preposição-substantivo-verbo em alemão. As construções são identificadas através de análise sintática e classificadas como literais, composicionais ou desconhecido. Verbos leves em japonês são estudados por Uchiyama et al. (2005), que propõem dois métodos de desambiguação: um método estatístico e outro baseado em regras. Cook et al. (2007) investigam a idiomaticidade de pares do tipo verbo-substantivo nos quais o substantivo é objeto direto do verbo e pode ser interpretado de forma idiomática (*fazer onda*) ou literal (*fazer um bolo*). Fazly and Stevenson (2007) propõem uma classificação mais fina para construções envolvendo verbos leves e substantivos, empregando aprendizado supervisionado para realizar uma desambiguação semântica com quatro classes.
– **Representação**: A representação de unidades léxicas polilexicais é um problema que tem dado muita dor de cabeça aos lexicógrafos durante a compilação de recursos lexicais. Sag et al. (2002) propõem duas abordagens: palavras com espaços e composicional. No entanto, entre esses dois extremos do espectro de composicionalidade, existem outras possibilidades sugeridas na literatura. Laporte and Voyatzi (2008) descrevem um dicionário de expressões adverbiais em francês e seus respectivos padrões morfossintáticos no formato léxico-gramática. Graliński et al. (2010) apresentam uma comparação quantitativa e qualitativa entre duas representações estruturadas para EPLs em polonês, Multiflex e POLENG. Grégoire (2007; 2010) usa um método baseado em classes de equivalência que agrupam expressões similares de acordo com suas características sintáticas. Izumi et al. (2010) sugerem um método baseado em regras para normalizar e consequentemente otimizar a representação de expressões funcionais em japonês. Schuler and Joshi (2011) propões o uso de gramáticas de reescrita de árvores para representar EPLs.
– **Aplicações**:
  Existem algumas aplicações de PLN para as quais foram obtidos resultados concretos em termos de integração de EPLs. Por exemplo, na análise sintática, Constant and Sigogne (2011) apresentam resultados promissores para a etiquetagem morfossintática do francês. Korkontzelos and Manandhar (2010) obtêm melhorias impressionantes na qualidade de um analisador raso comum por meio da inserção de EPLs. Zhang and Kordoni (2006) e Villavicencio et al. (2007) obtêm uma melhoria significativa de cobertura pela extensão do léxico de um analisador HSPG do inglês com entradas polilexicais.
  Wehrli et al. (2010) demonstram que EPLs não são "carne de pescoço", conforme descrito no célebre artigo de Sag et al. (2002), mas na verdade são uma fonte de informação valiosa para reduzir ambiguidade sintática. Outro exemplo de aplicação em que EPLs foram integradas com sucesso é recuperação de informações. Acosta et al. (2011) unem as palavras que compõem uma EPL antes de indexar os documentos, e isso aumenta a precisão média do sistema. Xu et al. (2010) propõem uma nova medida de coesão para sequências de quatro caracteres em chinês, obtendo igualmente um aumento na precisão média do sistema sobre um conjunto de testes.

B.4 Avaliação da aquisição de EPLs

A avaliação da aquisição de EPLs é um problema bastante complexo, porque os resultados dependem de muitos parâmetros do contexto de aquisição, tornando os resultados obtidos em um dado contexto difíceis de generalizar. Na bibliografia, encontram-se diversos estilos de avaliação: analisar listas ordenadas das top-*k* EPLs retornadas (da Silva et al. 1999), anotar manualmente esses top-*k* candidatos (Seretan 2008), medir a precisão e a revocação com relação a um dicionário (Ramisch 2009), comparar a qualidade das medidas de associação por meio da sua precisão média (Evert and Krenn 2005), comparar diferentes métodos (Pearce 2002, Ramisch et al. 2008a), e medir o impacto das EPLs adquiridas em aplicações reais de PLN (Finlayson and Kulkarni 2011, Xu et al. 2010, Carpuat and Diab 2010). No presente trabalho, propõe-se a seguinte tipologia para classificar o *contexto de avaliação*, de maneira a proporcionar uma visão mais estruturada da avaliação.

1. **De acordo com o objetivo da aquisição**
   - **Intrínseca**. Os resultados são apresentados por meio da avaliação das próprias EPLs adquiridas, diretamente, como um produto final de um processo. A avaliação intrínseca é fortemente baseada na aplicação alvo e na coerência das instruções fornecidas aos anotadores, mas ainda assim fornece uma estimativa útil da qualidade das EPLs adquiridas.
   - **Extrínseca**. A avaliação extrínseca pode ser realizada integrando-se EPLs em aplicações de PLN, e após verificando-se se elas melhoram a qualidade da saída produzida. Eventualmente, pode ser mais simples avaliar uma aplicação de PLN do que as listas de EPLs isoladas. A avaliação extrínseca pode ser bastante conclusiva para demonstrar a utilidade das EPLs adquiridas.
2. **De acordo com a natureza das medidas**
   - **Quantitativa**. Esse estilo de avaliação usa medidas objetivas como precisão, revocação, F-medida e precisão média. Apesar de diversos artigos apenas apresentarem a precisão para os primeiros *k* candidatos, é igualmente importante avaliar a revocação, pois o número de expressões (novas) descobertas é tão importante quanto a sua qualidade.
   - **Qualitativa**. O objetivo desse estilo de avaliação é entender os erros realizados pelo método de aquisição. Para isso, são observados os resultados em termos de sequências de etiquetas morfossintáticas, de distribuição de frequências, de contexto, etc. A análise qualitativa é complementar à quantitativa, e ambas podem ser realizadas simultaneamente e/ou iterativamente.
3. **De acordo com os recursos disponíveis**
   - **Anotação manual**. Um grupo de falantes nativos e/ou especialistas percorre a lista de EPLs candidatas, de maneira a decidir se uma dada combinação é uma EPL ou não. A anotação pode requerer bastante tempo, dependendo da disponibilidade de anotadores, e por isso é frequentemente efetuada sobre uma pequena amostra da saída do sistema.
   - **Anotação automática**. Na avaliação automática, considera-se que existe um dicionário completo ou de ampla cobertura contendo as EPLs alvo. Assim, considera-se que as candidatas que ocorrem no dicionário são verdadeiras positivas (EPLs interessantes/genuínas), enquanto as demais são falsas EPLs.
4. **De acordo com a classe de EPL**

– **Avaliação baseada em tipos**. Expressões não ambíguas como substantivos compostos, terminologia e construções de verbo-suporte podem ser anotadas independentemente do seu contexto de ocorrência. Existem diversos léxicos disponíveis que podem servir como referência para a avaliação baseada em tipos. Quando tais léxicos não existem para as expressões alvo, a avaliação deve ser manual.
– **Avaliação baseada em instâncias**. A avaliação baseada em instâncias (ou *tokens*) é requerida quando as EPLs alvo são ambíguas e podem ter várias interpretações de acordo com o contexto, como verbos frasais e expressões idiomáticas. Fora de contexto, é impossível decidir se as palavras devem ser processadas separadamente ou como uma unidade. Na avaliação baseada em instâncias, anotadores avaliam frases inteiras ao invés de EPLs candidatas isoladas.

Se modelarmos o resultado da aquisição de EPLs como uma lista C de candidatas ordenadas de acordo com um valor numérico, a precisão P(C) do sistema corresponde à proporção de candidatas avaliadas como EPLs verdadeiras dentre o conjunto de candidatas retornadas, $\mathrm{P}(C) = \frac{|\text{EPLs em } C|}{|C|}$. A precisão é uma estimativa da quantidade de trabalho necessária para transformar uma lista bruta de candidatas adquiridas automaticamente em uma lista de EPLs finalizada e validada por um especialista. No entanto, essa medida ignora as EPLs verdadeiras que não foram retornadas pelo sistema quando elas deveriam ter sido encontradas. Portanto, é crucial calcular a revocação $\mathrm{R}(C) = \frac{|\text{EPLs em } C|}{|\text{Total de EPLs a adquirir}|}$. Apesar da sua importância, R(C) é raramente calculada porque é difícil estimar o número total de EPLs que o sistema deveria adquirir.

Existem dois tipos de anotação: automática e manual. Na anotação automática, existe um *padrão de referência*, ou seja, um léxico contendo a lista completa de EPLs que devem ser retornadas pelo sistema ideal. Na anotação automática, P(C) e R(C) são subestimadas porque elas assumem que aquelas candidatas que não aparecem no padrão de referência são falsas EPLs. Apesar dessa simplificação, a anotação automática é utilizada com frequência, principalmente porque ela é rápida e pouco custosa quando existe um padrão de referência disponível de forma gratuita. A anotação manual é raramente realizada sobre a lista completa de EPLs candidatas resultantes de um sistema, mas em uma amostra. Se a lista está ordenada, as primeiras k candidatas podem ser anotadas, porém isso pode tornar a avaliação tendenciosa, avaliando apenas candidatas altamente frequentes, enquanto uma amostra equilibrada deveria incluir candidatas de todas as faixas de frequência. É importante conceber e testar cuidadosamente as instruções para o grupo de anotadores, que pode ser constituído por falantes nativos ou, caso as EPLs alvo sejam demasiado complexas, linguistas especializados. É recomendado permitir certa flexibilidade nas categorias de anotação, com múltiplas classes ao invés de uma escolha binária. A medida kappa de Fleiss é frequentemente usada para estimar a concordância entre os diversos anotadores, apesar de a sua interpretação ser controversa. Os estilos de anotação manual e automática são complementares, e é possível combiná-los, por exemplo, anotando manualmente entradas ausentes do padrão de referência.

O *contexto de aquisição* é definido como o conjunto de parâmetros que podem influenciar os resultados da avaliação e, consequentemente, limitam a generalização destes. A hipótese adotada nesse trabalho é de que uma avaliação realizada em um dado contexto de aquisição é dificilmente generalizável, porque ela depende de um número muito grande de parâmetros.

Alguns parâmetros do contexto de aquisição são características das próprias EPLs, como:

– **Tipo**. Diferentes tipos de EPLs requerem diferentes avaliações. Por exemplo, sequências de etiquetas morfossintáticas são frequentemente usadas com sucesso na aquisição de substantivos compostos, porém têm um desempenho muito ruim para expressões verbais (Villavicencio et al. 2012).
– **Língua**. Não somente as próprias EPLs, mas também os recursos de PLN são diferentes para cada língua. Por exemplo, o uso de um analisador sintático para a aquisição de colocações, como em Seretan (2008), é impossível para línguas pobremente instrumentadas, que requerem alternativas usando ferramentas de análise rasa.
– **Domínio**. O domínio da expressão deve ser levado em consideração. Por exemplo, a lista de padrões de etiquetas morfossintáticas sugeridas por Justeson and Katz (1995) para a aquisição de termos polilexicais genéricos obteve um baixo desempenho quando aplicada a um corpus do domínio biomédico (Ramisch 2009).

Alguns parâmetros do contexto de aquisição são características dos corpora, como:

– **Tamanho**. Corpora grandes contêm mais dados, e intuitivamente um método de aquisição automática poderá encontrar mais candidatas, obtendo assim uma melhor revocação. Métodos estatísticos podem ser sensíveis a dados esparsos, de forma que amostras maiores implicam em melhores estimativas e, por conseguinte, mais precisão.
– **Natureza**. Os resultados da aquisição dependem do domínio e do gênero dos textos. Experimentos mostram que, por exemplo, na extração de substantivos compostos especializados, o uso de contagens advindas da web como um corpus não é recomendado (Ramisch et al. 2010d).
– **Nível de análise**. Os métodos de aquisição de EPL atuais usam desde informações puramente superficiais e rasas (da Silva et al. 1999) até informações baseadas na análise profunda em um determinado formalismo sintático (Seretan 2008). No entanto, uma análise mais profunda não necessariamente gera melhores resultados (Baldwin 2005b).

A avaliação da aquisição de EPLs é ainda um problema em aberto. Se, por um lado, medidas como precisão e revocação aliadas a uma anotação automática dependem da disponibilidade de um padrão de referência completo, por outro lado a avaliação manual é trabalhosa e tende a dar mais importância à precisão do que ao número de EPLs novas descobertas. Na literatura, encontram-se artigos que descrevem avaliações comparativas (Schone and Jurafsky 2001, Pecina 2005, Ramisch et al. 2008a) e avaliações extrínsecas baseadas em aplicações como recuperação de informações (Doucet and Ahonen-Myka 2004, Xu et al. 2010, Acosta et al. 2011), desambiguação lexical (Finlayson and Kulkarni 2011), tradução automática (Carpuat and Diab 2010, Pal et al. 2010) e aprendizado de ontologias (Venkatsubramanyan and Perez-Carballo 2004).

B.5 Uma plataforma para a aquisição de EPLs

Uma das contribuições desta tese é a introdução de uma nova plataforma chamada `mwetoolkit`, que integra múltiplas técnicas e cobre todo o pipeline de aquisição de EPLs. É possível pré-processar corpora monolíngues brutos, caso haja ferramentas disponíveis para a língua alvo, enriquecendo-os com etiquetas morfossintáticas, lemas e dependências sintáticas. Então, com base em conhecimento de um especialista, em intuição,

em observação empírica e/ou em exemplos, o usuário define padrões multiníveis usando um formalismo similar às expressões regulares para descrever as EPLs alvo. A aplicação desses padrões sobre um corpus indexado gera uma lista de EPLs candidatas. Para filtrá-las, a plataforma disponibiliza uma miríade de métodos que vão desde simples limiares de frequência até listas de palavras proibidas, passando por medidas de associação sofisticadas. Finalmente, as candidatas filtradas resultantes podem ser diretamente injetadas em uma aplicação de PLN ou então validadas manualmente por um humano antes de serem integradas a uma aplicação. Um uso alternativo para as candidatas validadas é treinar um modelo de aprendizado de máquina, que pode ser aplicado em outros corpora para identificar e extrair automaticamente novas EPLs com base nas características das EPLs adquiridas anteriormente. O processo de aquisição é resumido na Figura 5.1. Mais detalhes sobre o funcionamento da plataforma podem ser encontrados no site da ferramenta ou em publicações anteriores (Ramisch et al. 2010b;c).

Atualmente, não há um consenso sobre a existência de um método único e ótimo para a aquisição de EPLs, ou sobre um subconjunto de métodos mais indicados para adquirir determinados tipos de EPLs. Uma das contribuições principais da metodologia proposta neste trabalho é uma integração sistemática dos processos e tarefas necessários à aquisição. Uma das grandes vantagens dessa plataforma é que ela modela todo o processo de aquisição de maneira modular, sendo assim customizável e permitindo a regulagem fina de um grande número de parâmetros. O `mwetoolkit` pode ser usado para acelerar o trabalho de lexicógrafos e terminógrafos e para contribuir com a adaptação de ferramentas de PLN a outras línguas e domínios. A metodologia empregada no toolkit não depende de conhecimento simbólico ou de dicionários preexistentes, e as técnicas implementadas são independentes de linguagem. Além disso, elas são independentes do comprimento dos *n*-gramas das candidatas e da adjacência entre as suas palavras. Graças a sua generalidade, essa metodologia pode ser aplicada praticamente a qualquer língua, tipo de EPL e domínio, sem estar condicionada ao uso de um dado formalismo ou ferramenta de análise. Em resumo, a metodologia do `mwetoolkit` permite que os usuários realizem uma aquisição de EPLs sistemática com arquivos intermediários consistentes e módulos com uma funcionalidade e um conjunto de parâmetros bem definidos.

A metodologia do `mwetoolkit` é inicialmente avaliada por meio de uma comparação com outras três ferramentas disponíveis gratuitamente, livres para download e abertamente documentadas: a implementação de referência do algoritmo LocalMaxs (`LocMax`), o pacote de estatísticas de *n*-gramas (`NSP`), e a ferramenta `UCS`. Os experimentos foram realizados em duas línguas, inglês (`en`) e francês (`fr`), analisando-se expressões verbais e nominais em inglês e apenas expressões nominais em francês. As EPLs adquiridas foram automaticamente avaliadas usando-se padrões de referência existentes.

A qualidade das candidatas extraídas do corpus de tamanho médio (M) varia de acordo com o tipo e a língua das EPLs, conforme mostrado na Figura 5.5. Para as EPLs nominais, os métodos possuem padrões de desempenho bastante similares, com alta revocação e baixa precisão. Para expressões verbais, `LocMax` obteve alta precisão (em torno de 70%), mas baixa revocação, enquanto os demais métodos possuem valores de P e R mais equilibrados, em torno de 20%. As técnicas se distinguem em termos de estratégia de extração: (i) `mwetoolkit` e `NSP` permitem a definição de filtros linguísticos, enquanto `LocMax` permite apenas a aplicação de filtros usando ferramentas como *grep* após a extração; (ii) não há nenhum tipo de filtragem preliminar no `mwetoolkit` e no `NSP`, eles simplesmente retornam todas as candidatas que correspondem a um padrão, porém `LocMax` usa o critério de máximos locais para filtrar as candidatas a priori; (iii)

LocMax apenas extrai unidades contíguas, mas as outras ferramentas permitem a extração de unidades descontínuas. A avaliação das candidatas nominais em en de acordo com o tamanho do corpus é mostrada na Tabela 5.4. Em todas as abordagens, a precisão decresce quando o corpus aumenta, enquanto a revocação aumenta em todas as abordagens exceto LocMax.

A Tabela 5.6 apresenta a avaliação das medidas de associação. A medida glue do LocMax tem o melhor desempenho para todos os tipos de EPLs, sugerindo que o critério de máximos locais é um bom indicador de EPLs genéricas, enquanto essa medida de associação é uma maneira eficiente de se obter resultados de alta precisão. Para o mwetoolkit, a melhor medida de associação é o coeficiente dice; as outras medidas não são consistentemente melhores do que a linha de base. A medida de Poisson-Stirling (Poisson) obteve um desempenho bastante bom, enquanto as outras medidas testadas para o NSP obtiveram um desempenho inferior ao da linha de base para alguns casos. Finalmente, todas as medidas testadas para o UCS obtiveram um desempenho significativamente superior ao da linha de base e, para as EPLs nominais, o desempenho é comparável ao da melhor medida de associação glue.

Aspectos como o grau de flexibilidade da EPL e o desempenho computacional do método podem influenciar a decisão da medida de associação adotada. Por exemplo, dice pode ser facilmente aplicado para qualquer tamanho de *n*-grama, enquanto medidas mais sofisticadas como Poisson são definidas apenas para 2-gramas e podem ser pesadas em termos de recursos computacionais. UCS não extrai candidatas de corpora, mas recebe como entrada uma lista de 2-gramas. NSP implementa algumas medidas de associação para 3 e 4-gramas e mwetoolkit e LocMax não possuem limitações quanto ao número de palavras das candidatas. LocMax extrai apenas EPLs contíguas, enquanto mwetoolkit e NSP permitem a extração de sequências de palavras não adjacentes. Somente o mwetoolkit integra filtros linguísticos baseados em lemas, etiquetas morfossintáticas e sintaxe. Para os outros métodos, isso deve ser realizado usando ferramentas externas como sed e grep.

O mwetoolkit é um primeiro passo importante no tratamento robusto e confiável de EPLs pelas aplicações de PLN. É uma plataforma disponível gratuitamente que fornece ferramentas poderosas e uma documentação coerente e frequentemente atualizada. Essas são características essenciais para a sua extensão e suporte ao uso, como para qualquer sistema computacional.

B.6 Aplicação 1: lexicografia

Uma primeira avaliação quantitativa e qualitativa da plataforma de aquisição de EPLs proposta foi realizada no contexto da lexicografia assistida por computador. Esse trabalho foi realizado em colaboração com colegas linguistas e lexicógrafos experientes, com o objetivo de criar recursos lexicais contendo EPLs em grego e em português. Os recursos resultantes dessa avaliação estão disponíveis gratuitamente. [6]

Para o grego, existe uma quantidade considerável de trabalhos publicados que estudam as propriedades linguísticas das EPLs, porém abordagens computacionais ainda são raras (Fotopoulou et al. 2008). Nos experimentos, usou-se o mwetoolkit para extrair uma lista inicial de EPLs candidatas da porção grega do corpus Europarl. Os padrões de extração usados foram os seguintes: adjetivo-substantivo, substantivo-substantivo, substantivo-

6. http://multiword.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets

artigo-substantivo, substantivo-preposição-substantivo, preposição-substantivo-substantivo, substantivo-adjetivo-substantivo e substantivo-conjunção-substantivo. Para filtrar as candidatas, aplicou-se um conjunto de medidas de associação estatísticas usando contagens coletadas no corpus original e na web. As primeiras 150 candidatas ordenadas de acordo com quatro medidas de associação nos dois corpora foram manualmente avaliadas por três falantes nativos. Cada anotador julgou aproximadamente 1.200 candidatas e ao final elas foram unidas, criando-se um léxico com 815 EPLs nominais em grego.

Com base nessas anotações, analisou-se a contribuição exata de cada uma das medidas de associação em termos de EPLs corretas encontradas. A medida de associação que apresentou os melhores resultados foi `dice`, que obteve um desempenho significativamente melhor que as outras medidas. O desempenho da medida `t-score` é o segundo melhor, porém surpreendentemente é também muito similar ao desempenho das contagens brutas dos *n*-gramas, sugerindo que medidas sofisticadas não são necessárias quando uma quantidade considerável de dados está disponível. O uso da web como um corpus apresenta diversas vantagens com relação a corpora tradicionais, dentre as quais se salienta sua acessibilidade e disponibilidade. No entanto, nos experimentos aqui discutidos, os resultados obtidos com as contagens da web não trouxeram nenhuma melhoria considerável. Em suma, os resultados obtidos indicam que métodos automáticos podem efetivamente ser usados para estender recursos de PLN com informações de EPLs, melhorando a qualidade dos sistemas de PLN da língua grega.

O objetivo do trabalho com predicados complexos (PCs) do português é realizar uma análise qualitativa dessas construções. Foram gerados dois recursos lexicais correspondendo a duas aplicações alvo: CP-SRL é concebido para a anotação de etiquetas de papel semântico, enquanto CP-SENT é concebido para análise de sentimentos. Para construir ambos os recursos, o corpus PLN-BR-FULL foi etiquetado morfossintaticamente a fim de serem extraídas sequências de palavras que correspondem a padrões específicos de etiquetas morfossintáticas usando o `mwetoolkit`.

A anotação de etiquetas de papéis semânticos depende da identificação correta dos predicados, antes de identificar os argumentos e atribuir as etiquetas de papel semântico. No entanto, vários predicados não são formados apenas por um verbo: eles são PCs que não aparecem nos léxicos computacionais. Para criar o dicionário CP-SRL, em vez de usar um conjunto fechado de verbos ou substantivos, foram usados os seguintes padrões morfossintáticos: verbo-[artigo]-substantivo-preposição, verbo-preposição-substantivo, verbo-[preposição/artigo]-advérbio e verbo-adjetivo. O processo de extração resultou em uma lista com 407,014 candidatas que foram filtradas usando-se medidas de associação. Um especialista humano anotou e validou manualmente 12,545 candidatas, das quais 699 foram anotadas como expressões verbais composicionais e 74 como expressões verbais idiomáticas. Os resultados incluem (mas não se limitam a) construções de verbos leves e de verbos-suporte. Os seguintes pares de paráfrases foram observados recorrentemente:

– V = V + N DEVERBAL: *tratar* = *dar tratamento*;
– V DENOMINAL = V + N: *amedrontar* = *dar medo*;
– V DEADJETIVAL = V + ADJ: *responsabilizar* = *tornar responsável*.

Para a criação de CP-SENT, o objetivo é investigar de que forma os sentimentos são exprimidos em português do Brasil. Verbos de sentimento como *temer, odiar* e *invejar* são exemplos de unidades lexicais usadas especificamente para expressar os sentimentos correspondentes. O mesmo sentido pode ser obtido pela associação de outros verbos com substantivos de sentimento. Esse estudo primeiramente identifica sete padrões recorrentes

que expressam sentimentos sem usar verbos de sentimento, e então aplica esses padrões para identificar substantivos de sentimento associados a eles. Isso foi realizado em cinco etapas. Primeiro, identificou-se padrões léxico-sintáticos recorrentes que expressam sentimentos usando substantivos de sentimento em vez de verbos. Segundo, foram usados os padrões identificados como argumentos de busca para identificar a expressão de sentimentos. Terceiro, um humano analisou a lista de candidatas resultante da etapa anterior, determinando se o substantivo colocado à direita de cada padrão exprime um sentimento ou não. Quarto, as candidatas foram analisadas, validadas e enriquecidas com um conjunto de atributos. Quinto, foram combinados os padrões da primeira etapa com os substantivos de sentimento identificados na etapa três e procurou-se a nova combinação na web. A análise desses padrões mostrou que a combinação de substantivos de sentimento com os sete padrões identificados pode ser útil para identificar automaticamente a expressão de sentimentos e adicionalmente descobrir quem está sentindo algo e o que ou quem está provocando o sentimento.

B.7 Aplicação 2: tradução automática empírica

Como uma segunda avaliação do `mwetoolkit`, foram realizados experimentos com um sistema empírico de tradução automática (TA), estudando a tradução para o português de verbos frasais em inglês como *give up* (*desistir*) e *get by [a name]* (*ser chamado de [um nome]*). A tradução dos verbos frasais é um desafio porque eles apresentam uma variabilidade sintática e semântica bastante ampla. Verbos frasais são muito comuns e aparecem frequentemente em inglês, ocorrendo em cerca de 17% das frases do corpus usado nos experimentos. Modelar o comportamento sintático e semântico complexo dos verbos frasais usando as sequências de palavras planas e contíguas dos sistemas atuais de TA empírica não é intuitivo. Apesar disso, é importante identificar os verbos frasais e processá-los de maneira correta para evitar traduções que pareçam pouco naturais ou agramaticais.

A representação e integração de EPLs em sistemas de TA têm sido estudadas em diversos projetos. O sistema de TA ITS-2 processa EPLs em dois níveis: durante a análise lexical para compostos contíguos, e durante a análise sintática para colocações (Wehrli 1998, Wehrli et al. 2010). Carpuat and Diab (2010) adotam duas estratégias complementares para integrar EPLs: uma estratégia estática de tokenização única, na qual as EPLs são tratadas como palavras com espaços; e uma estratégia dinâmica que adiciona as contagens do número de EPLs presentes no segmento fonte como um atributo do modelo. Morin and Daille (2010) obtêm uma melhoria de 33% na qualidade da tradução francês-japonês usando um método composicional baseado em morfologia como back-off quando não há dados suficientes no dicionário para traduzir uma EPL. Em linguagens morfologicamente ricas como alemão, em que um substantivo composto é na verdade uma unidade única formada por concatenação, Stymne (2011) divide o composto nas palavras simples que o compõem antes de traduzir, e então aplica pós-processamento como reordenação e união dos componentes após a tradução. Outra abordagem para lidar com dados esparsos é adotada por Nakov (2008a), que gera paráfrases monolíngues para aumentar o corpus de treinamento.

Nos experimentos realizados, treinou-se o sistema Moses sobre o corpus Europarl v6 inglês-português, gerando-se um sistema standard não fatorado de TA empírica baseado em segmentos. Os verbos frasais foram automaticamente identificados usando-se a ferramenta jMWE e um dicionário de verbos frasais. Foram comparadas cinco estratégias para

a integração no sistema dos verbos frasais identificados automaticamente. O conjunto de teste é formado por uma amostra de 1.000 frases, das quais metade contém ao menos um verbo frasal. As construções mais frequentemente encontradas incluem *lay down, set up, carry out* e *originate in*.

Como o fenômeno linguístico estudado é complexo, não é possível tirar conclusões somente por meio de medidas automáticas como BLEU e NIST, requerendo-se uma análise cuidadosa dos erros via avaliação manual da saída do tradutor. Problemas comumente encontrados nos sistemas de TA testados incluem erros de conjugação verbal, seleção errônea de partículas/preposições, tradução do verbo como um substantivo e adição de preposições espúrias ao verbo alvo. Os resultados preliminares da avaliação por humanos realizada sobre uma amostra de 100 frases mostram que, ao mesmo tempo em que a qualidade de algumas das traduções melhora, a de outras piora. Não foi observada nenhuma melhoria absoluta, mas acredita-se que isso se deve ao fato de que a avaliação deve considerar classes mais finas de verbos frasais em vez de misturá-los em um mesmo conjunto de teste. Além disso, seria necessário anotar mais dados para obter resultados mais representativos.

Descobriu-se uma forte correlação entre a qualidade das traduções de cada uma das estratégias testadas e a composicionalidade dos verbos frasais. As estratégias que produziram os melhores resultados para casos idiomáticos foram TOK e BILEX. Para os casos composicionais, TOK resultou em uma queda de desempenho. Apesar da estratégia PV? ter a tendência de traduzir os verbos frasais como uma unidade, ela é menos rígida que TOK, produzindo menos traduções incorretas de verbos frasais composicionais. A comparação dessas heurísticas mostra que elas têm vantagens complementares, que são relacionadas à composicionalidade e à frequência. Essa hipótese motiva a continuidade dessa pesquisa a fim de se obter uma compreensão mais profunda do impacto de cada uma das estratégias de integração em cada uma das etapas do sistema de TA.

B.8 Conclusões e trabalhos futuros

Os objetivos deste trabalho foram descritos anteriormente como: (a) desenvolver técnicas para a aquisição automática de EPLs a partir de corpora, (b) avaliar essas técnicas extrinsecamente medindo a sua utilidade em aplicações de PLN, e (c) investigar a aquisição e integração de EPLs em contextos multilíngues. No estado atual, o objetivo (a) pode ser considerado como alcançado, e a ferramenta de software resultante, o `mwetoolkit`, está disponível gratuitamente. Como a avaliação da aquisição de EPLs é um problema em aberto, sugeriu-se uma descrição teórica que, espera-se, ajudará a definir uma forma estruturada de descrever esse problema. Quanto ao objetivo (b), ele também pode ser considerado como atingido, porque a utilidade da plataforma `mwetoolkit` foi demonstrada na construção de três recursos lexicais. Outras aplicações do `mwetoolkit` não foram incluídas na presente tese (Villavicencio et al. 2012, Granada et al. 2012). Finalmente, no que diz respeito ao objetivo (c), foram fornecidos resultados preliminares sobre a integração das EPLs em um sistema de TA empírico. Este trabalho está em andamento, permitindo um grande número de melhorias, extensões e experimentos que serão realizados como trabalhos futuros.

Pode-se identificar dois objetivos principais da pesquisa em andamento e da pesquisa futura.

– **Melhor integração de EPLs em TA**: a integração de verbos frasais em um sistema de TA empírica é um problema árduo em decorrência da variabilidade dessas cons-

truções. Os experimentos mostraram que, enquanto um sistema de TA standard consegue traduzir corretamente alguns dos verbos frasais, ele costuma errar quando a construção é idiomática ou não contígua. Pretende-se investigar alternativas para a inserção de entradas na tabela de segmentos, como estimar as probabilidades para as novas entradas polilexicais (Bouamor et al. 2011), usá-las para guiar o alinhamento lexical, adicioná-las ao corpus de treino, e pós-processar o resultado da tradução. Também se quer melhorar a detecção dos verbos frasais usando um analisador profundo capaz de detectar dependências de longa distância em expressões com alta variabilidade sintática (Seretan 2008, Baldwin 2005a). Potencialmente, a informação da sintaxe pode fornecer atributos adicionais para o modelo de tradução. A detecção de composicionalidade em verbos frasais com base no corpus (McCarthy et al. 2003, Bannard et al. 2003, Baldwin et al. 2003) poderia ajudar a guiar a tradução, gerando traduções mais precisas. Planeja-se também investigar outros fenômenos polilexicais que influenciam as equivalências e assimetrias entre as línguas. O objetivo em longo prazo é integrar o tratamento de EPLs em sistemas de TA empírica a fim de obter uma tradução de alta qualidade por meio da combinação de informações estatísticas e linguísticas.

– **Projeto CAMELEON**: Um dos resultados da presente tese é o projeto CAMELEON, com um financiamento CAPES-COFECUB 707-11. [7] O objetivo do projeto é estudar técnicas automáticas e colaborativas para o desenvolvimento de recursos lexicais e ontológicos para aplicações multilíngues. O objetivo é investigar a viabilidade de usar um sistema de gestão colaborativa de recursos lexicais para filtrar e validar as EPLs adquiridas automaticamente. Experimentos em andamento, cujos resultados preliminares foram recentemente publicados, exploram a aquisição automática de um corpus comparável representando uma amostra da linguagem típica do domínio da organização de conferências científicas em português, inglês e francês (Granada et al. 2012). Simultaneamente, estuda-se hoje a viabilidade de uma abordagem para a criação de recursos lexicais para o português baseada no jogo lexical sério JeuxDeMots [8] (Mangeot and Ramisch 2012). Esses experimentos representam um contexto experimental interessante para pesquisas futuras sobre a presença e importância das EPLs nas aplicações e recursos criados nessas três línguas.

Apesar de um grande número de trabalhos publicados na área, o tratamento das EPLs nas aplicações de PLN ainda é um problema em aberto e representa um grande desafio! Esse fato está longe de ser surpreendente, dado que a natureza complexa e heterogênea das EPLs tem sido demonstrada por diversos estudos linguísticos. No início dos anos 2000, Schone and Jurafsky (2001) perguntavam se a identificação de EPLs era um problema solucionado, ao que o artigo respondia "não, não é". Publicações especializadas mais recentes dão indícios de que essa resposta ainda se mantém. Por exemplo, o prefácio das recentes edições especiais sobre EPLs de periódicos (Villavicencio et al. 2005b, Rayson et al. 2010b) e dos anais do workshop sobre EPLs (Laporte et al. 2010, Kordoni et al. 2011b) listam diversos desafios a serem abordados no tratamento de EPLs, como multilinguismo, representação em léxicos e avaliação baseada em aplicações.

Uma das contribuições mais importantes desta tese é que ela representa um passo significativo na direção de uma integração completa de EPLs extraídas automaticamente em aplicações reais de PLN. Apesar disso, dada a complexidade do fenômeno, há uma

7. http://cameleon.imag.fr
8. http://jeuxdemots.imag.fr/por

necessidade constante de melhorias e, ao que tudo indica, é pouco provável que, num futuro próximo, seja proposta uma solução simples e única. Portanto, o objetivo em longo prazo desse trabalho pode ser resumido como estender e melhorar o trabalho apresentado nesta tese. Se por um lado um primeiro passo importante foi dado, por outro lado a estrada a percorrer ainda é longa.

APPENDIX C FURTHER READING: MWE ACQUISITION

C.1 Chinese

– Piao et al. (2006)
– Huang et al. (2005)

C.2 Japanese

– Ikehara et al. (2008)
– Hazelbeck and Saito (2010)
– Haugereid and Bond (2011)

C.3 Korean

– Lee (2011)
– Kim et al. (1999)
– Shimohata et al. (1997)

C.4 Bengali

– Das et al. (2010)
– Chakraborty and Bandyopadhyay (2010)
– Chakraborty et al. (2011)

C.5 Hindi

– Mukerjee et al. (2006)
– Sinha (2009)
– Sinha (2011)

C.6 Urdu

– Hautli and Sulger (2011)

C.7 Arabic

– Boulaknadel et al. (2008)

– Attia et al. (2010)

C.8 Greek

– Fotopoulou et al. (2008)
– Michou and Seretan (2009)

C.9 Czech

– Pecina (2005)
– Pecina (2008a)
– Pecina (2010)

C.10 Basque

– Alegria et al. (2004)
– Gurrutxaga and Alegria (2011)

C.11 Portuguese

– Silva and Lopes (1999)
– Gil and Dias (2003)
– Baptista et al. (2004)
– Silva and Lopes (2010)
– de Medeiros Caseli et al. (2009)
– Ramisch et al. (2010a)

C.12 Spanish

– Català and Baptista (2007)

C.13 Italian

– (Calzolari and Bindi 1990)
– Basili et al. (1994)
– Bonin et al. (2010a)
– Bonin et al. (2010b)
– Zaninello and Nissim (2010)
– Spina (2010)

C.14 French

– Daille (2003)
– Laporte et al. (2008)
– Green et al. (2011)
– Seretan (2008)
– Seretan and Wehrli (2009)

C.15 Dutch

- Bouma and Moirón (2001)
- de Cruys and Moirón (2007)

C.16 German

- Heid and Weller (2008)
- Weller and Heid (2010)
- Evert and Krenn (2005)
- Evert (2004)

C.17 English

- Piao et al. (2003)
- Baldwin and Villavicencio (2002)
- Baldwin (2005b)
- Kim and Baldwin (2010)
- Kim and Kan (2009)
- Cook et al. (2007)
- Fazly and Stevenson (2006)
- Fazly et al. (2009)
- North (2005)
- Stevenson et al. (2004)
- Fazly et al. (2007)
- Bannard (2007)

APPENDIX D RESOURCES USED IN THE EXPERIMENTS

D.1 Data

D.1.1 Monolingual corpora

- **English**
  - *British National Corpus (BNC)*. The BNC is a general-purpose corpus of British English, containing around 100 million words, mixing several genres like literature and newspapers (Burnard 2007). It is one of the most popular corpora in NLP for English. It is annotated with POS.
  - *Genia corpus*. It is composed of a set of 2,000 English abstracts of scientific articles from the biomedical domain (Ohta et al. 2002). It contains around 18K sentences and around 490K tokens. The corpus contains information about sentence and word boundaries, POS tags and terminological annotation with respect to the Genia ontology.
- **Portuguese**
  - *PLNBR-FULL corpus*. This corpus was built in the context of the PLNBR project (`www.nilc.icmc.usp.br/plnbr`). It contains 29,014,089 tokens of news text from *Folha de São Paulo*, a Brazilian newspaper, from 1994 to 2005. It can be considered as a general-purpose corpus of Brazilian Portuguese.

D.1.2 Multilingual corpora

- *Europarl corpus (EP)* The EP corpus contains transcriptions of the sessions of the European Parliament (Koehn 2005). It contains around 50 million words of parallel text in 11 languages of the European Union, including Portugues, English and French, plus around 10 million words for other languages of countries that recently joined the European Union. It can be viewed as a general-purpose corpus as it runs over more than 10 years and the political debates have a wide range of discussion subjects. It is one of the most popular resources for SMT. We used two versions: the older one, v3 , consists of extracts from the proceedings of the European Parliament during the period Apr/1996 – Oct/2006; and the more recent version, v6, contains the same texts as in v3 plus recent transcriptions up to Dec/2010. EP is publicly available at `http://www.statmt.org/europarl/`.
- *The web as a corpus*. The web contains a large amount of textual data in several languages. As discussed in Appendix F, it can be used to overcome data sparseness in traditional corpora. It is not a parallel corpus, but comparable corpora may be extracted from the web (Granada et al. 2012). It can also be thought of as a set containing several monolingual corpora, each one with millions of words. It is

practically impossible to crawl and download all the ever-growing text of the web, but search engines can be used to estimate the counts of words in the web. We use Google and Yahoo! search APIs and the implementation of the `mwetoolkit`.

D.2 Software

D.2.1 Analysis tools

- *Europarl corpus tools*. The EP corpus comes with some scripts for text tokenisation, sentence splitting and sentence alignment. These were used in some experiments.
- *TreeTagger*. The TreeTagger is a free downloadable POS tagger available for several languages, and with a good performance for English (Schmid 1994). It performs not only POS tagging but also sentence splitting, tokenisation and lemmatisation of the text. The TreeTagger is freely available at `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`. The tagset used by the TreeTagger in English is available at `ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz`.
- *PALAVRAS parser*. This deep syntactic parsing tool of Portuguese was used for the analysis of Portuguese text (Bick 2000). It supports tokenisation, sentence splitting, POS tagging, lemmatisation, dependency parsing annotation and shallow semantic annotation. In most cases, only the first four features were used.
- *RASP parser*. The RASP parser is a free downloadable tool for the syntactic analysis of English text (Briscoe et al. 2006). It provides not only POS tagging but also constituent and dependency trees. It is available at `http://www.informatics.susx.ac.uk/research/groups/nlp/rasp/`.

APPENDIX E THE MWETOOLKIT: DOCUMENTATION

This appendix contains a copy of the mwetoolkit documentation. Part of this material is available on the website http://mwetoolkit.sf.net. This documentation was produced with the help of the people cited in Section E.8.

E.1 Design choices

The mwetoolkit manipulates intermediary candidate lists and related elements as XML files. The use of XML as intermediary format has the advantage that it is readable and easy to validate according to a *document type definition* (DTD). It is also easy to import and export XML documents from and to other tools, as we describe in the next section. However, in terms of computational performance, the choice of an interpreted programming language like Python combined with a verbose file format like XML made some modules very slow and/or memory-consuming, requiring some optimisations. For example, the first versions of the indexing and candidate generation scripts were not able to deal with large corpora such as Europarl and the BNC. Therefore, some parts of the mwetoolkit were re-implemented in C. With the C indexing routine, for instance, indexing the BNC corpus takes about 5 minutes per attribute on a 3GB RAM computer.

In the implementation, instead of using the XML corpus and external matching procedures, we match candidates using Python's built-in regular expressions directly on the corpus index. This avoids parsing a huge XML file and speeds up pattern matching. On a small corpus, the current implementation takes about 72% the original time (using the XML file) to perform pattern-based extraction. On the BNC, extraction of is currently possible and takes from some minutes to a couple of hours.

Our target users are researchers with a background in computational linguistics and with some experience using command-line tools. The method is not a push-button utility that acquires any type of MWE from any type of corpus: it requires some manual tuning, pattern definition and parameter tuning. In sum, some trial and error iterations are needed in order to obtain the desired output.

Although no graphical user interface is available, we developed a "friendlier" command line interface. In the original version, one needed to manually invoke the Python scripts passing the correct options. The current version provides an interactive command-based interface which allows simple commands to be run on data files, while keeping the generation of intermediary files and the pipeline between the different phases of MWE extraction implicit. At the end, a user may want to save the session and restart the work later. Although it is not a graphical interface like some users requested, it is far easier to use than previous versions. In the future, we would like to develop a graphical interface, so that the toolkit can be used by researchers who are not at ease with the command line.

The mwetoolkit is a downloadable, freely available and open-source set of scripts. However, for more up-to-date documentation, as well as for downloading and testing the tool, one should prefer the official project website hosted at http://mwetoolkit.sourceforge.net/. Previous publications also describe earlier versions and punctual improvements of the methodology and of the tool (Ramisch et al. 2010b;c, Araujo et al. 2011).

E.2 Installing the mwetoolkit

E.2.1 Windows

Unfortunately, there is *NO WINDOWS VERSION AVAILABLE* of the mwetoolkit for the time being.

E.2.2 Linux and Mac OS

To install the mwetoolkit, just download it from the SVN repository using the following command:

```
svn co https://mwetoolkit.svn.sourceforge.net/svnroot/mwetoolkit mwetoolkit
```

The toolkit is also availaible as a stable release at https://sourceforge.net/projects/mwetoolkit/files/latest/download. However, as the code evolves fast, we recommend you to use the SVN version instead.

Once you have downloaded (and unzipped, in the case of a release) the toolkit, navigate to the main folder and run the command

```
make
```

for compiling the C libraries used by the toolkit. Do not worry about the warnings, they are normal. If you do not run this command, or if the command fails to compile the library, the toolkit will still work but it will use a Python version (much slower and possibly obsolete!) of the indexing and counting scripts. This may be OK for small corpora.

E.2.3 Mac OS dependencies

In addition to mwetoolkit itself, you will need to download and to configure some specific libraries.

E.2.3.1 Coreutils Package (through MacPorts)

To get this done is pretty simple, once you have MacPorts set up correctly (you can type man port and get a manual page), just run the following command:

```
sudo port install coreutils
```

If you don't have MacPorts yet, install it from http://www.macports.org/install.php/.

E.2.3.2 Simplejson (Python)

The Python installation comes with a handy utility called easy_install, which easily installs missing components: `sudo easy_install simplejson`

E.2.4 Testing your installation

The test folder contains regression tests for most scripts. In order to test your installation of the `mwetoolkit`, navigate to this folder and then call the script `testAll.sh`:

```
cd test ./testAll.sh
```

Should one of the tests fail, please send a copy of the output and a brief description of your configurations (operating system, version, machine) to our gmail, our username is `mwetoolkit`.

E.3 Getting started

`mwetoolkit` works by extracting MWE candidates from a corpus using a set of morphosyntactic patterns. Then it can apply a number of statistics to filter the extracted candidates. Input corpora, patterns and candidates are stored as XML files, following the format described by the DTDs in the `dtd` directory in the distribution. The toolkit consists of a set of scripts performing each phase of candidate extraction and analysis; these scripts are in the `bin` directory.

`mwetoolkit` receives as input a corpus as a XML file. This file contains a list of the sentences of the corpus. Each sentence is a list of words, and each word has a set of attributes (surface form, lemma, part of speech, and syntax information, if available). To obtain this information from a plain textual corpus without annotations, usually a part-of-speech tagger is used, which takes care of separating the input in tokens (words) and assigning a part-of-speech tag to each word.

To obtain a XML corpus from a plain textual corpus, you will usually use a tagger program or parser, such as explained in Section E.5 and in Section E.6.

E.3.1 An example

The toolkit comes with example files for a toy experiment in the directory `toy/genia`:

- `corpus.xml` — A small subset of the Genia corpus.
- `patterns.xml` — A set of patterns for matching noun compounds.
- `reference.xml` — A MWE reference (gold standard) for comparing the results of the candidate extraction against.

This directory also contains a script, `testAll.sh`, which runs a number of scripts on the example files. For each script run, it displays the action performed and the full command line used to run the script. It creates an output directory where it places the output files of each command.

Let's analyse each command that is run by `testAll.sh`. First, it runs `index.py` to generate an index for the corpus. This index contains suffix arrays for each word attribute in the corpus (lemma, surface form, part-of-speech, syntax annotation), which are used to search for and count the occurrences of an *n*-gram in the corpus. The full command executed is `index.py -v -i index/corpus corpus.xml`. The option `-i index/corpus` tells the script to use `index/corpus` as the prefix pathname for all index files (the `index` folder must exist). The `-v` option tells it to run in verbose mode (this is valid for all scripts).

After generating the index for the Genia fragment, it performs a candidate extraction by running:

```
candidates.py -p patterns.xml -i index/corpus >
candidates.xml
```

This invokes the candidate extraction script, telling it to use the patterns described in the file `patterns.xml`, and to use the corpus contained in the index files whose prefix is `corpus` (this is the same name given to the `index.py` script). Instead of using a patterns file, you could specify the `-n min:max` option to extract all *n*-grams with size between `min` and `max`.

Once candidates have been extracted, the counts of the individual words in each candidates are computed with the command:

```
counter.py -i index/corpus candidates.xml >
candidates-counted.xml
```

These counts are used by other scripts to compute statistics on the candidates. Word frequency cannot be computed directly from the XML file (it is done through binary search on the index). Instead of a corpus, you can count estimated word frequencies from the web, using either the Yahoo (option `-y` - DEPRECATED) or Google (option `-w`) search engine. You can also count word frequencies from an indexed corpus different from the one used for the extraction.

After word frequencies have been counted, association measures are calculated with the command:

```
feat_association.py -m mle:pmi:ll:t:dice
candidates-counted.xml >candidates-featureful.xml
```

The `-m measures` option is a colon-separated list specifying which measures are to be computed: Maximum Likelihood Estimator (`mle`), Pointwise Mutual Information (`pmi`), Student's t test score (`t`), Dice's Coefficient (`dice`), and Log-likelihood (`ll`, for bigrams only).

The association measures can be used in several ways. Here, we simply chose an association measure that we consider good, the t score, and sort the candidates according to this score, with the command:

```
sort.py -f t_corpus candidates-featureful.xml >
candidates-sorted.xml
```

The next script then works as Linux `head` command, cropping the sorted file and keeping only candidates with higher t score values. Finally, we compare the resulting candidates with a reference list containing some expressions that are already in a dictionary for the Genia biomedical domain. This is quite standard in MWE extraction, even though it only gives you an underestimation of the quality of the candidates as dictionaries are not complete. The command used in the evaluation is:

```
eval_automatic.py -r reference.xml -g candidates-crop.xml
> eval.xml 2> eval-stats.txt
```

The `-g` option tells the script to ignore parts of speech while the `-r` option indicates the file containing the reference gold standard in XML format. The final figures of precision and recall is in file `eval-stats.txt`. Remember that this is only a toy experiment and that with such a small corpus, the association measures cannot be trusted

For more advanced options, you can call the scripts using the `--help` option. This will print a message telling what the script does, what are the mandatory arguments and optional parameters. If you still have questions, write to our gmail address, username `mwetoolkit`, and we'll be happy to help!

E.4 Defining patterns for extraction

mwetoolit extracts MWE candidates by matching each sentence in the corpus against a set of patterns specified by the user. These patterns are read from XML files. This section describes the format of such files.

The root element of the XML patterns file is `<patterns>`. Inside this element comes a list of patterns, introduced by the tag `<pat>`. The `candidates.py` script will try to match each sentence of the corpus against each pattern listed:

```
<patterns>
    <pat>...</pat>
    <pat>...</pat>
    ...
</patterns>
```

E.4.1 Literal matches

The simplest kind of pattern is one that matches literal occurrences of one or more attributes in the corpus. This is done with the tag `<w attribute="value" .../>`. For example, to match an adjective followed by a noun, one could use the pattern: [1]

```
<pat>
    <w pos="J" />
    <w pos="N" />
</pat>
```

E.4.2 Repetitions and optional elements

It is possible to define regular-expression-like patterns, containing elements that can appear a variable number of times. This is done with the `repeat` attribute of the pat tag and with the either element. Note that `pat` elements can be nested.

```
<patterns>
    <!-- Pattern for matching a simple noun phrase. -->
    <pat>
        <!-- optional determiner (appearing 0 or 1 times) -->
        <pat repeat="?"><w pos="DT" /></pat>
        <!-- any number (including zero) of adjectives -->
        <pat repeat="*"><w pos="J" /></pat>
        <!-- one or more nouns -->
        <pat repeat="+"><w pos="N" /></pat>
    </pat>

    <pat>
        <!-- 3 to 5 adjectives -->
        <pat repeat="{3,5}"><w pos="J" /></pat>
        <!-- followed by the noun "dog" -->
        <w pos="N" lemma="dog" />
```

1. The actual part-of-speech tags depends on the convention used to tag the corpus, of course. Some tagging tools tag nouns with SUBST or NN, for instance.

```
    </pat>

    <!-- A sequence of nouns or adjectives
         followed by a final noun -->
    <pat>
        <pat repeat="*">
        <either>
            <pat>
                <w pos="N"/>
            </pat>
            <pat>
                <w pos="J"/>
            </pat>
        </either>
        </pat>
        <w pos="N"/>
    </pat>
</patterns>
```

E.4.3 Ignoring parts of the match

You can discard parts of a match by specifying an `ignore` attribute to the `<pat>` element:

```
<pat>
    <!-- Match a determiner, followed by any number
         of adjectives, followed by a noun. The
         adjectives are discarded from the match. -->
    <w pos="DT" />
    <pat repeat="*" ignore="true"><w pos="J" /></pat>
    <w pos="N" />
</pat>
```

E.4.4 Backpatterns

It is possible to create patterns with backreferences. For instance, you can match a word that has the same lemma as a previously matched word. To do this, you assign an `id` to the first word, and use `back:id.attribute` as the value of an attribute in a subsequent word:

```
<pat>
    <!-- Match N1-prep-N1 compounds (e.g.,
         step by step, day after day) -->
    <!-- Match a noun, labeled n1 -->
    <w pos="N" id="n1" />
    <!-- Match a preposition -->
    <w pos="P" />
    <!-- Match a noun whose lemma is the same as
         the lemma of n1 -->
    <w pos="N" lemma="back:n1.lemma" />
</pat>
```

Previous versions of the toolkit used <backw lemma="n1" /> instead of <w lemma="back:n1.lemma" />. There is no way of specifying both a literal attribute and a backreference with the old syntax.

E.4.5 Syntactic patterns

The toolkit supports corpora with syntactic annotations: the <w> element can contain a syn attribute, which contains a list of the syntactic dependencies of the word in the sentence, in the format deptype1:wordnum1;deptype2:wordnum2;..., where deptypen is the type of the dependency, and wordnumn is the number of the word that is the target of the dependency (first word is 1). For example, <w lemma="book" pos="N" syn="dobj:4" /> in the corpus represents a noun, book, which is the direct object of the fourth word in the sentence. (Again, the syntactic tag will vary depending on the convention used in the corpus.)

You can specify a pattern with syntactic dependencies with the attribute syndep in the <w> element of the patterns file. First you assign an id to a word, and then you refer back to it with the syntax <w syndep="deptype:id">. This is so that the pattern is not dependent on the actual word numbers. For example:

```
<!-- Match a verb and its direct object, with possible
     irrelevant intervening material. -->
<pat>
    <w pos="V" id="v1"/>
    <pat repeat="*" ignore="true"><w/></pat>
    <w pos="N" syndep="dobj:v1" />
</pat>
```

Currently only "backward" syntactic dependencies are supported. Support for forward dependencies is planned.

E.5 Preprocessing a corpus using TreeTagger

This section explains how to use the POS tagger, TreeTagger, to obtain a XML corpus from a plain textual corpus.

E.5.1 Installing TreeTagger

To install TreeTagger, just follow the instructions in the "Download" section of TreeTagger's webpage [2]. In addition to TreeTagger itself, you will need to download parameter files for each language you wish to use the tagger with. We recommend that you add the path to TreeTagger to your PATH variable as suggested by the TreeTagger installation script, this will allow you to call it without using the full path.

E.5.2 Converting TreeTagger's output to XML

After installing TreeTagger, you can run it by running path-to-tree-tagger/cmd/tree-tagger-language input-file, where language is the language of the input file. TreeTagger will read the corpus from input-file and print each word, together with its surface form and part of speech, as a separate line to standard output.

2. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

mwetoolkit comes with a script, treetagger2xml.sh, which takes TreeTagger's output and converts it to XML. All you have to do is feed TreeTagger's output to it:

```
path-to-tree-tagger/cmd/tree-tagger-english corpus.txt |
python path-to-mwetoolkit/bin/treetagger2xml.py >corpus.xml
```

From there on you can process the XML corpus using mwetoolkit tools, such as is shown in Section E.3.

E.6 Preprocessing a corpus using RASP

This page explains how to use the Parser, RASP, to obtain a XML corpus from a plain textual corpus.

E.6.1 Installing RASP

RASP doesn't need to be installed. Just download it from RASP Download [3]. However, it assumes that you have downloaded it to your home directory, so do it and save yourself some headache.

E.6.2 Converting RASP's output to XML

After downloading RASP, you can run it by running path-to-rasp/scripts/rasp.sh < input-file. RASP will read the corpus from input-file and print for each sentence it's words, together with surface form, lemma and part of speech. Then will print the grammatical relations, which can be viewed as a kind of dependency tree, from where will be extracted the syntactic property, in separate lines to standard output.

mwetoolkit comes with a script, rasp2mwe.py, which takes RASP's output and converts it to XML. All you have to do is feed RASP's output to it:

```
path-to-rasp/scripts/rasp.sh < corpus.txt |
python path-to-mwetoolkit/bin/rasp2mwe.py >corpus.xml
```

From there on you can process the XML corpus using mwetoolkit tools, such as is shown in Section E.3.

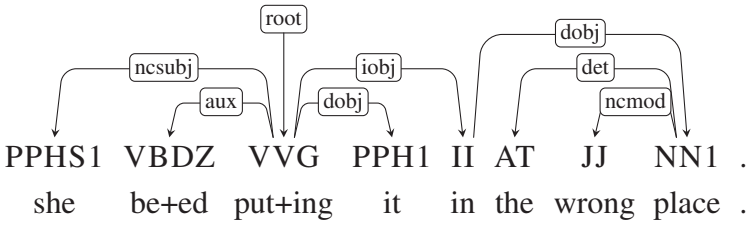
E.7 Examples of XML files

Figure E.1 shows an example of sentence in a XML corpus file. There are four possible attributes that can be defined at the word level: surface for the surface form, lemma, pos for the part of speech and syn for the dependency syntactic relation. Syntactic relations are represented as a pair type:parent where the first element is the type of syntactic relation and the second element is the position of the parent word on which the current word depends. This example sentence was parsed using the RASP parser. Word indices start at 1. The corresponding tree representation would be:

3. http://ilexir.co.uk/applications/rasp/download/

```
<s s_id="4">
    <w surface="She" lemma="she" pos="PPHS1" syn="ncsubj:3" />
    <w surface="was" lemma="be" pos="VBDZ" syn="aux:3" />
    <w surface="putting" lemma="put" pos="VVG" syn="" />
    <w surface="it" lemma="it" pos="PPH1" syn="dobj:3" />
    <w surface="in" lemma="in" pos="II" syn="iobj:3" />
    <w surface="the" lemma="the" pos="AT" syn="det:8" />
    <w surface="wrong" lemma="wrong" pos="JJ" syn="ncmod:8" />
    <w surface="place" lemma="place" pos="NN1" syn="dobj:5" />
    <w surface="." lemma="." pos="." syn="" />
</s>
```

Figure E.1: Example of sentence in a corpus.



E.8 Developers

The `mwetoolkit` is developed and maintained by:

- Carlos Ramisch
- Vitor De Araujo
- Sandra Castellanos
- Maitê Dupont

APPENDIX F THE WEB AS A CORPUS

This chapter discusses the use of the web as a corpus. Most of the materials presented here were taken from the paper *Web-based and combined language models: a case study on noun compound identification* (Ramisch et al. 2010d). Please refer to the original paper for more details and experimental results.

F.1 Introduction to the web as a corpus

Corpora have been extensively employed in several NLP tasks as the basis for automatically learning models for language analysis and generation. They are the basis of the work presented in this thesis. In theory, *data-driven* (*empirical* or *statistical*) approaches are well suited to take intrinsic characteristics of human language into account. In practice, external factors also determine to what extent they will be popular and/or effective for a given task, so that they have shown different performances according to the availability of corpora and to the linguistic complexity of the task.

An essential component of most empirical systems is the *language model* (LM) and, in particular, *n-gram language models*. It is the LM that tells the system how likely a word or *n*-gram is in that language, based on the counts obtained from corpora. However, corpora represent a sample of a language and will be sparse, that is, certain words or expressions will not occur. One alternative to minimise the negative effects of data sparseness and account for the probability of out-of-vocabulary words is to use discounting techniques, where a constant probability mass is discounted from each *n*-gram and assigned to unseen *n*-grams. Another strategy is to estimate the probability of an unseen *n*-gram by backing off to the probability of the smaller *n*-grams that compose it.

In recent years, there has also been some effort in using the web to overcome data sparseness, given that the web is several orders of magnitude larger than any available corpus. However, it is not straightforward to decide whether (a) it is better to use the web than a standard corpus for a given task or not, and (b) whether corpus and web counts should be combined and how this should be done (e.g., using interpolation or back-off techniques). As a consequence there is a strong need for better understanding of the impacts of web frequencies in NLP systems and tasks.

F.2 Related work

Conventional and, in particular, domain-specific corpora, are valuable resources which provide a closed-world environment where precise *n*-gram counts can be obtained. As they tend to be smaller than general purpose corpora, data sparseness can considerably

hinder the results of statistical methods. For instance, in the biomedical Genia corpus (Ohta et al. 2002), 45% of the words occur only once (so-called *hapax legomena*), and this is a very poor basis for a statistical method to decide whether this is a significant event or just random noise.

One possible solution is to see the web as a very large corpus containing pages written in several languages and being representative of a large fraction of human knowledge. However, there are some differences between using regular corpora and the web as a corpus, as discussed by Kilgarriff and Grefenstette (2003). One assumption, in particular, is that page counts can approximate word counts, so that the total number of pages is used as an estimator of the *n*-gram count, regardless of how many occurrences of the *n*-gram they contain.

This simple underlying assumption has been employed for several tasks. For example, Grefenstette (1999), in the context of example-based machine translation, uses web counts to decide which of a set of possible translations is the most natural one for a given sequence of words (e.g., *groupe de travail* as *work group* vs *labour collective*). Likewise, Keller and Lapata (2003) use the web to estimate the frequencies of unseen nominal bigrams, while Nicholson and Baldwin (2006) look at the interpretation of noun compounds based on the individual counts of the nouns and on the global count of the compound estimated from the web as a large corpus.

Villavicencio et al. (2007) show that the web and the BNC could be used interchangeably to identify general-purpose and type-independent multiword expressions. Lapata and Keller (2005) perform a careful and systematic evaluation of the web as a corpus in other general-purpose tasks both for analysis and generation, comparing it with a standard corpus (the BNC) and using two different techniques to combine them: linear interpolation and back-off. Their results show that, while web counts are not as effective for some tasks as standard counts, the combined counts can generate results, for most tasks, that are as good as the results produced by the best individual corpus between the BNC and the web. Nakov (2007) further investigates these tasks and finds that, for many of them, effective attribute selection can produce results that are at least comparable to those from the BNC using counts obtained from the web.

On the one hand, the web can minimise the problem of sparse data, helping distinguish rare from invalid cases. Moreover, a search engine allows access to ever increasing quantities of data, even for rare constructions and words, which counts are usually equated to the number of pages in which they occur. On the other hand, *n*-grams in the highest frequency ranges, such as the words *the*, *up* and *down*, are often assigned the estimated size of the web, uniformly. While this still gives an idea of their massive occurrence, it does not provide a finer grained distinction among them (e.g., in the BNC, *the*, *down* and *up* occur 6,187,267, 84,446 and 195,426 times, respectively, while in Yahoo! they all occur in 2,147,483,647 pages).

F.3 Standard vs web corpora

When we compare *n*-gram counts estimated from the web with counts taken from a well-formed standard corpus, we notice that web counts are "estimated" or "approximated" as page counts, whereas standard corpus counts are the exact number of occurrences of the *n*-gram. In this way, web counts are dependent on the particular search engine's algorithms and representations, and these may perform approximations to handle the large size of their indexing structures and procedures, such as ignoring punctuation

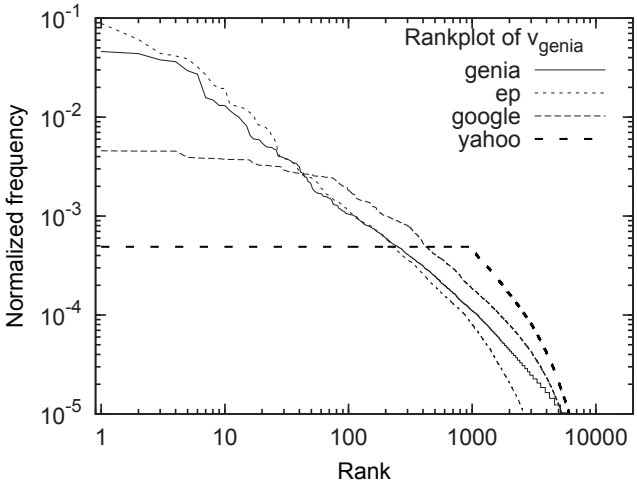


Figure F.1: Plot of normalised frequencies of the Genia vocabulary according to rank positions, log-log scale.

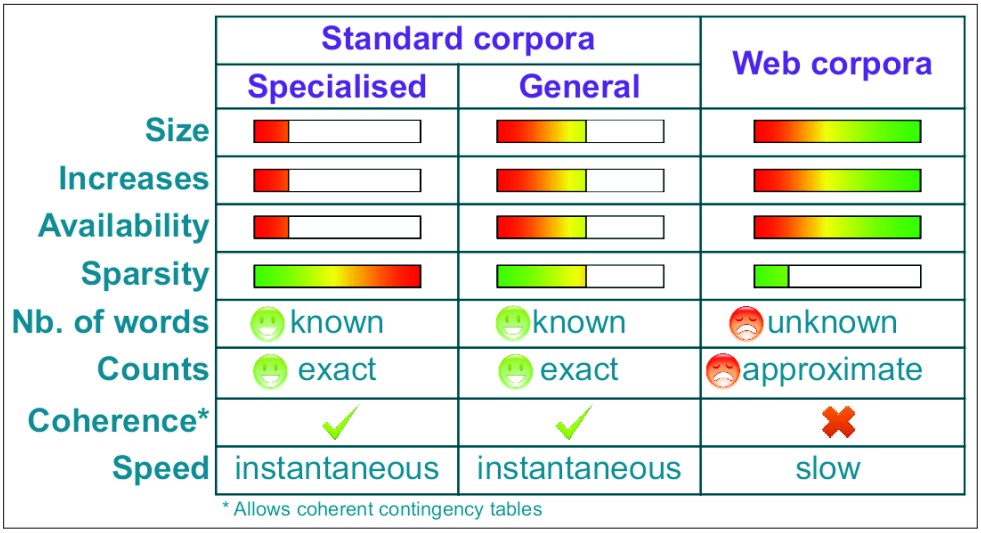


Figure F.2: Summary of differences between a specialised corpus, a general-purpose corpus and the web as a corpus.

and using stopword lists (Kilgarriff 2007). This assumption, as well as the following discussion, are not valid for for controlled data sets derived from web data, such as the Google 1 trillion *n*-grams [1] (Bergsma et al. 2009).

In data-driven techniques, some statistical measures are based on contingency tables, and the counts for each of the table cells can be straightforwardly computed from a standard corpus. However, this is not the case for the web, where the occurrences of an *n*-gram are not precisely calculated in relation to the occurrences of the $(n-1)$-grams composing it. For instance, the *n*-gram *the man* may appear in 200,000 pages, while the words *the* and *man* appear in respectively 1,000,000 and 200,000 pages, implying that the word *man* occurs with no other word than *the*. [2]

In addition, the distribution of words in a standard corpus follows the well known

1. This dataset is released through LDC and is not freely available. Therefore, we do not consider it in our evaluation.

2. In practice, this procedure can lead to negative counts.

Zipfian distribution (Baayen 2001) while, in the web, it is very difficult to distinguish frequent words or *n*-grams as they are often estimated as the size of the web. For instance, the Yahoo! frequencies plotted in Figure F.1 are flattened in the upper part, giving the same page counts for more than 700 of the most frequent words. Another issue is the size of the corpus, which is an important information, often needed to compute frequencies from counts or to estimate probabilities in *n*-gram models. Unlike the size of a standard corpus, which is easily obtained, it is very difficult to estimate how many pages exist on the web, especially as this number is always increasing.

But perhaps the biggest advantage of the web is its availability, even for resource-poor languages and domains. It is a free, expanding and easily accessible resource that is representative of language use, in the sense that it contains a great variability of writing styles, text genres, language levels and knowledge domains. Figure F.2 summarises the pros and cons of using the web as a corpus compared to standard general-purpose and specialised corpora.

APPENDIX G DETAILED LEXICON DESCRIPTIONS

G.1 Dimensions of Fujitsu's ATLAS lexicon

Dimensions of Fujitsu's technical dictionaries v14. Source: `http://www.fujitsu.com/global/services/software/translation/atlas/system/techdics.html`

Field	`en-jp` **entries**	`jp-en` **entries**
Chemistry	226000	222000
Medicine: Disease, Symptoms	216000	218000
Person's and Place Name	209000	211000
Medicine: Biochemistry	205000	207000
Physics and Atomic Energy	176000	178000
Information Processing	174000	175000
Biology	162000	164000
Business	149000	151000
Unno's Business Dictionary	149000	39000
Electrical Engineering	125000	127000
Mechanical Engineering	115000	117000
Construction, Architecture	90000	91000
Medicine: Pharmacology	83000	84000
Earth Science and Astronomy	81000	83000
PC Dialog Messages	76000	76000
Agriculture and Fisheries	70000	71000
Finance and Economics	65000	66000
Factory Facilities	63000	64000
Transportation	63000	64000
Medicine: Anatomy	57000	59000
Medicine: Medical Equipment	52000	54000
Metal	46000	47000
Motor Vehicles	43000	45000
Environment	35000	36000
Medicine: Psychiatry	32000	33000
Biology and Biochemistry	31000	32000
Military	26000	28000
Law	18000	19000
Total	2837000	2761000

G.2 Sentiment verbs extracted from Brazilian WordNet

List of sentiment verbs extracted from the Brazilian version of WordNet. We intend to investigate the relation between these verbs and the complex predicates extracted in Section 6.2.

abalar	abominar	aborrecer-se	abrandar
acalmar	acalmar-se	acender-se	acovardar-se
adorar	afligir	agitar-se	agradar
alarmar	alarmar-se	alegrar	aliviar
alterar	alucinar	alvoroçar	animar
antipatizar	apiedar	apoquentar	apreciar
arrasar	assanhar	atormentar-se	atraiçoar
atrair	atrapalhar-se	babar-se	cativar
chatear	cobiçar	comover	comover-se
compadecer-se	conciliar	confortar	conquistar
consolar	consolar-se	consumir-se	decepcionar
decepcionar-se	deleitar-se	desadorar	desagradar
desagradar-se	desagradecer	desalentar-se	desangustiar
desanimar	desapoquentar	desassossegar	desconfortar
desejar	desemburrar	desemburrar-se	desencabular
desencorajar	desenjoar	desesperar-se	desfazer-se
desiludir	desinteressar	desmotivar	despertar
despreocupar	desprezar	distrair-se	doer
embaraçar	emburrar	encantar	encantar-se
encorajar	enfurecer	enfurecer-se	enlouquecer
enlouquecer-se	enlutar	enlutar-se	entristecer
entristecer-se	entusiasmar	entusiasmar-se	envaidecer-se
envergonhar	espezinhar	estimar	estimular-se
exasperar	exasperar-se	excitar	expectar
expiar	fascinar	frustrar	fustigar
horrorizar	horrorizar-se	humilhar-se	impacientar-se
incomodar	inferiorizar-se	inquietar-se	intimidar
intimidar-se	invejar	irar-se	irritar-se
irromper	lastimar	magoar-se	malucar
nublar	nublar-se	obsequiar	orgulhar-se
penitenciar-se	perrengar	perturbar	perturbar-se
pirraçar	preferir	preocupar-se	rebaixar-se
simpatizar	sossegar	temer	torturar
venerar	zangar		

G.3 Sentiment nouns identified

List of Brazilian Portuguese nouns expressing sentiments. These nouns were identified using the morphosyntactic patterns described in Section 6.2.1.2 and manual validation.

admiração	adoração	ambição	amor
angústia	ansiedade	antipatia	apego
apelo	apreço	asco	aspiração
atração	bronca	carinho	certeza
cheiro	choque	ciúme	compaixão
complexo	confiança	consciência	constrangimento
convicção	coragem	culpa	curiosidade
desejo	desespero	desprezo	devoção
dificuldade	disposição	dó	dor
dor-de-cabeça	dúvida	esperança	expectativa
fadiga	falta	fascinação	fobia
fome	frio	gosto	horror
ímpeto	impressão	instinto	interesse
inveja	irritação	mágoa	medo
moleza	necessidade	nojo	nostalgia
obsessão	ódio	orgulho	paciência
paixão	pânico	pavor	pena
piedade	prazer	predileção	preguiça
preocupação	pudor	raiva	rancor
receio	rejeição	remorso	repugnância
repulsa	respeito	responsabilidade	sabor
saudade	segurança	sensação	sentimento
simpatia	sintoma	suador	suspeita
tentação	tranquilidade	trauma	tristeza
vergonha	vontade		