Collocation Extraction
Based on Syntactic Parsing

THÈSE
présentée à la Faculté des lettres de l’Université de Genève
pour obtenir le grade de Docteur ès lettres

par
Violeta Seretan

Thèse N° 653

GENÈVE
2008
La Faculté des lettres, sur le préavis d’une commission composée de MM. les professeurs Jacques MOESCHLER, président du jury; Eric WEHRLI, directeur de thèse; Christian BOITET (IMAG, Grenoble); Ulrich HEID (IMS, Stuttgart); Paola MERLO (Genève) autorise l’impression de la présente thèse, sans exprimer d’opinion sur les propositions qui y sont énoncées.

Genève, le 9 juin, 2008

*Le Doyen:* Eric WEHRLI

**Thèse N° 653**
To my daughter, Nadia Giulia,
for sweet collocation
during thesis writing
Abstract

Pervasive across texts of different genres and domains, collocations (typical lexical associations like *to wreak havoc, to meet a condition, to believe firmly, a deep concern, highly controversial*) constitute a large proportion of the multi-word expressions in a language. Due to their encoding idiomaticity, collocations are of paramount importance to text production tasks. Their recognition and appropriate usage is essential, for instance, in Foreign Language Learning or in Natural Language Processing applications such as machine translation and natural language generation. At the same time, collocations have a wide applicability to tasks concerned with the opposite process of text analysis.

The problem that is tackled in this thesis is the automatic acquisition of accurate collocational information from text corpora. More specifically, the thesis provides a methodological framework for the syntax-based identification of collocation candidates in the source text, prior to the statistical computation step. The development of syntax-based approaches to collocation extraction, which has traditionally been hindered by the absence of appropriate linguistic tools, is nowadays possible thanks to the advances achieved in parsing. Until now, the absence of sufficiently robust parsers was typically circumvented by applying linear proximity constraints in order to detect syntactic links between words. This method is relatively successful for English, but for languages with a richer morphology and a freer word order, parsing is a prerequisite for a good performance.

The thesis proposes (and fully evaluates on data in four different languages, English, French, Spanish and Italian) a core extraction procedure for discovering binary collocations, which is based on imposing syntactic constraints on the component items.
instead of linear proximity constraints. This procedure is further employed in several methods of advanced extraction, whose aim is to cover a broader spectrum of collocational phenomena in text. Three distinct but complementary extension directions have been considered in this thesis: extraction of \( n \)-ary collocations (\( n > 2 \)), data-driven induction of collocationally relevant syntactic configurations, and collocation mining from an alternative source corpus, the World Wide Web. The possibility to abstract away from the surface text form and to recover, thanks to parsing, the syntactic links between discontinuous elements in text, plays a crucial role in achieving highly efficient results.

The methods proposed in this study were adopted in the development of an integrated system of collocation extraction and visualization in parallel corpora, a system which was intended to enrich the workbench of translators or other users (e.g., terminologists, lexicographers, language learners) wanting to exploit their text archives. Finally, the thesis gives an example of a practical application that builds on this system in order to further process the extracted collocations, by automatically translating them when parallel corpora are available.
Résumé

Largement présentes dans les textes de tout genre et de tout domaine, les collocations se taillent la part du lion dans l’inventaire des expressions à mots multiples d’une langue. En raison du caractère idiomatique de leur encodage, les collocations jouent un rôle primordial dans la production de texte (par exemple, dans des applications telles que la traduction automatique et la génération de texte, ainsi que dans l’apprentissage de langues étrangères). En même temps, elles présentent également un grand intérêt du point de vue de l’analyse du texte.

Le sujet de cette thèse est l’acquisition automatique, basée sur corpus, de ressources lexicales collocationnelles. Dans la thèse, nous proposons des méthodes précises pour l’identification de collocations candidates, qui s’appuient sur une analyse syntaxique détaillée du texte source effectuée préalablement aux calculs statistiques. Le développement des approches syntaxiques pour l’extraction a été jusque maintenant entravé par l’absence d’outils d’analyse appropriés, mais cette situation est en train de changer grâce aux progrès réalisés dans le domaine du parsing. Si pour l’anglais le manque d’analyseurs suffisamment robustes a pu être compensé partiellement par l’application de contraintes de proximité linéaire afin de détecter des liens syntaxiques entre les mots, pour d’autres langues il est impératif de faire appel au parsing afin d’obtenir des résultats d’extraction fiables.

Cette thèse propose (et évalue sur des données en 4 langues, l’anglais, le français, l’espagnol et l’italien) une procédure principale d’extraction de collocations binaires qui se base sur l’application de la contrainte de proximité syntaxique aux éléments d’une collocation candidate, à la place de la contrainte de proximité linéaire qui est la plus répandue dans les travaux existants. Cette procédure est ensuite utilisée dans
la conception de plusieurs méthodes d’extraction avancées, qui visent à couvrir un éventail plus large de phénomènes collocationnels dans le texte. Ces méthodes vont dans trois directions distinctes, mais complémentaires : l’extraction de collocations n-aires ($n > 2$), l’induction dirigée par les données des configurations syntaxiques appropriées aux collocations, et la détection de collocations à partir d’un corpus alternatif, le Web. La possibilité de faire abstraction de la forme superficielle du texte et celle de récupérer, grâce au parsing, les liens syntaxiques entre des éléments lexicaux qui ne sont pas forcément contigus dans le texte jouent un rôle déterminant dans l’obtention de résultats performants.

La méthodologie présentée a été adoptée dans la création d’un système intégré d’extraction et de visualisation de collocations dans des corpus multilingues, destiné à enrichir l’environnement de travail des traducteurs, et, en général, celui des utilisateurs intéressés à exploiter leurs archives textuelles (terminologues, lexicographes, ou apprenants de langues étrangères). La thèse présente aussi une application basée sur ce système, qui traite les résultats d’extraction afin d’effectuer leur traduction automatique lorsque des corpus parallèles sont disponibles.
Acknowledgements

I feel very fortunate of having found at LATL, the Language Technology Laboratory of the University of Geneva, a great working environment in which I could conduct my PhD studies during the past years. I am most grateful to my supervisor, Eric Wehrli, who gave me the opportunity to join his team and work on a fascinating topic—first, in the framework of the project “Linguistic Analysis and Collocation Extraction”, then as part of my thesis. Eric guided my work with his vast knowledge and infinite patience. I appreciated his good advices, his constant support and quick feedback, as much as I appreciated the freedom he gave me to develop my own ideas, or his permanent good mood and optimism. Thank you very much!

Since part of my thesis work took already shape during the project I mentioned, I wish to gratefully acknowledge financial support from the RUIG-GIAN organisation. I also want to thank our project partner, Olivier Pasteur, for his active involvement in this project and for suggestions on the design of our system. Together with him, with Luka Nerima and Eric Wehrli, we had lengthy regular meetings which set the framework of my thesis. Luka further accompanied my work with his expertise in database technology, and was constantly available for discussions. Thank you for your contribution!

During these years, I had the opportunity to present my work in many local or non-local meetings. The feedback I received from colleagues during research seminars and doctoral summer schools, from invited researchers in our lab, from fellow conference attendees and, particularly, from the anonymous reviewers of my papers, left a definite mark on my work. I wish to thank all the people involved for their invaluable help.
Many of my colleagues in the Department of Linguistics provided me with useful comments and suggestions on various occasions. I received valuable criticism, in particular, from (in office-room order): Paola Merlo, Gabriele Musillo, Eric Joanis, Gerold Schneider, Jacques Möeschler, Christopher Laenzlinger, and Antonio Leoni. My computer-scientist office mates, Jean-Philippe Goldman, Mar Ndiaye, and Yves Scherrer, always gave me good advices and were ready to lend me a hand. JP made many useful suggestions for the extraction tool, especially in the early stage. Mar tested this tool several times, and together with Yves and Antonio, helped me annotate the multilingual data I used in evaluation (my husband Vincenzo did not escape the torture; he had to annotate the Italian data).

Other colleagues have contributed to my work in different ways. Occasionally, Gabriela Soare, Stephanie Durrleman-Tame, Catherine Walther Green and Genoveva Puskas helped me proofread my papers. I am grateful to them, and in general to all the members of our Department for maintaining a stimulating working environment. But more than a team, I found here a family and friends. Since Eva Capitão is our friendship hub, I wish to thank her in particular for her kindness, as well as for her support with various administrative matters.

A different type of support came from the Commission of Equality of our University, which granted me a temporary exemption from teaching so that I could focus on my thesis. During the last stages of thesis writing, Vincenzo’s help was essential, as he spend so much time taking care of our new-born daughter, Nadia Giulia. Thank you both for the joy you are bringing into my life!

Special thanks are also due to my Romanian family for love and encouragements; to my first teacher, Vasile Frumos, for making me love simplicity and rigor; and to Dan Cristea from the University of Iași, for introducing me to Computational Linguistics.

Extra special thanks go to my friends around the world, who persistently inquired about the progress of my PhD. I am also indebted to Alex Clark, Laura Hasler and Eric Joanis, who proofread a previous version of this manuscript.

Finally, I wish to express my gratitude to Christian Boitet, Ulrich Heid and Paola Merlo, who kindly agreed to be part of the thesis committee, and to Jacques Möeschler, who accepted the role of president of the jury.
## Contents

Abstract vii
Résumé vii
Acknowledgements ix

1 Introduction 1
  1.1 Motivation .......................................................... 1
  1.2 Objectives .......................................................... 5
  1.3 Chapters outline .................................................... 7
  1.4 Published work ..................................................... 8

2 On Collocations 9
  2.1 Introduction .......................................................... 9
  2.2 A survey of definitions ............................................. 10
    2.2.1 Statistical approaches ......................................... 12
    2.2.2 Linguistic approaches ......................................... 14
    2.2.3 Collocation vs. co-occurrence ............................... 16
  2.3 Towards a core collocation concept ............................. 16
  2.4 Theoretical perspectives on collocations ........................ 20
    2.4.1 Contextualism .................................................. 21
    2.4.2 Text cohesion .................................................. 22
    2.4.3 Meaning-Text Theory ........................................... 22
    2.4.4 Other perspectives ............................................. 25