Computer-assisted translation (CAT) has always used translation memories, which require the translator to have a corpus of previous translations that the CAT software can use to generate bilingual lexicons. This can be problematic when the translator does not have such a corpus, for instance, when the text belongs to an emerging field. To solve this issue, CAT research has looked into the leveraging of comparable corpora, i.e. a set of texts, in two or more languages, which deal with the same topic but are not translations of one another.

This work had two primary objectives. The first is to assess the input of lexicons extracted from comparable corpora in the context of a specialized human translation task. The second objective is to identify bilingual-lexicon-extraction methods which best match the translators’ needs, determining the current limits of these techniques and suggesting improvements. The author focuses, in particular, on the identification of fertile translations, the management of multiple morphological structures, and the ranking of candidate translations.

The experiments are carried out on two language pairs (English–French and English–German) and on specialized texts dealing with breast cancer. This research puts significant emphasis on applicability – methodological choices are guided by the needs of the final users. This book is organized in two parts: the first part presents the applicative and scientific context of the research, and the second part is given over to efforts to improve compositional translation.

The research work presented in this book received the PhD Thesis award 2014 from the French association for natural language processing (ATALA).
Comparable Corpora and Computer-assisted Translation
To Elia
Comparable Corpora and Computer-assisted Translation

Estelle Maryline Delpech
# Table of Contents

## ACKNOWLEDGMENTS

ix

## INTRODUCTION

xi

## PART 1. APPLICATIVE AND SCIENTIFIC CONTEXT

1

## CHAPTER 1. LEVERAGING COMPARABLE CORPORA FOR COMPUTER-ASSISTED TRANSLATION

3

1.1. Introduction ........................................... 3

1.2. From the beginnings of machine translation to comparable corpora processing ........................................... 3

1.2.1. The dawn of machine translation .......................... 3

1.2.2. The development of computer-assisted translation ........ 5

1.2.3. Drawbacks of parallel corpora and advantages of comparable corpora ............................ 7

1.2.4. Difficulties of technical translation ........................ 9

1.2.5. Industrial context ...................................... 13

1.3. Term alignment from comparable corpora: a state-of-the-art .............................. 15

1.3.1. Distributional approach principle .......................... 15

1.3.2. Term alignment evaluation ................................ 18

1.3.3. Improvement and variants of the distributional approach .............................. 20

1.3.4. The influence data and parameters on alignment quality .............................. 28

1.3.5. Limits of the distributional approach ........................ 30

1.4. CAT software prototype for comparable corpora processing .............................. 32

1.4.1. Implementation of a term alignment method .............................. 32

1.4.2. Terminological records extraction .............................. 36
1.4.3. Lexicon consultation interface .......................... 38
1.5. Summary ................................................. 38

CHAPTER 2. USER-CENTERED EVALUATION OF LEXICONS
EXTRACTED FROM COMPARABLE CORPORA ...................... 41
  2.1. Introduction ............................................. 41
  2.2. Translation quality evaluation methodologies ............... 42
    2.2.1. Machine translation evaluation ......................... 42
    2.2.2. Human translation evaluation .......................... 46
    2.2.3. Discussion ......................................... 49
  2.3. Design and experimentation of a user-centered evaluation ...... 50
    2.3.1. Methodological aspects ................................ 50
    2.3.2. Experimentation protocol .............................. 54
    2.3.3. Results ............................................ 58
  2.4. Discussion .............................................. 64

CHAPTER 3. AUTOMATIC GENERATION OF TERM TRANSLATIONS ...... 67
  3.1. Introduction ............................................. 67
  3.2. Compositional approaches ................................ 67
    3.2.1. Compositional translation principle ..................... 68
    3.2.2. Polylexical units compositional translation ............. 70
    3.2.3. Monolexical units compositional translation ............. 75
    3.2.4. Candidate translation filtering ......................... 81
  3.3. Data-driven approaches ................................ 84
    3.3.1. Analogy-based translation ............................. 85
    3.3.2. Rewriting rules learning .............................. 87
    3.3.3. Dealing with morphological variation .................... 88
  3.4. Evaluation of term translator generation methods ............. 91
  3.5. Research perspectives ................................ 95

PART 2. CONTRIBUTIONS TO COMPOSITIONAL TRANSLATION ........ 99

CHAPTER 4. MORPH-COMPOSITIONAL TRANSLATION:
METHODOLOGICAL FRAMEWORK .................................. 101
  4.1. Introduction ............................................. 101
  4.2. Morpho-compositional translation method .................... 101
    4.2.1. Scientific positioning ................................ 102
    4.2.2. Definitions and terminology .......................... 105
    4.2.3. Underlying assumptions ............................... 108
    4.2.4. Advantages of the proposed approach for processing comparable corpora .................................. 109
  4.3. Addressed issues and contributions ........................ 110
    4.3.1. Generating fertile translations ......................... 110
I would like to give Béatrice Daille and Emmanuel Morin my deepest thanks for having supervised and co-supervised this doctoral research. I felt so honored to work and learn by their sides. They both have shown an effective combination of academic rigor and pedagogy, which has helped me to progress over the past three years. Béatrice, thank you for suggesting this thesis to me – it was a very enriching experience. Manu, thanks again for being there whenever I needed you.

I extend my warm thanks to Nabil Hathout, Élisabeth Lavault-Olléon, Emmanuel Planas and Michel Simard for honoring me by attending my viva and being part of my committee in spite of their busy schedules. Their constructive comments were especially useful. I am glad to have profited from so many complementary points of view on my work. Special thanks to Michel Simard for coming to Nantes all the way from Canada.

I am particularly thankful to Emmanuel Planas, the former Scientific Director of Lingua et Machina, for trusting me and hiring me as a research engineer five years ago. Otherwise, I would probably not have had the chance to carry out a dissertation with LINA or to work on such a fascinating research subject within such a stimulating industrial environment.

Several people contributed to the work presented in this book. I would first like to thank Claire Lemaire from the University Stendhal in Grenoble, because she was an amazing colleague and co-doctoral candidate; and also for creating the resources for processing and assessing the German language. This would not have been possible without her and I am very thankful.

I would also like to thank Geoffrey Williams and Pierre Zweigenbaum for agreeing to be a part of my doctorate monitoring committee. Their shrewd comments and advice guided me in the right direction during this research project.
I would also like to thank Léa Laporte of the Toulouse Information Technology Research Institute (Institut de Research en Informatique de Toulouse (IRIT)) and Damien François from the Catholic University in Louvain for answering my questions on the statistical data-processing. Thanks also to Van Dang, from the University of Massachusetts, for answering my questions on the use of learning-to-rank algorithms.

I am immensely grateful to Clémence de Baudus, Kiril Isakov, Mathieu Delage from the Higher Education Institute of Translation and Interpreting (ISIT) and Nicolas Auger, for their extremely detailed annotation, without which it would not have been possible to evaluate the translation system.

My thanks go also to my colleagues at Lingua et Machina, François, Étienne and Jean-François, with whom I learned a lot and who I thank for their support. François’s advice and experience were invaluable during the last year of my doctorate.

Unfortunately, I did not have the chance to spend a lot of time in the laboratory but it was always a pleasure to come to the team meetings. The welcome and atmosphere of the LINA is fantastic, and I greatly enjoyed talking with my colleagues – especially Amir Hazem and Prajol Shrestha, who were delightful fellow doctoral candidates.

Finally, I would like to thank my partner, Nicolas, for his unfailing support; my friends Émilie and Nathalie and my sister Laureen for their understanding about my lack of availability, as well as for their presence and logistical support on the day of the viva. Thanks to Loki, who is a wonderful alarm clock.
I.1. Socio-economic stakes of multilingualism management

In the days of the globalization of exchanges, multilingualism is an undeniable socio-cultural asset, but it presents many challenges to our society.

First of all, the lack of knowledge of a language is often synonymous with limited access to information, and it is generally linguistic communities with little economic power, or whose language is not a prestigious one, who suffer as a result.

The case of the Internet is a good example: English – the most represented language on the web (54.8%)\(^1\) – is the first language of only 26.8% of web users\(^2\) whereas Chinese – the first language of 24.2% of the web users – is only sixth in terms of presence on the Internet (4%).

A significant portion of web-based information is therefore unavailable to many web users because of the language barrier.

In countries which are officially bilingual or multilingual or in international organizations such as the European Union, managing multilingualism falls within the remit of democracy: it is meant to ensure that each citizen has access to administrative services and legal texts in his own first language so she/he knows his/her rights and can benefit from the government’s services in a language she/he speaks fluently. This has a considerable cost: the European Union spends 1 billion Euros every year in translation and interpretation costs [FID 11].

---

1 In May 2011, according to WEB TECHNOLOGY SURVEYS http://w3techs.com/technologies/overview/content_language/all.
Multilingualism also has an impact on our economy: the ELAN report [HAG 06] claimed that in 2006 the lack of language skills had cost on average 325,000 Euros to a European SMB over three years.

To deal with this social and economic cost, research has been performed to speed up and improve the process of human translation. Today, there is a whole industry devoted to this issue. The language industry provides both human translation services and a wide range of software packages intended to bring translation costs down: translation memories, bilingual terminology-extraction and management software, localization software, etc.

This is the framework of research and development in computer-assisted translation (CAT) within which my doctoral research has taken place. This research was partially funded by Lingua et Machina³ – a company specializing in multilingual content management in a corporate environment, and by the ANR project Metricc,⁴ devoted to the leveraging of comparable corpora.

I.2. Motivation and goals

CAT has always used translation memories. This technique requires the translator to have a corpus of previous translations available, which the CAT software can use to generate bilingual lexicons, for example. This reality is problematic when the translator does not have such a corpus. This situation arises when the texts to be translated belong to an emerging field or to several languages for which few resources are available. To solve this issue, CAT research has looked into the leveraging of comparable corpora, i.e. a set of texts, in two or more languages, which deal with the same topic but are not translations of one another.

Comparable corpora have been the focus of academic research since the 1990s [FUN 95, RAP 99], and the existence of the Workshop on Building and Using Comparable Corpora (BUCC), organized every year since 2008 on the fringe of major conferences, shows the dynamism of this research topic.

The current research mainly aims at extracting aligned pairs of terms or sentences, which are then used in cross-lingual information retrieval (CLIR) systems [REN 03, CHI 04, LI 11] or in machine translation (MT) systems [RAU 09, CAR 12]. While CAT is often mentioned as a potential applicative field, the input of comparable corpora has not, to our knowledge, been genuinely studied within this application framework. Yet it presents several issues such as scaling or the adaptation to the needs of the final users.

³ http://www.lingua-et-machina.com,
This book had two primary objectives. The first objective is to assess the input of lexicons extracted from comparable corpora in the context of a specialized human translation task. Care has been taken to highlight the needs of translators and to understand how the comparable corpora can be best leveraged for CAT.

The second objective is to identify bilingual-lexicon-extraction methods, which best match the translators’ needs. Determining the current limits of these techniques and suggesting improvements is the focus of this research. We will focus, in particular, on the identification of fertile translations (cases in which the target term has more words than the source term), the management of multiple morphological structures and the ranking of candidate translations (the algorithms usually return several candidate translations for a single-source term).

The experiments are carried out in two language pairs (English–French and English–German) and on specialized texts dealing with breast cancer. This research has significant emphasis on applicability, and our methodological choices are guided by the needs of the final users.

I.3. Outline

This book is organized in two parts:

Part 1 presents the applicative and scientific context of the research. In Chapter 1, a historical overview of the beginning of MT is presented and we show how the focus of research efforts gradually turn toward CAT and the leveraging of comparable corpora. This book presents the current techniques to extract bilingual lexicons and detail the way in which the writer created the prototype of a CAT tool meant to leverage comparable corpora. Chapter 2 is devoted to the applicative assessment of this tool: we observe how the lexicons, thus extracted, enable translators to work more efficiently in their work. This assessment highlights the specific needs of human translation which are not dealt with in the classical techniques of term alignment. This is why this research took a different path, toward a different type of method, which aims to generate the translations of terms which can then be filtered using the corpus rather than to align terms that had been previously extracted from corpora. These techniques are described in Chapter 3. In this chapter, the focus is mainly on the so-called compositional approaches. Their limits are explored and Part 1 concludes with an indication of possible fruitful avenues for future research.

Part 2 of the book is given over to the efforts to improve compositional translation. Chapter 4 presents the methodological framework of the research: it describes the principle behind this approach, and attempts to highlight the contributions this work makes to compositional translation in terms of fertility, variety of the morphological structures processed and ranking of the candidate