U.PORTO
FACULDADE DE LETRAS
UNIVERSIDADE DO PORTO

MAKE IT SIMPLE WITH PARAPHRASES

•

AUTOMATED PARAPHRASING FOR AUTHORING AIDS AND MACHINE TRANSLATION

Anabela Marques Barreiro

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

2008

U.PORTO
FACULDADE DE LETRAS
UNIVERSIDADE DO PORTO

MAKE IT SIMPLE WITH PARAPHRASES

.

AUTOMATED PARAPHRASING FOR AUTHORING AIDS AND MACHINE TRANSLATION

Anabela Marques Barreiro

Advisors:

Professor Belinda Mary Harper Sousa Maia

Faculdade de Letras da Universidade do Porto

Professor Adam Meyers

New York University

2008

Words are the leaves of the tree of language, of which, if some fall away, a new succession takes their place.

– John French

In honor of my father Luiz

In memory of my brother Luís José

To my son Luís

Acknowledgments

The present study would not have been possible without the generous support of several people and institutions to whom I would like to express my gratitude.

I would like to thank my research advisors Professors Belinda Maia and Adam Meyers whom I genuinely respect and admire, for their guidance, insight, and support throughout this research.

I should also thank Professor Ralph Grishman for the opportunity to work as a visiting scholar in his department: the Department of Computer Science at the Courant Institute of Mathematical Sciences, at New York University, where I felt significant progress was made in my research. My gratitude also goes to the faculty members of the Proteus Project, Satoshi Sekine and Dan Melamed and to my colleagues especially, Cristina Mota, Heng Ji, and Yusuke Shinyama for their support, helpful discussions and congeniality.

I am grateful to Professor Max Silberztein and his department LASELDI at the Université de Franche-Comté, for the funded visits to Paris and his visits to Portugal, from which fruitful research collaboration has taken place, and in particular, for the funding to participate in the 2006 Colloquium "NooJ tools for Machine Translation: What is next?" at the the University of Paris 1-Panthéon-Sorbonne, and the NooJ international conferences, more precisely the 2007 and 2008 conferences in Barcelona and Budapest. I appreciate his patience in understanding and answering all my numerous e-mails with questions about NooJ and the challenging discussions about strategies for implementing the Portuguese linguistic resources.

Many thanks too to Diana Santos, my first mentor and friend for over 15 years, for the opportunity to participate in the doctoral seminars she organized and for reading and commenting on this dissertation and providing useful information to improve it. I am grateful to her and my Linguateca colleagues, especially Luís Costa and Luís Miguel Cabral, for making possible the enhancement and publication of this research's resources and for helping with the development of the interface for the authoring aid software tool ReEscreve.

Many thanks also go to my former colleagues at Logos Corporation, in particular, to Mike Dillinger, the Director of Linguistics back then, for reading this dissertation and providing relevant feedback. I thank him and Bud Scott, the father of the Logos system, for their support and encouragement in using OpenLogos resources, and to Andrei Telenkov for helping in the extraction of the lexica from the Oracle database.

I should also like to express my gratitude to:

- Professor Elisabete Ranchhod for the opportunity to work at the Laboratory of Language Engineering (LabEL), at the Instituto Superior Técnico - Technical University of Lisbon, where I had the chance of working with excellent researchers;

- Paula Carvalho, my friend and former colleague from Logos Corporation and LabEL, for teaching me the fundamental principles of the Lexicon-Grammar and for her comments and suggestions as I worked on this dissertation;
- Slim Mesfar, for the technical support on NooJ and for the encouragement when I was "fighting with my grammars";
- Leonor Davies, Allyson and Alexandra Barreiro, and the students of the Masters in Translation and Linguistic Services at the Faculdade de Letras of Porto University, namely Ágata Esmeriz, Ana Paula Mendes, Ana Polido, Andreia Martins, Isabel Coelho, Isabel Guedes, Ivone Oliveira, Luciana Peixoto, Madalena Ribeiro, Marco Costa, Natália Fernandes, Regina Ventura, and Sandrine Pires for helping with the machine translation assessment experiments.

My thanks also go to Professors and fellow researchers who in one way or another were receptive to my research work: Professor Philip Resnik, from University of Maryland, Walter Kasper, from DFKI, Professor Christiane Fellbaum from Princeton University, Professor Mário Silva from University of Lisbon, Professor Odile Piton from University of Paris 1, Professor Maria do Carmo Oliveira from PUC-Rio, Luís Sarmento from University of Porto, Sérgio Matos, from University of Aveiro, Professor Jorge Baptista from University of Algarve, Professor Nuno Mamede from INESC & Technical University of Lisbon, Tamás Váradi, from the Linguistics Institute of the Hungarian Academy of Sciences and Bernardo Recamán Santos from University Sergio Arboleda.

The research work presented in this dissertation was partly supported by PhD grant SFRH/BD/14076/2003 from *Fundação para a Ciência e a Tecnologia* (FCT), co-financed by POSI. This funding was essential to the concretization of the project presented here. I also need to thank FCT for the subsidy to support the dissertation printing costs.

***

On a more personal note, I would like to thank all my friends who were ever present: Alda Maria (my son's adopted grandmother) and her family, Ju, Fernando, Mariana and Tomás de Jesus for the amazing support and babysitting; Sonja and Pasquale Gangala for the treasured help and kindness in New Jersey; Beatriz Padilla and Pedro Correia for the encouraging conversations and generous help with Luís; Franco Colasuonno for keeping our business running smoothly in Swaziland; Pedro Bastos for helping with the language business; Jordina Guitart for letting me stay in her wonderful ski-rise apartment in New York City; Gursimranjot Singh Raipuria, "my NYC pal", for the companionship and fun moments; Rosa Bastos Pinto, Isabel Ilhorca, Helena Marques for

their support when Luís was born and all those who expressed their words of encouragement throughout the difficult moments, I truly appreciate their friendship.

A special thank you to Gordon Ashworth, my son's father, from whom I have learnt to turn adversity into something positive and who unconventionally taught me to see the strength and determination I did not know I had. His support and enthusiasm were indispensible to this achievement. He also helped turning my "Enguese" (English words and grammar with a flavor of Portuguese writing style) into native-like fluent and idiomatic English.

Last, but not least, I am extremely grateful to my family to whom I dedicate this work: my parents Luiz and Emília; my sister and brothers Luisa, Luís José, Vitor, and Henrique; my brother and sisters-in law José, Allyson and Maria do Carmo; my nieces and nephews Alexandra, Anabela, Ana Catarina, Pedro, Luís Filipe, João and Tomás, and to my little son, Luís. Special encouragement came from three of these extraordinary people: my father Luiz, a leading, visionary and inspirational figure, my brother Luís José who God called during this research to be my guiding star, and my son Luís the most precious gift I was offered during this research, who makes my work and my life fun and meaningful. He is the finest and most wonderful motivator for this effort.

Anabela

Lisbon, December 2008.

Resumo

Esta tese apresenta uma nova abordagem que permite melhorar a tradução automática baseada no parafraseamento de construções com verbos suporte. O desafio consistiu em parafrasear expressões nominais predicativas como *fazer uma análise*, em construções verbais como *analisar*, tirando partido das potencialidades parafrásticas da língua. Em casos particulares, o parafraseamento consistiu em substituir o verbo suporte da construção nominal, semanticamente fraco, por uma variante lexical ou estilística, por exemplo *realizar uma análise* ou *efectuar uma análise*. Quando as construções com verbos suporte foram identificadas e substituídas por verbos lexicais ou por outras expressões verbais semanticamente equivalentes ou próximas, numa fase de pré-processamento do texto, obteve-se aproximadamente 21% de melhoria na qualidade dos resultados avaliados da tradução automática do português para o inglês e aproximadamente 31% na dos resultados avaliados da tradução automática do inglês para o português. A investigação baseou-se numa análise linguística contrastiva, em que as construções com verbos suporte foram organizadas em várias subclasses sintáctico-semânticas de acordo com os princípios teóricos e metodológicos do Léxico-Gramática, estabelecidos no quadro da gramática transformacional harrissiana. Este estudo incidiu sobre um tipo particular de expressões multipalavra, as construções com verbos suporte, mas é extensível a outros tipos de expressões multipalavra, nomeadamente a expressões idiomáticas, tais como *dar o braço a torcer*, e construções sintáticas livres, tais como a coordenação de sintagmas nominais e a passiva. A informação linguística foi formalizada em dicionários e gramáticas desenvolvidos no ambiente linguístico NooJ e utilizados em várias tarefas de processamento de língua natural, sob o ponto de vista monolingue e bilingue. Os recursos bilingues português-inglês do sistema de processamento Port4NooJ, disponível em domínio público, integram a ontologia SAL do modelo OpenLogos e foram construídos como o alicerce deste estudo. As ferramentas de parafraseamento ReWriter e ParaMT foram criadas para reescrever e traduzir automaticamente as construções com verbos suporte e estão descritas nesta tese. O ReEscreve, versão portuguesa do ReWriter, está disponível na Internet como um serviço público de ajuda à autoria e a sua interface está descrita nesta tese. O parafraseamento automático de construções com verbos suporte através do ReEscreve permite melhorar em 40% os resultados da tradução automática no contexto em que estas aparecem.

Palavras-chave: construções com verbos suporte, expressões multipalavra, paráfrases, ferramentas de parafraseamento, tradução automática, sistemas de ajuda e sugestão à estrita, linguagem controlada, ReWriter, ReEscreve, ParaMT, DicTUM, Port4NooJ, dicionários electrónicos, gramáticas.

Abstract

This dissertation introduces a novel approach to improving machine translation by focusing on paraphrasing of support verb constructions. The challenge of the research was to paraphrase predicate nominal expressions such as *fazer uma análise* (*to do an analysis*) with predicate verbals, such as *analisar* (*to analyse*), applying language paraphrasing capabilities to produce better machine translation results. In particular cases, the paraphrasing consisted in replacing the semantically weak support verb of the predicate nominal construction with lexical-syntactic and stylistic variants, such as *realizar uma análise* or *efectuar uma análise* (*to perform an analysis*). When support verb constructions were identified and replaced with semantically equivalent or similar verbal expressions as a pre-processing step to translating, an average 21% improvement was observed in the evaluated quality of the results of Portuguese-English machine translation and, an average 31% improvement in the results of English-Portuguese machine translation. The research was based on a contrastive linguistic analysis of support verb constructions and of their paraphrases, which were organized in several syntactic-semantic subclasses according to the theoretical and methodological principles of the Lexicon-Grammar Theory, established in the Harrisian framework of transformational operator grammar. This study looked into one particular category of multiword expression, support verb construction, but it was designed to be repeatable and extensible to other types of multiword expression, namely to idiomatic expressions such as *dar o braço a torcer* (*to give up*) and to syntactically free constructions, such as noun phrase coordination or the passive voice. All linguistic information was formalized in dictionaries and grammars developed with the NooJ linguistic environment. This linguistic information was explored for several natural language processing tasks, from both a monolingual and a bilingual perspective. The Portuguese-English bilingual resources of the open source Port4NooJ natural language processing system were built as groundwork for the study. They integrate the SAL ontology of the OpenLogos system. Based on Port4NooJ, automated paraphrasing software tools ReWriter and ParaMT were also created to re-write and translate support verb constructions. ReEscreve, the Portuguese version of ReWriter, is being used as an authoring aid online public service and its interface is described in this dissertation. The automated paraphrasing of support verb constructions through ReEscreve allows a 40% improvement of the quality of the machine translation results in that context.

Keywords: support verb constructions, multiword expressions, paraphrases, paraphrasing software tools, machine translation, authoring aids, controlled language, ReWriter, ReEscreve, ParaMT, DicTUM, Port4NooJ, electronic dictionaries, grammars.

List of Symbols and Conventions

Symbols and conventions which were used in this dissertation are listed below.

= ‘identical’; strong similarity

≃ isomorphicity, approximate equality, congruence

≡ ‘is a paraphrase of’; similarity

Л stylistic or metaphorical equivalent

=> ‘translates into’

Ŧ transliteration

Examples

[ ] to the right side of an example, these parentheses indicate the source of the example. There are three types of sources:

1. examples extracted from corpora – they contain a mnemonic with the hyperlink to the URL where the example was found
2. examples extracted from other authors – they contain the reference
3. our own examples – they have no reference following them

Translation of Examples

Whenever appropriate, interlinear glosses (transliteration lines) are placed between the line of the original Portuguese example and the English correct translation. These transliteration or transcription lines help the reader follow the relationship between the original text in Portuguese and its English translation and understand the different structure of the Portuguese example being glossed. The translation in the interlinear glosses is a literal, word-for-word, translation that may not be coherent or grammatically correct in English. In such cases, they are marked with non-grammaticality (*) or questionable acceptability (?) signs. The objective is to show non-speakers of Portuguese to understand why machine translation may have problems with the translation of such expressions and why literal translation cannot be applied to such examples.

Table of Contents

List of Figures

List of Tables

PART ONE

Theoretical Background and Conceptual Problems

Chapter 1

Introduction

*

Chapter One presents a general overview of the main objective of this dissertation, which is to show how linguistic knowledge of paraphrases of support verb constructions can improve the quality of machine translation. It explains the motivation behind the research, defines the problem and its context, and describes the objectives and the scope of the research. A general preview of the research and a brief description of the structure of the dissertation are also presented in this chapter.

*

Delivering quality machine translation has eluded researchers for over fifty years. In response to the need for improved machine translation, developers have tried new methods to develop systems more quickly than using traditional methods. However, the same problem remains. Language is so complex that it is difficult to incorporate it into natural language processing programs, including machine translation. Machine translation models based on the transfer of linguistic knowledge use strategies analogous to the ones employed by human trained translators. This research discusses the classification and formalization of subsets of linguistic knowledge using empirical methods for lexical, syntactic and semantic description. It incorporates representations of linguistic knowledge into machine translation in search of better results. Enhanced dictionaries and computational grammars with syntactic-semantic and other linguistic knowledge have been created to support the model. With time and collaborative effort, these natural language resources will hopefully be integrated with machine translation systems, which will have substantial linguistic knowledge. We believe that these systems will perform qualitatively better than non-linguistic systems because they focus on precision criteria, such as the semantic relationships or syntactic properties of each individual linguistic unit.

The research described here is based on a system of observing phenomena to acquire and formalize a body of linguistic knowledge using linguistic analysis. Knowledge about

support verb constructions and their paraphrases has been collected and applied to machine translation systems to determine whether this knowledge improves translation quality. The **hypothesis** is that linguistic knowledge of paraphrases applied to a machine translation system improves the output quality. The **method** focused on the analysis and extraction of *support verb constructions* such as *fazer uma proposta* (*make a proposal*) from corpora, and the subsequent formalization of these constructions and their paraphrases within the framework of the Lexicon Grammar Theory [Gross, 1975] [Gross, 1981]. The **goal** was to apply the linguistic knowledge of support verb construction paraphrases to create the automated monolingual and bilingual/multilingual paraphrasing software tools, *ReWriter* and *ParaMT*. *ReWriter* is used as a standalone tool and *ParaMT* is for integration with machine translation systems.

1.1. **Motivation**

Before the Internet, machine translation was used to translate technical texts using specialized terminologies, such as user manuals for software products. These texts let companies make the best use of computer-assisted tools such as machine translation. Technical texts usually contain conventional language that proves to be simpler for machine translation to handle. Machine translation of this type was limited to a few applications and users were required to write in controlled technical language to obtain better quality machine translation. Controlled language helped reduce the time and the costs of post-editing.

The internet provided machine translation for the general public and it became a popular natural language processing application. Although translation of these texts is approximate (gist), the users find value in the results. This application has been accepted for more ephemeral text-types. Machine translation is driving internet communications between diverse populations of users with multilingual translation and editing services. It is being used to answer email in multiple languages, to support multilingual websites, to help people read and translate blogs, enroll in chats, etc.

Automated translation of varied texts requires more robust and linguistically sophisticated systems responsive to user expectations of quality and performance in the translation process. It is significantly more challenging to translate informal, non-standardized everyday language, because this language presents more clichés, idioms, set

expressions (frozen and semi-frozen), including certain types of support verb constructions. Informal language is more difficult to translate than the standardized, formal language with restricted lexicons or terminologies, where machine translation can work reasonably well. Linguistically enriched systems which handle multiword and idiomatic expressions correctly will enable the general public to communicate more freely and more understandably across different languages. A machine translation program that offers correct translation of support verb constructions, either via direct phrasal translation or via paraphrases demonstrates how applied linguistic knowledge helps improve output quality.

1.2. **Objectives**

One of the main objectives of this research is to show that experience with knowledge-based machine translation systems can be useful in adding linguistic quality and precision to machine translation in general. This research seeks improvement in machine translation output following integration with compatible linguistic resources. Despite difficulties associated with building such resources, positive results have been achieved from an understanding of our field of study. In line with the way that research goals are defined in [Dillinger, 2008], we have set ourselves the challenge of answering theoretical, empirical, methodological and practical questions, such as: where can we find a better understanding of the linguistic phenomenon? Which new data will be available to the research community for which purposes? Which research methods do we know better and are they suitable? Which practical problems can we approach and resolve/solve better, and how can we apply the linguistic knowledge to real applications? The answers for these questions range from the application of a theory that produces useful practical results based on previously objectively analysed empirical linguistic data, the creation of data-driven, non-algorithmic tools, procedural approaches to solving until now unsolved natural language processing and machine translation problems and identification of candidates for further study. These major objectives can be assisted by the complementary **sub-goals** of the research project that led to this dissertation. The sub-goals were: (i) assembling a bilingual list of support verb construction lexicalizations, such as *fazer uma apresentação de* (≠ *to make a presentation of*) or *to take a look* (*at*); (ii) creating rules and methodologies for finding short paraphrases for the support verb constructions; (iii) creating pooled bilingual lexical resources – by compiling sets of support verb constructions translations and paraphrasing

capabilities for the Portuguese-English language pair; (iv) building tools for automated extraction and generation of paraphrases; (v) creating a stable and repeatable methodology for anyone working on machine translation or bilingual lexicography; (vi) identification and building of English-Portuguese and Portuguese-English syntactic-semantic paraphrasing capability rules for the support verb constructions and study transfer mechanisms between the two languages; (vii) applying the results to machine translation; (viii) evaluating results.

1.3. **Scope**

The scope of this research is to show how to create support verb construction paraphrases automatically and how machine translation results can improve significantly with these paraphrases. Primary consequences of paraphrase knowledge include gains in language comprehensibility to non-native speakers and easier handling of "paraphrased/pre-edited" texts in natural language processing applications, especially machine translation, but also in more intelligent information extraction and retrieval.

1.4. **Preview**

The main challenge of this research was to paraphrase support verb constructions such as *fazer uma operação* (≠ *to make an operation*) with corresponding correct paraphrasing capabilities. The paraphrasing capabilities can be strong verbs, such as *operar* (*to operate on*) or lexical-syntactic and stylistic variants of the original support verb construction, such as *realizar uma operação* (*to perform an operation*) or *submeter-se/sujeitar-se a uma operação* = *ser operado* (*to undergo/have an operation* = *to be operated on*), depending on the context, and on the semantic classes of the arguments. The task involved steps such as identification of support verb contructions, identification of possible paraphrasing capabilities and formalization of the relation between these elements in the lexicon and syntactic-semantic grammars. Explicit marking of derivation and support verb and semantic verb association at the lexical level is the main characteristic of the method adopted. Identifying source language multiword expressions such as support verb constructions is not a trivial task, although it is the starting point for paraphrasal knowledge, as it is for translation.

As demonstrated *inter alia* by [Santos, 1988], [Santos, 1990] and [Santos, 1992], the suggestion of conceptually separating monolingual paraphrasing from translation in machine translation has been put forward by the insertion of a "style transfer" module which selects the "*best* or *chosen* translation" from multiple "*possible*" translations. The idea of dynamically invoking monolingual grammars to perform translation of multiword expressions was raised by developers on the working prototype built by the IBM-INESC Scientific Group back in the late eighties (ibidem). The approach of the present study uses monolingual grammars for the identification of support verb constructions and monolingual paraphrasing and bilingual/multilingual grammars for translation. Paraphrasing can be used in a monolingual text as a pre-editing procedure for more direct and controlled-language writing, or paraphrases can be generated and translated and inserted directly into Portuguese to English machine translation (and vice versa). The Portuguese-English bilingual resources of the open source, ontology-driven **Port4NooJ** system [Barreiro, 2008b], which was built as groundwork for the study, were employed. New automated paraphrasing software tools, such as the monolingual **ReWriter** [Barreiro, 2008c] and the bilingual/multilingual **ParaMT** [Barreiro, 2008a] were built and used to re-write and translate support verb constructions respectively. Both the monolingual and bilingual/multilingual automated paraphrasing tools, ReWriter and ParaMT are based on the NooJ linguistic environment [Silberztein, 2004]. ReWriter will be demonstrated with Portuguese data and ParaMT will be demonstrated with Portuguese to English data. The formalization and translation of support verb constructions are done through finite-state transducers (local grammars) for both monolingual and bilingual/multilingual purposes.

The theoretical framework adopted in this study is the **Lexicon-Grammar** [Gross, 1975] [Gross, 1981], which stands on the principles of the Harrisian transformational grammar [Harris, 1968], but the work is also based on the application of predicate-argument structure knowledge [Meyers et al., 2004a] [Meyers et al., 2004b]. According to the Lexicon-Grammar, simple or elementary sentences (predicate and its arguments), and not the individual words, represent basic syntactic-semantic units. These elements are entered into a dictionary. Each dictionary entry includes a description of the associated distributional and transformational properties. In this study, lexical, syntactic and semantic knowledge is added to these phrases. Support verb construction paraphrases are

constructed from a combination of enhanced linguistic (syntactic-semantic) resources with existing bilingual resources developed for Portuguese-English machine translation.

We consider that a detailed linguistic analysis of increasingly complex language leads to progress in language science and to fine-tuning of systems. This knowledge applies to machine translation, summarization or information retrieval and extraction systems and independently of the approach of any particular application. Effective results from such linguistically based research can save substantial effort and resources employed by statistically based machine translation systems. Such support verb construction paraphrases are also important for machine translation evaluation, because they provide the means to measure/compare the results of the systems automatically. The importance of paraphrases for automated machine translation evaluation has been confirmed by other authors: "models employing paraphrase-based features correlate better with human judgments than models based purely on existing automatic MT metrics" [Russo-Lassner et al., 2005: 1].

1.5. **Dissertation Structure**

This dissertation is organized in two main sections. Part ONE describes the theoretical background and conceptual problems from a linguistic point of view. Part TWO presents the application of the linguistic knowledge discussed in Part ONE to the development of paraphrasing software tools which use this linguistic knowledge. The state of the art in natural language processing applications is presented. Our resources, methodology and paraphrasing tools are described in detail. The assessment on support verb construction paraphrases and the evaluation of the new paraphrasing tools is also performed.

Part ONE presents the theoretical background and the conceptual problems that involve the object of the research. It contains six chapters.

Chapter One, the current chapter, introduces the research problem and describes the main issues that are addressed in the study. It defines the problem of achieving machine translation improvements by paraphrasing. It also presents the motivation, objectives, the scope of the research and a brief preview, and discusses resultant contributions.

Chapter Two focuses on the contrast between translation and paraphrasing. It presents different views of translation, familiar problems and models. It contrasts human with machine translation, pointing out the main difficulties and challenges translation presents

to machines. It explains what paraphrases are in general, and which types of paraphrases there are. It discusses which elements play a key role in paraphrasing and specifies the different granularities of meaning representation. And lastly, it stresses the importance of paraphrasing multiword expressions for machine translation.

Chapter Three describes multiword expressions. It discusses the importance of multiword expressions for natural language processing. It describes the properties that characterize multiword expressions: the properties of non-compositionality, non-substitutability, non-modifiability and non-translatability. It presents the major classes of multiword expressions, distinguishing between lexical units, frozen and semi-frozen expressions, idioms, phraseology, proverbs, collocations, metaphors, and lexical bundles. It ends up with comments on solutions to handle idiomaticity in translation.

Chapter Four defines support verb constructions and describes them in detail. It presents some theoretical background describing how support verb constructions were dealt with in distinct frameworks. Then, it discusses the syntactic and transformational properties of support verb constructions in the light of the Lexicon-Grammar Theory. Issues such as definition of role and classes of support verb constructions based on semantic content, definition of predicate nouns and nominalizations, comparison between predicate noun constructions and predicate adjective constructions, the role of determiners and prepositions, pre- and post-modifiers will be discussed, among others. Predicate-argument structure phenomena will be presented as an important linguistic concept in disambiguating support verb constructions for machine translation.

Chapter Five presents a typology of paraphrasing capabilities for support verb constructions. Support verb constructions are linked to predicates (verbs or verbal phrases) with equivalent or similar meaning. The usefulness of these paraphrasing capabilities will be discussed. Finally, the challenges that support verb constructions present to machine translation will be comprehensively described.

Part TWO of this dissertation is dedicated to natural language processing applications. The state of the art in machine translation and paraphrasing is presented as an introduction to this part. The remaining chapters describe in detail the empirical and practical work of our research, specifically, our resources, methodology and paraphrasing tools. Some evaluation is also performed. Part TWO contains five chapters.

Chapter Six presents the main approaches to machine translation: the grammar-based, example-based, statistical, and hybrid approaches and the three important methods for automatic acquisition of paraphrases: the corpora-based, statistical-based and dictionary-based methods. The importance of corpora to most of these approaches and methods is discussed. The chapter also presents some techniques related to the one described in our study. Finally, a sketch of the ideal machine translation is offered.

Chapter Seven presents the linguistic resources, tools and methodology, used in processing support verb constructions. It also shows how support verb construction paraphrases are generated and implemented. It first describes the original sources: OpenLogos and NooJ, and then the augmented linguistic resources, and goes on to demonstrate the new features. The dictionaries and grammars used to create paraphrases are illustrated and explained.

Chapter Eight describes the new resources and applications: Port4NooJ, DicTUM, ReWriter, and ParaMT. Port4NooJ is an ontology-based open source natural language processing system which includes bilingual resources for machine translation. DicTUM is a dictionary of multiword expressions. ReWriter and ParaMT are two new automated tools that recognize multiword expressions and generate paraphrases from them. ReWriter is an automated monolingual paraphraser designed as a writing aid, and ParaMT is an automated bilingual/multilingual paraphraser designed for machine translation. The interface for ReEscreve, the Portuguese version of ReWriter, is presented, together with its applicability to controlled versions of general or technical language and paraphrasing of extensive linguistic phenomena.

Chapter Nine is about evaluation. First, it describes some methods for evaluating machine translation. Then it presents some evaluation experiments and illustrates how support verb construction paraphrases can be used to improve machine translation. In addition, suitability parameters will be defined and used as part of the requirements in building automated paraphrasing software.

Chapter Ten describes the conclusions of the research presented in this dissertation. The main points will be summarized. The importance of paraphrases for natural language processing, and particularly for machine translation, will be emphasized in a list of benefits we found most important. Finally, we present a synopsis of future research work that the current study can sustain.

Chapter 2

Translation and Paraphrasing

*

Chapter Two establishes the parallelism between translation and paraphrasing. First it exposes misconceptions about translation and presents distinct perceptions about the nature of translation. It also contrasts the characteristics and purposes of machine translation with those of human translation, focusing on the major obstacles and challenges for successful machine translation. Then, it introduces paraphrases in *sensu lato*, explains the meaning of paraphrases in linguistics by reviewing the relevant literature, and describes the importance of paraphrasing in translation. The basic elements of paraphrasing are presented and the different levels of granularity for semantic representation in paraphrasing demonstrated; namely referential, lexical, collocational and syntactic paraphrasing. The chapter ends by stressing the importance of paraphrasing multiword expressions.

*

In the past, translation and paraphrasing were considered separate subjects in the field of natural language processing. [Callison-Burch, 2007: 1] affirms: "Paraphrasing and translation have previously been treated as unconnected natural language processing tasks. Whereas translation represents the preservation of meaning when an idea is rendered in the words of a different language, paraphrasing represents the preservation of meaning when an idea is expressed using different words in the same language". In effect, one could say that paraphrasing is at the level of one language and translation is at the level of two languages or, in other words, that translation is the paraphrasing of meaning in one language applied to another language. The same kind of problems and linguistic phenomena take place in paraphrasing and in translation.

Even though we use a different approach and methodology from the aforementioned author, one common goal is to improve machine translation based on paraphrasal

knowledge, and thus prove that both tasks are allied and cannot be detached. Translation and paraphrasing complement each other in improving machine translation output quality and enhancing natural language processing.

Before pointing out the major topics entailed in paraphrasing, we need to discuss the nature of translation. Translation has been misunderstood by non-experts in the field, the nature of translation has given rise to divergent opinions even among experts, and a clear distinction needs to be made between human and machine translation. After approaching these problems, paraphrasing is defined and studied in detail. A few questions are raised for which answers will be attempted: to what extent can a sentence conveying the same idea be considered a paraphrase? What makes two sentences a paraphrase of each other? Is it enough to change a word in a sentence to generate a paraphrase?

2.1. **Translation Misunderstood**

When studying translation, an objective and purpose distinction must be established between human translation and machine translation. Even though this dissertation is about improving machine translation with linguistic knowledge, some topics discussed in the theory of human translation seem relevant to the study of translation in general.

The aptitude to translate and the process of translation are complex and not always seem to be understood by the general public. There are two common translation fallacies. First, any person who speaks two languages is able to translate easily between these two languages. Second, translation is a mechanical process and the correspondence between two languages can be done merely word-for-word (*literal translation* or *metaphrase*).

As to the first myth, any person who speaks two or more languages well can only consider him/herself a candidate for the profession of translator. There are several skills that have to be learned for the profession that often can only address particular niche topics, and not work with the full scope of two languages in contrast. With the development of theories of linguistics and applied linguistics, translation was seen as a specialized and licensed activity performed by professional translators. Professional translation requires a profound knowledge of the source language as well as native proficiency of the target language. Additionally, it requires above-average writing skills, and an insightful knowledge of the social-cultural aspects of the source and target

languages. Because total fluency is uncommon among even highly skilled second-language learners, translation experts not only master completely the linguistic spectrum of two idioms in contrast, but they need to be native speakers of, at least, the target language [Kasparek, 1983]. Translators translating into a second-language are prone to the influence of their native languages' structures, idioms, etc. even if they are very skillful. Translation also implies knowledge of the grammar of the two languages, their writing conventions, and the situational and cultural context. In the case of scientific and technical translation, subject matter knowledge is required, including terminologies of the field or knowledge domain.

With regard to the second myth, the translation community has already made clear that literal translation is not acceptable [Maia, 2007]. Attempts to translate literally fail whenever the translator is confronted with expressions that have no correspondence in the target language. A literal translation is impractical because words can be ambiguous and because languages have different grammars, different lexicons, and different ways of expression, idioms, cultural background, etc. Polysemic words carry different meanings depending on context. For example, the word *sound* can mean *tone*, *body of water*, or *medical probe* (examples from [Scott, 2003]), depending on the context in which the word appears. On the other hand, one single meaning can be conveyed using distinct words or phrases. For example, the concept of *creation* can be expressed in words such as *formation*, *making*, *conception*, *construction*, *design*, among many others that can be found in any English dictionary. Different scopes of meaning represent different choices and possibilities for translation of a certain word, expression, sentence, or passage.

Different possibilities of translation result from **dynamic equivalence**. The term dynamic equivalence was introduced by [Nida, 1969] as opposing to formal equivalence. His use of the expression **formal equivalence** or a translation that is as close to the source text as possible. Formal equivalence translations help understand how meaning was expressed in the source language text. They try to show the original aesthetics, the beauty of original idioms, figures of speech such as metaphorical or symbolic patterns, allegories, etc., and also help perceive the author's style and use of unique vocabulary terms because they preserve similar grammar, style, voice, or order, features that are related to form (syntax). Nida's definition of dynamic equivalence moved towards the idea of the functionalists, such as [Reiss & Vermeer 1991] and [Nord, 1997]. He applied it

essentially to Bible translation and adapting the Bible (or not) to the local culture. For example, the image of *Lamb of God* was replaced with *Seal of God* (translated *kotik*, for *young seal*) for the Eskimos because "there are no lambs trotting on the ice meadows of the Arctic where the Eskimos live" [Barnstone, 1993: 41]. Bible translators have always argued about translating '*the word of God*' (as literally as possible) or doing an 'interpretative' version that attracts the public it is aimed at. For example, some of the modern translations aimed at attracting a younger generation and they are selective in their use of words and style. Functionalists normally start at the level of the text and only work down to the word level gradually. This preference is related to the fact that functionalists are always more interested in analyzing the pragmatic, social or communicative aspects of the text, and the reason why it is being translated. Functional equivalence is a top-down approach to translation where the word level comes at the end of the analysis. Functional translations distance themselves from the original form, and emphasize understandability of the original meaning, and readability. In functional equivalence translation, the translator attempts to achieve a target language text that functions the same way the source language text functioned for the original readers. Nowadays, more modern culturally orientated theories deal with the concept of equivalence in translation in different terms. [Venuti, 2004] and [Venuti, 2008] claim that a translator should not simplify in favour of making the translation fluent or easier to read but should make the reader appreciate the 'foreignness' of the text and accept a different cultural way of looking at the world.

Theory of translation has been dealing with controversial issues such as problems related to privileging meaning over form, visibility or invisibility of the translator, being faithful to the author or trying to make the text accessible to the reader (and which kind of reader), giving value to the source language culture (foreignise) or making the text suitable for the target language culture (domesticate) or allowing a 'big' language / culture to predominate over a 'small' language / culture or being creative, etc.

Because the nature of machine translation is different from that of human translation, many of the problems discussed in the theory of translation are not applicable to machine translation. An important common issue in human and machine translation is to define equivalence and to define and establish **paraphrasing capabilities** [Polguère, 2000] [Yamamoto, 2004]. Equivalence relates to the success of paraphrasing. It can be

evaluated by humans without recourse to linguistics, but establishing an equivalent by a human is not a repeatable process, but a random, sometimes arbitrary and spontaneous practice. Each time a human paraphrases, a different paraphrasing strategy may be adopted and be more or less effective. Paraphrasing capabilities concern the resources and can become repeatable. Paraphrases obtained without human intervention for a specific phrase or sentence can be used by machine translation systems consistently, in one or varied combinations. Linguistics offers the resources and the ability to acquire knowledge concerning these paraphrases so that they can be used, either by translators for stylistic reasons, or by machine translation for simplifying meaning. In both cases, the goal is to get close to equivalence. So, paraphrasing capabilities are used to set a standard against which both human and machine translation efforts can be measured.

Traditionally, a translation equivalent is the target language expression, which corresponds to the source language expression. Essentially, equivalence concerns translation units. As in [Teubert, 2001: 145], "the equivalence of a translation unit in the target language is called a translation equivalent" and "the translation equivalent is regarded as the "paraphrase" of the meaning of a translation unit but in the target language". A translation unit is a segment of a text that represents a single cognitive unit when establishing equivalence. It can be a single word, a multiword expression, a sentence, or even a larger unit. An idiomatic or non-literal translation comprises larger translation units. A translation unit is different from a unit of meaning [Sinclair, 1996]. A unit of meaning is a meaningful linguistic unit and may be as small as a morpheme. For example, the plural morpheme *-s* is a unit of meaning but it is not a translation unit. Other units of meaning may be expressions that cut across traditional constituency boundaries, such as lexical bundles, viz. *dizem que* (*they say that*) or support verb constructions, viz. *ter uma conversa com* (*to have a conversation with*).

2.2. **On the Nature of Translation**

Language is very complex and the same idea can be expressed in more than one way. With translation, that flexibility of language is tested incessantly. In the words of [Newmark, 1988: 9], "What translation theory does is, first to identify and define a translation problem; second, to indicate all the factors that have to be taken into account in solving the problem; third, to list all the possible translation procedures; finally, to

recommend the most suitable translation procedure, plus the appropriate translation". It is difficult to conceive translation as a science, because it lacks the precision and predictability that characterizes other sciences. As pointed out by [Azizinezhad, 2006], "it is impossible to devise a scientific equation that would work in the same way, every time, for each problem in all languages due to the inescapable differences among languages as well as their cultural contexts throughout the world".

The lack of consensus on the nature of translation is related to different translation methods and the meaning conveyed in each particular type of text. The most relevant aspect in translation is to define the purpose of each translation, which is related to the characteristics of each text. For example, a certain subjectivity and distance from the source language text is allowed in translation of literary text for the sake of maintaining the artistic and aesthetic aspects of the target language text [Hermans, 1985] [Landers, 2001]. Literary translation may be considered an art [Leighton, 1990] [Weaver, 2002], where the translator has more freedom of expression. On the contrary, technical, commercial, and legal translators, like the authors of the original texts, are more restrained in their use of language, and they need to be precise and convey the exact meaning of the original text. Technical texts are not meant to be beautiful but rather to be informative, instructive and explanatory. Their main function is to be clear, so the easier they are to read, the better they are understood. Technical translation may be regarded as a craft [Newmark, 1988] [Biguenet & Schulte, 1989] for which both technical and linguistic competence is essential, but creativity and vagueness prohibited.

With more translation being performed by machines, new challenges are imposed on the field, theoretical traditions shaken and the need to rethink the status of translation becomes more evident. Of all automated applications, machine translation compels us to reconsider the nature of translation. Art and craft are not appropriate concepts for machine translation because it has necessarily to rely on linguistics and computer science.

2.3. **Machine Translation *versus* Human Translation**

For most of human history, translation was an exclusively human activity. Translation is as ancient as written literature. Historiography documents several translations of ancient literary texts. As interest in different languages developed, so did translation philosophy, its theory and its methods. With the appearance of sophisticated computers and the

continuous development of computer software capabilities, machine translation has become available to the general public in the last few years. Before that, machine translation was only accessible to a very restricted niche of the market, and computer-aided translation was used only by professional translators. Since the internet boom, the world of translation has changed. Machine translation has become widely used and it often replaces human translation when the client requires gisting.

Machine translation is the automated translation of text or speech from one natural language into another. It is an important tool that assists human translators. However, computational linguists have not yet discovered how to formally model or represent with success the linguistic information that human translators can convey without any mathematical formalization. As expressed decades ago in [Bar-Hillel, 1959], some still believe that general-purpose fully-automatic high quality machine translation is unfeasible. Formalizing language and programming computers to use these formalizations efficiently is a difficult process and it may take more time that non-skepticals initially thought. Dynamic and ever changing language remains at the center of scientific investigations, but because it is so complex, there has been no unified way of studying and modeling it in a way that permits machines to use language more efficiently. Various attempts have been made by linguists to systematize language, but none has been able to fully succeed in natural language processing. Chomsky's proposed universals turned out to be not directly applicable to translatability and language formalization cannot be reduced to a set of "mechanical" rules as proposed in [Chomsky, 1980]. In Chomsky's own words: "The existence of deep seated formal universals implies that all languages are cut to the same pattern, but does not imply that there is any point by point correspondence between particular languages. It does not, for example, imply that there must be some reasonable procedure for translating between languages" [Chomsky, 1965: 30] (…) "by a "reasonable procedure" I mean one that does not involve extra-linguistic information. That is one that does not incorporate an "encyclopedia") [Chomsky, 1965: 202; footnote 17].

Theories and linguistic schools which recognize a vital connection between lexicon, syntax and other language levels are now generally considered to provide more satisfactory reflections on how real language works. A middle ground solution does not exclude the possibility of formalizing the underlying sharing of common features among

languages so that they can be translated among themselves. Efforts in language systematization, such as those of Harris and his followers, helped develop the study of the structure of different languages and the efforts of all linguists working at systematization, whether morphological syntactic, or semantic, helps machine translation by adding to the pool of useable linguistic knowledge. However, language always escapes from total classification, due to its creativity and dynamics and because humans usually communicate in a context of shared information where social and cultural aspects cannot be filtered out. Also, many of the studies in philosophy, psychology, neurology and linguistics over the last couple of decades have demonstrated that human beings do not always behave 'logically' and that 'emotion' plays a major part in our behavior and in our ways of expression. This is one reason why it is so difficult to program computers to use language as we do. Machine translation faces different barriers from human translators. Machine translation cannot grasp humour, sarcasm, and other human feelings expressed in sophisticated linguistic expression, while human translators use ingenuity and skill to artfully reproduce feeling and sound. Until now, researchers have not been successful at producing machine translation capable of translating certain types of text, resolving anaphora, and handling other extra-sentential and extra-textual information. Human beings can easily and intuitively retrieve information from even distant parts of the text, or from extra-textual knowledge. This is the big advantage human translators have over machines and one of the reasons why human and machine translation do not compete. Extra-linguistic factors make machine translation different from and more difficult than human translation. Human translation can easily solve most ambiguity problems as well as problems of culture and context that machine translation cannot. Human use of language assumes knowledge of the world or user-centric particular worlds, something that is difficult or impossible to program in a machine. Machine translation functions best in a situation where the writer cannot assume shared knowledge of the world. Domain specific contexts and controlled language make machine translation reliable, proving the advantages of a scientific approach to language. Human translation cannot be replaced by machine translation, at least until there are breakthroughs in the limitation of machine translation to sentence level translation, and in artificial intelligence.

Computational linguistics is a young field and applications such as data mining and information extraction are developing and employing recognition and tagging of named

entities [Nadeau & Sekine, 2007] [Mota, 2008]; the semantic web [Berners-Lee et al., 2001] [Antoniou & van Harmelen, 2004] [Allemang & Hendler, 2008] and automatically generated semantic 'clouds', also known by the metaphorically-based term 'cloud computing' [Brandel, 2008], are still experimental and 'stormy'. The various 'manually' prepared Wordnets have to be language specific and even Framenet [Fillmore et al., 2002] is limited by the lexicon. Machine translation deals with all these questions and that makes it a much more complex natural language processing application. As the subtasks become more complete, more can be done, better, by machines. The task of delivering high-quality machine translation of certain types of texts and complex linguistic phenomena is difficult, but not impossible.

2.3.1. **Complexity and Ambiguity of Natural Language**

Despite the availability of funding and many talented researchers worldwide, most efforts to build cost-effective, industrial strength, high-quality machine translation have fallen short of their goals, since first attempts in the 1950's. Successful machine translation has been difficult to achieve because of two major hurdles: complexity and ambiguity of natural language. Translation in general is a very difficult task because language is very complex. Every sentence can be expressed in a variety of lexical and syntactic ways, depending often on what each individual has absorbed and processed as suitable for the occasion from previous experience, and according to contextual and cultural associations that each individual is more or less aware of, and processes in a less-than-linear fashion. Different human translators can produce a wide variety of different and equally acceptable or good translations (different target sentences) from the same text by using different words or synonyms, different word order and word combinations, different stylistic choices, etc. This flexibility and appropriateness is difficult to achieve in machine translation. Current results show that machine translation has difficulties in producing even one translation which respects entirely the meaning and grammaticality of the source text. This is mainly because natural language is full of ambiguous words and structures that frequently defy computer analysis. Researchers have not found a way to program computers to handle certain types of ambiguity. Any linguistic unit, word, expression or sentence that has more than one possible interpretation is ambiguous. Ambiguity can pertain both to grammatical function (PoS ambiguity) and to meaning

(semantics). [Scott, 2003] presents examples of ambiguities operating at several levels: the lexical syntactic level, the sentential syntactic level, the lexical semantic level, the sentential semantic level and the extra-sentential level. Humans are not even aware of most ambiguities when they are using language because they are conscious about the context. But it is difficult to program the computer in a way that is easy to determine which sense of a word is triggered by the use of that word in a particular context. When a machine translation system deals with the multiple layers of meaning with which it is confronted, the opportunity for error is high and a single incorrect decision can damage the machine's parsing of the sentence, causing unexpected results in the translation.

2.3.2. **Challenges to Machine Translation**

Challenges to machine translation engines pointed out by [Barreiro & Ranchhod, 2005] and [Maia & Barreiro, 2007], stress the importance of linguistic formalizations in order to improve translation quality in areas which statistical machine translation systems find difficult to handle. A few critical areas have been identified and we will present them and others in the next few paragraphs.

Homographs represent remarkable challenges to part of speech disambiguation, because they often trigger the wrong analysis. Since machine translation systems lack interpretation capabilities, error probability is significant in these cases. If a part of speech is not analyzed in the correct way in the source language, it leads to an incorrect resolution and consequent change of the part of speech in the target language. Consequently, sequences of words without a coherent syntactic structure might be generated, resulting in translation errors, and therefore, in meaningless translations. This is still a problem for many current systems.

Many systems still lack the ability to deal satisfactorily with commonly recognized differences between source and target languages (cross-language phenomena as defined in the FEMTI guidelines), such as lexical divergences and idioms and cross-language syntactic transformations, such as passives. Idiomacity, improper identification of named entities, insufficient dictionaries and grammars to recognize and translate multiword expressions, and incapacity to deal with long sentences, are factors that often cause errors in machine translation. Unusual alterations to the order of words in the target language are also a common problem in machine translation.

It is beyond the current capability of machine translation systems to process language at the discourse level. More challenging machine translation grey areas are anaphora resolution, common-noun nuance resolution, and the handling of ellipsis, which constitute a class of more advanced ambiguity problems. To solve them, analysis must go beyond the sentence level to the extra-sentential (discourse) level. They relate to referential associations of utterances in neighboring clauses or the text. The computer simulates human processing of sentences by capturing referential associations. In (1) there is a referential connection between the noun *países* (*countries*) and the indefinite pronoun *nenhum* (*any*).

(1) O João visitou muitos **países** do mundo. A Maria não visitou **nenhum**.
=> *João has visited many **countries** in the world. Maria hasn't visited **any**.*

The ambiguity in (1) is more difficult for the machine than for a human (native) speaker. While for the human mind it is clear that the anaphor "*nenhum*" has as its antecedent the word *países* and not *mundo* (*world*), this connection has to be formalized and transferred to the machine.

In Portuguese, the words *primos* and *inteiros* in (2) are ambiguous. In the context, they are predicate adjectives, qualifying numbers, viz. *números primos* (*prime numbers*) and *números inteiros* (*whole numbers*) where the noun in the noun phrase is elided (see [Carvalho, 2007] on noun elided constructions). In the absence of the noun, the adjective becomes the head of the noun phrase. But the word *primos* is also ambiguous regarding part of speech. It can also be a noun, meaning *cousins*. Without context knowledge, the machine does not know which translation should assign to that word, and most likely will assign the wrong one, *cousins*, because it "sees" it as a noun instead of an adjective. On the other hand, *inteiros* is an adjective with several meanings (polysemic). Without knowledge of the missing noun, the machine has to decide among the meanings which one to assign it and translate it accordingly. The choice in that case is purely arbitrary.

(2) São chamados de **primos** os [números] **inteiros** diferentes de 1 que só são divisíveis por 1 e por si próprios. [Escolinha]
Ŧ *Are called **primes**, the **whole** [numbers] that are different from 1 and divisible only by 1 and themselves.

=> A prime number is any whole number which is different from 1 and divisible only by itself and 1.

The question is insoluble without reference to the contextual clue given extra-sententially, or to clues presented on discourse analysis, the presence of revealing neighbouring words, such as the number *1* or the adjective *divisíveis* (*divisible*). To resolve the ambiguity raised by the elided term *números* (*numbers*) in example (2), analysis would benefit from the access to paraphrasal knowledge. Paraphrasing helps create semantic associations that otherwise would be extremely complex for the computer to make, especially if the contextual clues are remote.

The complexity of language reflects the complexity of our own minds, so it is difficult to have as the instrument of investigation, the object of that same investigation. Clearly, difficult issues constantly arise in the study of language, and linguistics provides a key element of repeatable science to translation. Our paraphrasing tools show that reliable linguistic knowledge can be used to improve text quality and drive it closer to the target language text. Although this study is limited to one linguistic phenomenon, it is designed in such a way as to be a useful resource according to the forward-thinking design of the resource objects, such as the Nooj linguistic environment, that we employ in this research. Building meaningful resource objects that support machine translation is leading to increasingly powerful natural language processing tools in the hands of end-users and usable by machines.

Successful machine translation requires a complete and systematic analysis of the contrasts between source and target languages at all levels. If correspondence relationships and linguistic representations of both languages are established correctly, more satisfactory results will be produced. The essential key is to build complete and integrated linguistic intelligence into these systems. We believe that most of the problems mentioned above have solutions. Larger and better linguistic resources, a more conscious use of the written word, better quality texts as input to the systems and paraphrasing techniques are some of the ways to handle immediate problems.

2.4. **Paraphrase *Sensu Lato***

The etymology of the word 'paraphrase' confers some understanding to its meaning. The roots of the word are found in online dictionaries, such as Merriam-Webster OnLine, MSN Encarta and Dictionary.com. According to these sources, the word appeared in 1548 Middle French texts. It derived from the Latin *paraphrasis* (a paraphrase), originated from the Greek *paraphrazein*, which means *to tell in other words*, from the prefix *para-* (beside) + *phrazein* (*to tell*; *to point out*).

In a general sense, paraphrasing means rephrasing or rewording. It is a technique that consists of changing the words (the form, syntax) of a text, yet preserving its meaning (the semantics). Language is so vast that there are many ways of conveying the same meaning. Humans use various linguistic strategies when expressing their knowledge of the world or of real-world situations. Paraphrasing is a valuable skill and paraphrases are often sought when searching for different ways to articulate the same meaning or idea. The use of paraphrases relates to the difference between people. Different social, cultural and psychological characteristics are expressed according to the individual. Paraphrasing facilitates comprehension by providing alternative expressions that might be more easily understood by each individual.

Paraphrases fulfill several functions and are used for explanatory, interpretational or simplifying purposes, depending on the needs of the individual. Simplification of concepts and ideas, as a means of communicating clearly, is a prevalent reason for paraphrasing, and it is used as an exercise in writing classes. Paraphrases are also used to improve clarity, to establish a certain style or to avoid repetition. The use of paraphrases to make a text less wordy or more concise is a common practice in controlled writing, where the technical writer deliberately chooses words that express a more succinct and clear meaning. It is also common to paraphrase technical terms in words that are easier to understand to an unfamiliar user. For example, in ophthalmology, the term *periorbital infection* means an '*infection of the dermis and associated tissues around the eyes or an infection of the lacrimal system*'. Thus, the adjective *periorbital* is paraphrased by the expression *around the eye*, so that a non-specialist may understand the term. In addition, an identical message can be conveyed in different styles, more or less formal, straightforward, metaphorical, etc. Paraphrasing is testimony to the richness of language and the opportunities to make texts that have more value to the audience.

2.5. **Paraphrase in Linguistics**

The paraphrase is the central issue to many linguistic theories, which regard it both as an integral part of a speaker's linguistic competence and a central tool for language modeling. Some linguistic trends became particularly focused on the study of paraphrases. A special focus has been given to paraphrases in generative semantics [McCawley, 1976a], and in systemic linguistics [Halliday, 1985c], [Hasan, 1987], [Cross 1992]. Paraphrasing was also an appealing topic for some academics of the School of Prague [Sgall et al., 1986]. Individual scholars such as Fuchs [Fuchs, 1982], [Fuchs, 1985], [Fuchs, 1987] and [Fuchs, 1994]; Martin [Martin, 1976] and [Martin, 1983]; and Culioli [Culioli, 1990], have also developed interesting approaches. A detailed diachronic overview of works related to, or on paraphrasing, can be found in [Fuchs 1982: 109-227]. [Milićević, 2007a] presents succinctly the main approaches on paraphrase, from Harris until now, and answers interesting questions on paraphrase modeling.

The Meaning-Text Theory [Zholkovskij & Mel'čuk, 1965] brings probably the most important contribution and pertinent notions for the study of paraphrases. According to this theory, language is represented as a "paraphrasing system". This paraphrasing system "operates on the principle that language consists of a mapping from the content or meaning (semantics) of an utterance to its form or text (phonetics). Intermediate between these poles are additional levels of representation" [Beck, 1997]. The correspondence between levels is made by means of bi-directional rules that translate one level into the next level. They are interpretive devices establishing equivalencies between the symbolic conventions of one level and those of another, and they can operate in both the direction of meaning to text or in the direction of text to meaning.

Many existing systems use paraphrases based on the Meaning-Text Theory, such as [Boyer & Lapalme, 1985], [Iordanskaja & Polguère, 1988], [Iordanskaja et al., 1991], [Iordanskaja et al., 1996], [Nasr, 1996], [Lareau, 2002]. [Milićević, 2007a], establishes a typology of paraphrases based on lexical-syntactic rules ('*système de paraphrasage lexico-syntaxique*'). [Milićević, 2007b] proposes new paraphrasing rules, which establish equivalences between "*les fragments des représentations sémantiques des énoncés*", i.e., fragments operating at the semantic level of representation of "utterances". These fragments represent a novelty in relation to previous theoretical concepts. They relate

the propositional aspect of meaning with its communicative and rhetorical aspects in paraphrase production, relating style and the flexible concept of paraphrases that takes into account an approximate character (*paraphrases approximatives*). They make the notion of paraphrase broader, covering not only quasi-synonym phrases, but also other linguistic expressions, from one language (*paraphrases intralinguistiques*), or from different languages (*paraphrases interlinguistiques*).

Although it is not the goal of this dissertation to present a detailed description of all these works or of their various results, the most relevant ones to the present study will be discussed. The Lexicon-Grammar Theory framework was tested and found to offer, according to the particular needs of this research, the most adequate treatment of the phenomena involved in the type of paraphrases studied here. We reiterate that this substrate of paraphrase is very specific and partial and the results cannot be extrapolated to discourse paraphrase. However, the additional theoretical background provided in Chapter 4 shows that this methodology is suitable for the study of support verb construction paraphrases in particular and can be extended to other types of multiword expression, namely to idiomatic expressions such as *dar o braço a torcer* (*to give up*) and to syntactically free constructions, such as noun phrase coordination or the passive (phrase and sentence-level paraphrasing), as will be explained in § 8.5.1.

The creation of a model for paraphrasing is not a simple task. There are many problems and difficulties associated with the analysis of them. Paraphrases are produced by complex phenomena and multiple factors that determine the result, even for relatively simple phrases such as support verb constructions. Some of the problems of establishing a typology will be discussed in Chapter 5 and the general architecture of the paraphraser will be drafted in Chapter 8.

2.6. **Paraphrase and Translation**

Paraphrases are important within the same language, but they are critically important in translation from one language to another. Respect for the delivered word is a foundation of civilized society and translators seek first to understand the text before rendering a translation. However, this understanding of the text often implies rewording it so that it is more likely to be understood in the target language. Different languages use different ways of expression, and present different syntactic-semantic behavior. Some languages

are more alike than others, some have a simple grammar, and others have more complex grammatical patterns. In addition, words and expressions can be ambiguous and may have more than one synonym (cf. § 2.7), and therefore, more than one translation.

Synonyms are the primary source of paraphrases. Translation implies cross-language synonymy. Cross-language synonyms are words that mean the same thing in the source and target languages. For example, the English time-related noun *period*, can be translated by the synonymous nouns *período* or *época* in Portuguese. However, a synonym may not always exist in the target language and a literal translation is not always possible or desired. The source language may contain words or expressions for concepts that have no direct equivalent in the target language. In these cases, translation has to be more dynamic and less attached to the original. This is related to the referential problems that will be presented in § 2.8.1. Even though this is a study of paraphrases from a translation point of view, it is relevant to understand paraphrases in general, discuss their usefulness in natural language processing and then focus particularly on machine translation. The following paragraphs discuss the basic elements that come into play when humans paraphrase, and distinguish the main levels of granularity of meaning representation in paraphrasing.

2.7. **Elements Playing an Important Role in Paraphrasing**

There are several basic "*elements*" in creating paraphrases: synonyms, near-synonyms, connotation, figurative meaning, and metaphors. They all play an important role in paraphrasing. Synonyms are probably the most straightforward source of paraphrase. They are used to create more varied and fluent text. Synonyms are found in dictionaries and thesauri, and are identified and characterized with regard to certain senses of a word. A sense of the word is contained in the meaning of that word, which is also defined (by degree) by the use of that word in context. Words are interchangeable to some extent. The interchangeable character of words gives rise to paraphrases. For example, the English words *wage*, *salary*, *income* [examples from Haas, 2000: 6] and *pay* can be used almost interchangeably to express a form of periodic payment established between an employer and an employee. So, in this specific context of usage they are all synonyms and identical in meaning. That does not mean that they have exactly the same meaning in all contexts or social levels of language. Several linguistic aspects make them unique:

etymology, orthography, phonetics, ambiguous meanings, usage, etc. Also, different words that are similar in meaning often differ for reasons such as style, formality or politeness, alternative usage, doublespeak, euphemism, etc. For example, the Portuguese word *invisual* (*visually impaired*) is more formal and politically correct than the word *cego* (*blind*); the adjectives *usado* (*used*) and *em segunda mão* (*second-hand*) are only synonyms in certain usages and not in others, such as in *veículo usado* (*used vehicle*) or *roupas usadas* (*used clothes*) and *veículo em segunda mão* (*second-hand vehicle*) ou *roupas em segunda mão* (*second-hand clothes*). The adjective *usado* in *solas usadas* (∓ **used soles*) means *worn-out*.

Some words may not be synonyms when appearing by themselves, but are synonyms when combined with other words. In both general language and specific or domain contexts, certain compounds are made up of distinct words, which in the context mean the same. For example, in the technical domain, the adjectives *health* and *medical* appear in synonym compounds. *Health care* is equivalent to *medical care* and *health insurance* is equivalent to *medical insurance* (examples from [Haas, 2000: 6]). These are examples of different names for the same concept - they mean exactly the same thing. There are also instances of phrases from general language mapping to a single word. The adverb phrase *in a successful way* can map to the single word adverb *successfully*. Support verb constructions are also examples of phrases that can be mapped to a single word. For example, the support verb construction *to make a presentation of* can be mapped into the single verb *to present*. In other cases, the same exact word can mean different things in different terminologies or scientific language domains. For example, in business, the word *client* means *costumer*, but in computer software terminology, it means the distributed part of the program that runs on the end-user's machine.

Although there are indeed many instances of fairly exact synonyms, there are also many examples in the dictionaries and thesauri of "*near-synonyms*" or "*quasi-synonyms*", which are often used in paraphrasing. The optimal paraphrase is the closest possible in meaning to the original sentence, but sometimes paraphrasing is difficult to achieve without resorting to the strategy of including near-synonyms. A near-synonym is a word or phrase that means not exactly the same as another, but almost the same [Ullmann, 1962] [Cruse, 1986] [Lyons, 1995] [Edmonds, 1999] and [Edmonds & Hirst 2002]. In near synonymy, "the difference may be one of tone or register (*chick* and *girl*), one may be

slightly broader or narrower in meaning (*nurse* and *RN* or *elderly* and *older American*) or the degree of synonymy may be dependent on context or usage (*retired* and *old*, or *unemployed* and *not in the labor force*). In addition, there are differences in dialect, local usage, age, social standing, native or non-native language, and so on that can complicate judgments of similarity. In these cases, identifying useful groupings may be less straightforward than in the first case. Obviously, there is some level of agreement when words mean the same thing – this is information that most dictionaries and general thesauri provide. Even there, however, the words listed as synonyms need to be interpreted for "flavor" or nuances of meaning. For example, *Roget's Thesaurus* (4th Edition, 1977, Harper & Row) lists these words in the same cluster as *unemployed*: *idle, fallow, otiose, unemployed, unoccupied, disengaged, jobless, out of work, out of employment, out of a job, out of harness, free, available, at liberty, at leisure, at loose ends, unemployable, lumpen, leisure, leisured, off duty, off work, off*. (Concept 708.17). Obviously, some of these are closer to the technical meaning of *unemployed* than others." [Haas, 2000: 14-15].

Language ambiguity also plays a role in paraphrasing. Some words or phrases in general language can be ambiguous, but are less ambiguous as a factor of the domain or when in context. For instance, the noun *share* can mean: (1) a part or portion belonging to, distributed to, contributed by, or owed by a person or group; (2) an equitable portion: do one's share of the work; (3) any of the equal parts into which the capital stock of a corporation or company is divided (definition extracted from the The Free Dictionary). If *share* occurs with the verb *to do* and is associated with a possessive pronoun in the general language, then it can only have the meaning represented in (2), but if it occurs with a percentage in the financial field, it has the meaning represented in (3). Other words may or may not be ambiguous in general, but be ambiguous in a specific context or field of knowledge or domain [Haas, 1999]; [Haas & Hert, 2000]. For example, the Portuguese word *corpo* may be ambiguous in the field of medicine. In internal medicine, *corpo* means *body* and in anatomy, it means *corpse*. Other words may be ambiguous in a specific field by a non-expert in that field. For example, the word *janela* (*window*) can be used as a concrete noun, in general language, or as an information-type noun, recorded data in the field of computers. A person who does not know anything about computers may be confused about what *janela* means, but a computer-informed person will have no

difficulty in disambiguating the meaning of that word in context. In these instances, paraphrases may use ambiguous terms and include additional information to disambiguate them and make them understandable by a non-expert.

The concepts of synonymy and ambiguity are associated with the concept of polysemy. Polysemy relates to the several related meanings of a word or expression. For example, the word *wood* can mean a piece of a tree or a geographical area with many trees (small forest). Most words are polysemous, that means they have more than one meaning in the same syntactic category. They contrast with the words that are monosemous, having a single meaning (not ambiguous). A synonym is an identical word to one of the meanings of a polysemous word.

Much research on paraphrases [Boonthum, 2004], semantic relatedness [Miller, 1995] [O'Grady et al., 1996] [Fellbaum, 1998] [van der Plas & Tiedemann, 2006] and entailment [Braz et al., 2005] [Roth & Sammons, 2007] include related terms, such as the relationship terms hypernyms, hyponyms, and antonyms. For example, some consider that a sentence such as *the animal is suffering* is a paraphrase of the sentence *the dog is suffering*, where one of its hypernyms, the word *animal*, is replacing the noun *dog* of the original sentence. Antonyms can also play an important role in paraphrasing. For example, the sentence *she is not stupid* can be considered a paraphrase of the sentence *she is smart* in certain contexts, especially when spoken in a given intonation.

In the figurative sense, it is also common to use the term synonym to relate terms that have the same connotation or are co-referents, for example, *Paris* and *The City of Lights or Eve* and *the first woman*. Co-reference is often used as a paraphrasing strategy. In addition, the term *mentally challenged* means the same thing as *retarded* in a different "register". Identity co-references between two entities or denotation of the same object are frequently used strategies to paraphrasing.

Metaphors are rhetorical devices used to compare unconventionally related subjects. [Lakoff & Johnson, 1980] claim that metaphor is omnipresent in everyday language. [Lakoff, 1992] [Lakoff, 1996] suggest that metaphorical expressions are cross-domain mappings in the conceptual system. They reveal how human cognition works. There are many different types of these cognitive mechanisms to explain similarity in domains such as artistic language [Davidson, 1978], political language [Gruber, 1993] [Zhang, 1998] [Kerzazi-Lasri, 2003], etc. Even terminology is often conditioned by metaphors

[Temmerman, 2001] [Temmerman, 2002]. Many basic concepts are established by our bodily experiences and metaphor involves a mapping of the structure of a source domain onto the structure of a target domain [Lakoff & Turner, 1989], where familiar notions of the physical world are compared to abstract conceptual domains, of the mental and emotional field. Metaphorical language is difficult to paraphrase, especially when metaphors have an artistic function. On the one hand, [Davidson, 1978] expresses that "understanding a metaphor is as much a creative endeavor as making a metaphor, and as little guided by rules" and attempting to paraphrase a metaphor makes as much sense as attempting to paraphrase a photograph. On the other hand, [Guttenplan, 2005] claims that paraphrasing a metaphor is not merely difficult or impossible; rather, it is "outright inappropriate". This constraint relies on employing a 'strict' sense of paraphrase -- "restating the sense of the passage in other words" [Guttenplan, 2005: 18]. In this sense, elucidation or explanation of the meaning of a metaphor is allowed for the purpose of providing clarity, but restatement is not acceptable.

Paraphrasing also deals with connotation. In linguistics, connotation refers to a certain intended meaning that is not the literal meaning (= denotation). In logic, connotation relates to intension. For example, in the sentence *he is a pig* the intension is not that *he is a four legged animal*, but that *he is dirty/unhygienic/filthy/a slob* or *he is not clean* or even *he has impure thoughts*. In a contextual situation, it would be possible that someone is representing the role of a pig in a play, movie or other entertainment event, and then the paraphrase would need to express that meaning, as for instance in *he is playing a pig*. When paraphrasing, it is obligatory to maintain the original intent. Humans grasp the meaning out of the context. A computer grammar would require a style tag for connotation, in order to properly represent the intended meaning. A rule of the type [ARG_0+be+PredAdj+AN+mammal], where ARG_0 is a human subject and [PredAdj+AN+mammal] is a predicative adjective of the type animal, more specifically a mammal, would allow disambiguation of the expression as an idiomatic support verb construction, with the meaning of an insult. A rule of the type [ARG_0+play+AN+mammal] would account for the meaning of a male person playing the role of a pig. Paraphrasing syntactic-semantic rules to allow disambiguation would be: to be a pig in + [IN+arts] = to play a pig in + [IN+arts].

2.8. **Granularity of Meaning Representation in Paraphrasing**

Granularity of representation is a term used by some authors [Edmonds & Hirst, 2002] [Albertoni et al., 2006]. [Edmonds & Hirst, 2002] define it as "the level of detail used to describe or represent the meanings of a word. A fine-grained representation can encode subtle distinctions, whereas a coarse-grained representation is crude and glosses over variation". The idea of different levels of granularity related to the meaning of a word or a term has its roots in the theory of meaning [Catford, 1965]. This author established the theoretical connection between translation and comprehension. His theory comprehends different levels in the meaning of a text. At one level, there are the meanings of individual words or phrases, and the meaning of clauses that form the sentence. On a higher level, there are the meanings of the individual sentences belonging to a passage. Finally, there is the meaning of the passage as a whole. For [Catford, 1965: 20], "the substitution or replacement of textual material in one language by equivalent textual material in another language" operates at the sentence level. Below that level, it is difficult to establish equivalence of meaning between a source and a target language. The argument for his claim relates to the fact that the translation of a word or of an entire clause in one language may be a phrase or simply a word in another language. However, translation of a source language into a target language does not operate necessarily only at the sentence level in all circumstances. In order to comprehend translation, it is important to understand how paraphrase operates at several different levels according to purpose and to the field of knowledge. In literary texts, paraphrasing relates to analysis, interpretation and synthesis and is at times associated with restatement of passages, more often used in *textual* or *sentential* and *clausal* contexts. From a linguistic standpoint, paraphrase is more often associated with synonymy, and usually operates at the *lexical* or *phrasal* level (words and multiword expressions, etc.). Paraphrasing at the word and at the phrase level operates in the following way: (i) the meaning (= content) remains equivalent and (ii) the structure (= form) of the sentence remains the same or presents a minimal change. Paraphrasing of a phrase often implies a change at the part-of-speech level. For instance, a (support) verb plus a noun can be transformed into a verb with identical value/meaning. Paraphrasing support verb constructions can operate at several levels. A distinction between different levels of meaning representation follows.

2.8.1. **Referential Paraphrasing**

Paraphrase is often used as a way of expressing and explaining different realities. Within the same language, a single sentence may express different real-world situations. One variant of the same language may use words unknown to another variant. One variant may contain a word for a concept, which has no direct equivalent in the other variant. In English there are words, concepts and expressions that are used in only one of the English speaking countries and not in the others. Likewise, in Portuguese, there are words that are specific to the lexicon of one of the variants of the language, because they refer to different realities. Certain names of plants, fruits or animals exist only in certain geographical regions. For example, *azinheira* (*holm oak*; Latin: *Quercus ilex*) is an unknown tree in Brazilian Portuguese and *sapoti* (Amazon exotic fruit; from the *sapodilla* tree; Latin: *Manilkara zapota*) is an unknown fruit in European Portuguese. *Alcatruz* (*the bucket of a water wheel*; from Arabic *al-qadus*) is an example of a word that is used in European Portuguese but not in Brazilian Portuguese and *abati* (Tupi word for *corn*) is a word, which is used in Brazil and is unknown in Portugal (examples from [Barreiro et al., 1996]). When one referent does not exist in the culture of one of the two variants of the language, or in different languages, there is no equivalent expression between them. Words that have no equivalent concept in a certain culture are often not translated and explanatory additional words are needed or useful to convey unconventional foreign meaning. This is a common way of dealing with referential differences.

2.8.2. **Lexical Paraphrasing**

Synonymy refers to a word with an identical or similar meaning, such as the verbs *buy* and *purchase*, where both pairs have the same part-of-speech or syntactic category. It operates at the lexical level. The paraphrases *John bought a car* and *John purchased a car* are identical at the lexical level, because they contain synonyms. Since the only difference in these sentences is one word, the paraphrase is the whole sentence. Both sentences could have the same translation in Portuguese, *O John comprou um carro/automóvel*.

All parts-of-speech elements have synonyms or can be paraphrased. For example, the preposition *other than* is similar to *besides*; the pronoun *they* could be used instead of [these+N], such as in *these men* in an anaphoric relation, etc.

2.8.3. **Phrasal Paraphrasing**

When two paraphrases have more than a word of difference, it is said that paraphrasing operates at the phrasal level. This is the level of paraphrasing where many include multiword expressions or the so-called collocations. Support verb construction paraphrases often operate at the phrasal level. However, they display both lexical and phrasal properties. Whenever the meaning of the support verb construction is non-compositional, they behave as lexical items. Whenever they occur in syntactic processes, they present phrasal properties. For example, the structure [N0 *fazer a apresentação de* N2] (N0 *to do the presentation of* N2) is equivalent to the structure [N0 *apresentar* N1] (N0 *to present* N1). The paraphrasing consists of phrasal replacement. In *fazer uma promessa* (*to make a promise*), a phrase is transformed into a single lexical item, *prometer* (*to promise*).

2.8.4. **Syntactic Paraphrasing**

When paraphrasing extends to more than a word or phrase, it operates at the sentence level. At sentence level, paraphrasing often implies reorganization of the syntactic structure of the original sentence. In the framework of Generative Transformational Grammar [Chomsky, 1957] [Chomsky, 1965] [Chomsky, 1975], transformational rules such as active to passive voice transformation, are instances of paraphrasal representation where paraphrases are surface structures for the same deep structure. Passive-active paraphrasing operates at the syntactic level. There is a structural transformation, but an identical or similar meaning is preserved. For example, the passive sentence *The medication was prescribed by the doctor* can be paraphrased by the active sentence *The doctor prescribed the medication*. This paraphrasing deals with structural changes at the sentence level. Arguments change positions. The verb in the passive voice is transformed into the simple past in the active voice. But the semantic roles have not changed. The subject (ARG0) is always *the doctor* and the object (ARG1) is always *the medication*. In both of the above sentences, *the doctor* has the semantic role of agent. Some languages accept passives better than others. Sometimes it is necessary to transform a source

language active sentence into a target language passive sentence or vice-versa to maintain a native fluency.

2.8.5. **Lexical-Syntactic Paraphrasing**

Most paraphrases operate at more than one level. The level in which the paraphrase is used is very important in defining the type of paraphrase. [Shinyama & Sekine, 2005: 2], point out that "if one found two sports articles which said "John won" and "John had a glorious victory", these expressions are correct paraphrases as there would be the same entity "John" filling the same table." This is an example of paraphrase operating at a lexical-syntactic level. However, they may have a different situational usage. *John won* is a neutral objective statement, but *John had a glorious victory* is clearly emphatic. Even if in both examples there is the meaning of an event that resulted in a victory and this victory is assigned to *John*, there is a clear distinction in the focus. This poses the question of meaning preservation in paraphrasing - a controversial topic, since the thought of what a paraphrase is may vary among authors. Recent works confirm that the rank of meaning preservation in a paraphrase is relative to the application for which the paraphrase is used. [Barzilay & McKeown, 2001] consider that the paraphrases retain approximate conceptual equivalence, and are not limited only to synonymy relations. When it comes to sentences or broader linguistic context levels, the number of possible alternatives is larger and the more words or phrases changed in the same sentence, the greater the risk of introducing differences in meaning. This is the reason why it is so important to define a linguistic unit properly. Multiword expressions are the linguistic units that present most problems in translation, especially in machine translation. Paraphrasing them is the only possible solution to overcome many of those problems.

2.8.6. **Paraphrasing of Multiword Expressions**

As [Santos, 1999] points out, multiword expressions are language-pair specific and often understandable only by reference to the real world situation of the text (textual context), rather than the co-text. This means that each language has its own ways of conveying expressions, but also the same language can use different expressions for the same

meaning depending on the style chosen for a particular text. Chapter 3 will be dedicated to multiword expressions.

Chapter 3

Multiword Expressions

*

Chapter Three describes multiword expressions. It discusses the interest in multiword expressions for natural language processing. It exposes the non-compositionality, the non-substitutability, the non-modifiability and the non-translatability properties that characterize multiword expressions overall. Likewise, it establishes a classification of multiword expressions, distinguishing the differences between major classes: lexical units, frozen and semi-frozen expressions, idioms, phraseology, proverbs, collocations, metaphors, and lexical bundles. Finally, it discusses solutions for idiomacity in translation.

*

A multiword expression is a group of two or more words in a language lexicon that generally conveys a single meaning. Multiword expressions are abundant in language, but yet until recently they have been considered idiosyncratic linguistic objects. Little focus was given to them by traditional theoretical linguistics, grammars describe them inconsistently, and they are not formalized adequately in the dictionaries or applied successfully to machine translation. Certain linguistic trends attempted to do something useful with them. For example, within the area of corpus linguistics, multiword expressions, categorized as collocation phenomena, have been extracted by means of concordance tools, where the linguist specifies a key word in context (aka KWIC) and the tool identifies and retrieves the words immediately surrounding them. Some of the co-ocurrences are random and purely arbitrary; others are statistically relevant (collocation pairs) and give an idea of how words are used (measure of association parameter).

Most studies on collocations [Smadja, 1993] [Kilgarriff & Tugwell, 2001] [Heyer et al., 2001] [Evert & Kermes, 2003] [Orliac & Dillinger, 2003] [Orliac, 2004] consist in identifying collocates within a corpus, with the goal of including them in extended dictionaries. The term collocation is still generally used to define a sequence of two or more words or terms that ‘go together’ with a precise or conventional meaning. These sequences or

combinations of words often co-occur with high frequency, repeatedly/recurrently and in a predictable way. Even though the term became widespread, collocation is related to statistical analysis (co-location means positioning side by side or close together) and covers a broad set of multi-layered linguistic phenomena, which must be identified and studied individually.

3.1. Interest of Multiword Expressions for Natural Language Processing

Recently, in natural language processing, there has been interest in multiword expressions and in the problems they raise. Recent calls for papers for conferences on language resources and evaluation in the area of multiword expressions draw attention to different types of multiword expressions: phrasal verbs, light or support verb constructions, noun compounds, proper names, and non-compositional idioms. The specification of several classes of multiword expressions reflects some progress in their classification. A special issue on multiword expressions was published this year by the International Journal of Language Resources and Evaluation focusing on re-evaluation of extraction and classification techniques and offering motivation for languages other than English. Upcoming conferences will certainly focus on issues of context and language dependent expressions.

The lack of formalization or inadequate processing of multiword expressions triggers problems with the syntactic and semantic analysis of sentences where they occur and make difficult or even damage the performance of natural language processing systems. They are important in question answering, summarization, information extraction and, particularly, in machine translation.

3.2. Non-compositionality, non-substitutability, non-modifiability and non-translatability Properties

Multiword expressions have been characterized by properties of non-compositionality, non-substitutability and non-modifiability. These properties are relevant to a greater or smaller extent, depending on the type of expression, but they are emblematic and apply to the great majority of expressions.

The **non-compositionality property** refers to the fact that the words of the expression put together have a different meaning than the sum of their parts, which means that the meaning of the whole expression cannot be necessarily predicted from the meaning of the individual parts. In non-compositional expressions, such as the compound noun *lua cheia* (*full moon*) or the idiomatic expression *perder a cabeça* (∓ **to lose one's head*), the meanings of the words interrelate in such a way that a new meaning comes out which is very different from the meanings of the words in isolation. A full moon is not ("just") a "moon that is full" or round, but it is a lunar phase that occurs when the Moon is on the opposite side of the Earth from the Sun. Also, one's head or mind is something that is not possible to lose because it is attached to one's body.

The **non-substitutability property** means that it is not possible to substitute a word in a multiword expression with a related word and preserve the same meaning. Paradigmatic replacement does not work if, for example, we replace the adjective in the compound *vinho branco* (*white wine*) by another color adjective other than *verde* (∓ **green* => *young*), in Portuguese, to designate a very dry lightly bubbly white wine with a slight green hue to it, or *red*, in English, to specify another type of wine. No other color adjective could substitute *branco*, *verde* (in Portuguese) or *white* and *red* (in English), because the concept of a wine that is designated by a different color does not exist in the real world. And even the concept of color in these cases is relative because for example, neither the wine nor the grapes are "red" in the basic sense of "red" or "white" in the basic sense of "white", etc. This phenomenon is also known as paradigmatic rupture.

The **non-modifiability property** means that it is not possible to modify a multiword expression by freely introducing additional lexical material or to apply any syntactic transformations while preserving the same meaning. For example, the expression *to see the light at the end of the tunnel* in the sentence "*I saw the light at the end of the tunnel*" is a revelation, but "*I saw the red light at the end of the tunnel*" refers to a real light in a real tunnel and "*at the end of the tunnel is where I saw the light*" is not acceptable. The non-modifiability poperty is not always observed (viz. *pôr* (*mais*) *lenha na fogueira* - ∓ *to put* (*more*) *logs in the fire* => *to make a subject more controversial*), but it characterizes many multiword expressions, especially the more idiomatic ones.

In addition to the non-compositionality, the non-substitutability and the non-modifiability properties, described above, the **non-translatability property** applies to

many multiword expressions. The non-translatability property means that not all meaning can be translated across cultural and linguistic boundaries. Some natural language expressions are simply not translatable into other languages. Many proverbs and idioms are not understandable if they are translated. For example, the Portuguese idiomatic expression *ir à vida* (≠ **go to life*) represents nothing to an English speaker, unless it is explained that it means *to die* (when used with an animate subject) or *to finish* (when used with a non-animate subject), the opposite of what the English speaker might have thought. Similarly, the proverb *filho de peixe sabe nadar* (≠ **son/child of fish knows how to swim*) means nothing unless it is adapted and explained that *children usually have certain skills that they learn from their parents*. In a similar way, the English expression *to pay a visit* cannot be translated into Portuguese word for word (≠ **pagar uma visita*). *Pagar uma visita* in Portuguese may be understood as someone is paying for someone to visit somewhere, when in fact the visitor is visiting someone and not being paid. Translatability is an essential concept for understanding cross-cultural and cross-linguistic encounters. Adaptation and explanation is a common strategy to deal with the problems raised by non translatable expressions.

3.3. **Classification of Multiword Expressions**

Traditionally, the term compound lexical unit [Gross, 1986] was the one used for multiword expressions, but it did not have the same coverage as the current term. The terms **multi-word units** [Glass & Hazen, 1998] and **multiword expressions** [Sag et al., 2001] are more recent. They take account of support verb constructions, compound lexical units, semi-fixed expressions and other expressions.

Multiword expressions can be classified into three main categories: lexical units, frozen and semi-frozen expressions, including proverbs, and lexical bundles. Some multiword expressions do not fit into any of these three major types. For example, institutionalized utterances, such as *a ver vamos!* (*we'll see!*) or *se fosse a ti* (*if I were you*), sentence frames such as *a questão é a seguinte* (≠ *the question is the following*) and text frames such as *em primeiro lugar…, em segundo lugar, … e por último…* (*firstly…, secondly…, … and finally…* also classified independently as compound adverbs) can also be seen as special types of multiword expression. Idioms, such as *dar com a língua nos dentes* (≠ **to hit the tongue in the teeth => to talk when one should be quiet => to reveal a secret*); *estar*

com uma pedra no sapato (∓ **to have a stone in a shoe => to have a problem to solve*); *ir desta para melhor* (∓ **to go from this one to a better one => to die*) or *onde Judas perdeu as botas* (∓ **where Judas lost his boots => in the middle of nowhere => in a remote place*) are semi-frozen or frozen multiword expressions that can fit in one or another class. Many frozen multiword expressions are less variable types of support verb construction, where the support verb is the only variable word in the whole expression. Figure 1 (translated and adapted from [Ranchhod & Carvalho, 2006]) represents the major classes of multiword expression.

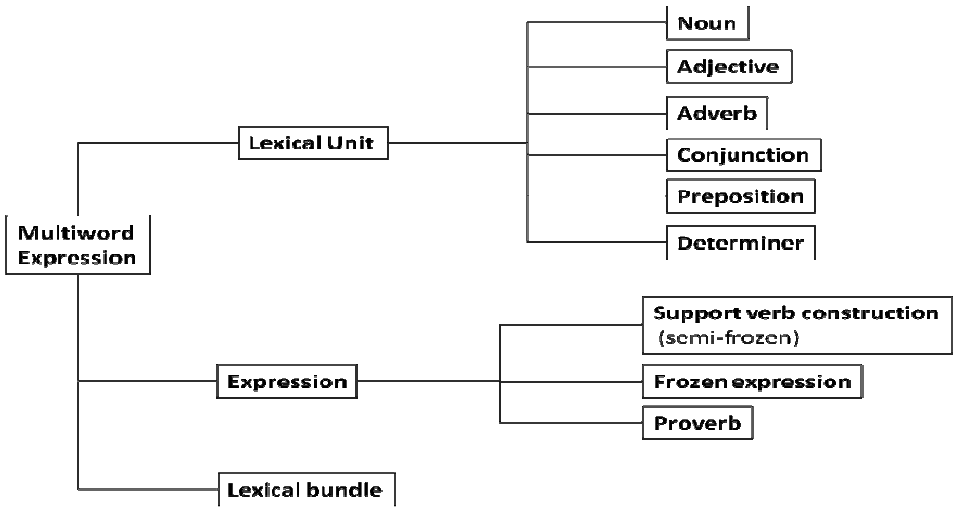

Figure 1: Types of multiword expression

3.3.1. **Lexical Units**

Lexical units are compounds that can be subcategorized into different parts of speech: nouns, adjectives, etc. Compound entries are often more numerous than the simple-word entries for the same part of speech. Some can belong to the general language, and some can be more specific, regarding a given field of knowledge and belonging to that field's terminology. Lexical units can be: compound nouns, viz. *Casa Branca* (*White House*); *piloto de motocross* (*motocross rider*), *bomba injectora* (*injection pump*); compound adjectives, viz. *amigo do ambiente* (*environmentally friendly*); compound adverbs, viz. *já agora* (*by the way*), *para cima e para baixo* (*up and down*), *do avesso* (*inside out*); compound prepositions, viz. *de acordo com* (*in accordance with*); compound conjunctions, viz. *de modo que* (*so that*), *a fim de que* (*with the purpose of*); compound

determiners, viz. *montes de* (*tons of*) and numerals, viz. *uma centena e meia* (*one hundred and fifty*).

Compound nouns are generally formed by a combination of simple words. Their meaning is usually non compositional. They are very frequent and occur in typical nominal contexts, but they are not considered free noun phrases. There are certain linguistic criteria that can facilitate the task of identifying compound nouns. These criteria can vary from morphological behavior of the compound elements, verification of total or partial loss of compositionality (lexical, syntactic and semantic). For example, the expression *poder paternal* (*parental control*) in sentence (3) is considered a compound noun. Certain linguistic tests (syntactic-semantic and morphological) can be used to confirm the compositionality of the expression (cf. (4)-(7)).

(3) Os filhos estão sujeitos ao poder paternal até à sua maioridade (artº 1877º C. Civ.), (…) [TRC]
∓ **Children are subject to parental control until their majority (art. 1877 C. Civ.) (…)*
=> *?Children are subject to parental control until they reach the age of majority (art. 1877 C. Civ.) (…)*

It is not possible to to transform the adjective of the compound (*paternal* => *parental*) into a relative, as can be confirmed in (4).

[Adjective Predicativity]

(4) Os filhos estão sujeitos ao *poder que é paternal
∓ *Children are subject to *control that is parental.*

It is not possible to coordinate another adjective of the same kind as that in the compound (cf. (5)).

[Adjective Coordination]

(5) Os filhos estão sujeitos ao *poder paternal e maternal
∓ *Children are subject to *parental and maternal control*

It is not possible to eliminate the adjective from the compound. The elided adjective results in an ambiguous sentence (cf. (6)).

[Adjective Elision]

(6) Os filhos estão sujeitos ao *poder
∓ *Children are subject to *control*

And finally, it is not possible to replace the adjective with another adjective of the same kind, as example (7) shows.

[Paradigmatic Rupture]

(7) Os filhos estão sujeitos ao poder (paternal + *maternal)
∓ *Children are subject to (parental + *maternal) control.*

[Ranchhod & Carvalho, 2006] include the "variation of adjective degree" as a test for idiomaticity. They claim that it is not possible to modify the adjective by means of co-ocurrence with an intensifier such as *muito* (*very*) as in (8).

[Variation of Adjective Degree]

(8) Os filhos estão sujeitos ao *poder muito paternal
∓ *Children are subject to *very paternal control.*

Even though the presence of an intensifier results in a non-grammatical sentence, the test is not sufficient because the sentence's non-gramaticallity seems to depend on whether a particular adjective is gradable and not necessarily whether it can be modified. Some adjectives do not accept degree variability, because they describe absolute qualities, not gradable ones. For example, the compound *voto secreto* (*secret vote*) is modifiable with adverbs such as *obviamente* (*obviously*), as in *voto obviamente secreto* (*obviously secret vote*). The criterium seems to be more semantic than syntactic.

A different example, the compound *braço de ferro* (∓ *arm of iron*; *arm wrestling*) in (9) also presents a certain morphological and syntactic behavior that makes it a compound. The example and linguistic tests presented in (10)-(14) are from [Ranchhod & Carvalho, 2006].

(9) Mantém-se o **braço de ferro** entre os dois sindicatos.
∓ *There is an **arm-wrestling** between the two unions.*
=> *Both unions refuse **to give in**.*

It is not possible to insert new elements in the noun phrase. The insertion of the adjective *forte* (*strong*) after the noun *braço* (*arm*), results in the non-grammatical sentence that example (10) illustrates.

[Insertion of Elements in the Noun Phrase]

(10) Mantém-se o ***braço** forte **de ferro** entre os dois sindicatos.
∓ **There is a strong arm-wrestling between the two unions.*

It is not possible to delete any element in the noun phrase. Attempts to elide the elements *de ferro* result in a meaningless sentence (cf. (11))

[Elision of Elements in the Noun Phrase]

(11) *Mantém-se o **braço** entre os dois sindicatos
∓ **The arm between the two unions is still going on.*

It is not possible to insert a determiner after the preposition that precedes the second noun (N2). When trying to insert the definite article *o* (*the*) before N2, sentence (12) results are non-grammatical.

[Insertion of a Determiner in Prep N2 Structure]

(12) *Mantém-se o **braço do ferro** entre os dois sindicatos.
∓ **There is wrestling of the arm between the two unions.*

It is not possible to coordinate the noun phrase with any other element. The coordination attempt is unsuccessful, as we can see in (13).

[Noun Phrase Coordination]

(13) *Mantém-se o **braço de ferro** e de aço entre os dois sindicatos.
≡**There is an arm-wrestling and twisting between the two unions.*

It is not possible to replace N2 with any other paradigmatic element (cf. (14)).

[Paradigmatic Rupture]

(14) *Mantém-se o **braço de (ferro** + *aço + *ferro forjado**)** entre os dois sindicatos.
≡ *There is wrestling of (arms + *legs + etc.) between the two unions.

Table 1 (extracted and translated from [Ranchhod & Carvalho, 2006]) represents the class of compound nouns in Portuguese.

Type	Structure	Examples
NA	**N**ome **A**djectivo	via verde; bem comum; desenvolvimento sustentável; poder paternal
NDN	**N**ome **D**e **N**ome	efeito de estufa; braço-de-ferro; segredo de justiça; pensão de alimentos
AN	**A**djectivo **N**ome	falsa modéstia; mau-olhado;
NPN	**N**ome **P**rep **N**ome	barco a remos; voto em branco; depósito a prazo
NPV	**N**ome **P**rep **V**erbo	canção de embalar; ferro de engomar
VN	**V**erbo **N**ome	coca-bichinhos; ganha-pão
PN	**P**rep **N**ome	sem-abrigo; à-vontade
NN	**N**ome **N**ome	cara-metade; raio laser; decreto-lei; queixa-crime
NCN	**N**ome **C**onj **N**ome	saia e casaco; prós e contras
XX	---	Habeas corpus; modus vivendi

Table 1: Classes of compound nouns according to their combinatorial rules

Compound adjectives are also common in Portuguese. The internal structure of compound adjectives is varied, and can be complex. As happens with all non-compositional expressions, there is either a total or a partial fixedness of the elements that constitute the compound. The most common structural types are represented in Table 2 (extracted and translated from [Ranchhod & Carvalho, 2006]).

Type	Structure	Examples
AdvA	Adv Adj	bem-parecido; mal-agradecido; muito visto
APrepC	Adj Prep C	duro de ouvido; novo em folha; baço para espelho
AConjA	Adj Conj Adj	certo e sabido; impávido e sereno; pobre e mal-agradecido
PrepN	Prep Noun	de luxo; de renome; sem retorno

Table 2: Classes of compound adjectives according to their combinatorial rules

Compound adjectives can be integrated into distinct syntactic classes. According to the following principles:

(i) construction of the adjectives with *ser* or *estar* (*to be*) or with only one of those verbs,

(ii) existence or not of free complements,

(iii) capacity to accept a completive in the subject and/or complement position.

Table 3 (extracted from [Ranchhod & Carvalho, 2006]) shows the syntactic classes of the adjectives.

Type	Structure	Examples
SA	N_0 ser Adj	O Zé é **maior e vacinado**
EA	N_0 estar Adj	O bife está **mal-passado**
SEA	N_0 (ser + estar) Adj	O Zé (é + está) **doido varrido**
QSA	(Que F_0) ser Adj	**É certo e sabido** que vai haver problemas
SAPN	N_0 ser Adj Prep N	A Ana é **mal-empregada** para o Zé
EAPN	N_0 estar Adj Prep N	O Zé está **bem visto** junto do eleitorado

Table 3: Syntactic classes of compound adjectives

Compound adverbs are the simplest type of compound. They fill syntactic positions that are characteristic of adverbs and circumstantial complements. In general, they are not compositionally interpretable. They are often introduced by a preposition, but they are not free prepositional phrases. They are facultative adjunctions that commute (have the same or similar value or approximate equality) as simple adverbs (cf. (15)-(16)).

(15) O Governo Japonês recebe de bom grado a posição do JBIC de estudar positivamente a concessão de créditos (…) [CasaCivil]
≃ O Governo Japonês recebe prazerosamente a posição do JBIC de estudar positivamente a concessão de créditos (…)
=> *The Japanese government receives willingly the JBIC position of studying in a positive way the concession of credits (…)*

(16) Prometo contar tudo tintim por tintim. [Blog-3B]
Л Prometo contar tudo ao pormenor / pormenorizadamente
=> *I promise to tell everything in detail / thoroughly*

Some of the properties of the compound adverbs: they do not allow insertions (cf. (17)), reductions (cf. (18)), commutations (cf. (19)), and morphological changes (cf. (20)).

[Paradigmatic Rupture]

(17) Foi uma candidatura contra ventos e marés, um verdadeiro milagre [DiárioIOL]
∓ **It was a candidature against winds and tides, a true miracle*
Foi uma candidatura contra *esses ventos e marés, um verdadeiro milagre
∓ **It was a candidature against those winds and tides, a true miracle*

[Reductions]

(18) *O Zé contou isso à Ana tintim
∓ **O Zé told that to Ana word*

[Comutations]

(19) O Zé expôs a questão de viva (voz + *presença)
∓ *Zé put the question in (person + *presence)*

[Morphologic alterations]

(20) *O Zé expôs a questão de vivas vozes
∓ * *Zé put the question in persons*

Table 4 (extracted from [Ranchhod & Carvalho, 2006]) presents the classes of the Portuguese compound adverbs.

Type	Structure	Examples
P-PADV	Adv	tão-somente
P-PC	Prep C	de rompante
P-PDETC	Prep Det C	à pressa
P-PAC	Prep Adj C	de bom grado
P-PCA	Prep C Adj	a olhos vistos
P-PCDC	Prep C de C	com pezinhos de lã
P-PCPC	Prep C Prep C	de alto a baixo
P-PCDN	Prep C de N	em matéria de *N*
P-PCPN	Prep C Prep N	no tocante a *N*
P-PCONJ	Prep C Conj C	contra ventos e marés
P-PV	Prep V W	a bem dizer
P-ACO	(Adj) como C	como uma porta
P-PVCO	(V) como C	como sopa no mel
P-PPCO	(V) como Prep C	como do dia para a noite
P-PJC	Conj C	e assim por diante
P-PF	F	sem tugir nem mugir

Table 4: Classes of compound adverbs

3.3.2. **(Semi) Frozen Expressions and Proverbs**

Non-compositional expressions such as frozen and semi-frozen expressions, phraseology and proverbs appear often in discourse, and include everyday vocabulary, technical terms, etc. Idioms are typically non-compositional. They are expressions where the meaning of the whole expression cannot be built up from the meaning of its component parts when they are used individually. Their meaning is derived from metaphor and other types of semantic extension. For example, the idiomatic expression *to get up on the wrong side of the bed* cannot be interpreted literally. It means *to be in a bad mood*. As pointed out by [Baptista, 2004], frozen sentences such as *passar de cavalo para burro* (∓ *go from horse to donkey => to be in a worst situation then before*) are elementary sentences that convey semantic predicates. They are different from free, distributional verbs, but they follow general syntactic rules for sentence building; they show important combinatorial constraints, namely, on distributional variation on argument positions and on the application of several transformations. Frozen expressions belong to varied registers and have one property in common: they contain verb-noun combinations that are not distributionally productive, and are not interpreted compositionally. Examples (21) and 0 illustrate idiomatic expressions.

(21) Adoeceu, bateu as botas e viajou para a cidade dos pés juntos [OCadáver]
Ŧ ?*(S)he got sick, kicked the bucket and turned up his/her toes.*
=> Л ?*(S)he got sick, kicked the bucket and went to meet her/his Maker.*

(22) Temos duas hipóteses, ou o governo **comprou gato por lebre** ao escolher a Ota, ou o governo está a **vender gato por lebre** ao país. [Blog-LQL]
Ŧ **We have two hypotheses, either the government bought cat instead of hare when choosing Ota, or the government is selling cat instead of hare to the country.*
=> *We have two hypotheses, either the government was deceived when choosing Ota, or the government is lying to the country.*

Frozen expression properties: they do not allow passivation (cf. (23)), pronominalization (cf. (24)), but in some cases they allow inserts. For example in (25), the insertion of an adverb between the verb and its complement nouns phrase is allowed.

[Passivation]

(23) Logo, logo ele iria esticar o pernil. [Blog-AT]
=> *Soon enough he would kick the bucket.*
*O pernil seria esticado por ele.
Ŧ **The bucket would be kicked by him.*

[Pronominalization]

(24) *Ele esticá-lo-ia
Ŧ *he would kick it

[Insertions – between verb and complement NPs]

(25) Ele iria esticar logo, logo, o pernil.
Ŧ **He would kick soon enough the bucket*

Table 5 presents (extracted from [Ranchhod & Carvalho, 2006]) the classes of frozen expressions.

Type	Structure	Examples
P- VC0	C_0 V W	O Senhor chamou o Zé à sua presença
P- VC1	N_0 V C_1	O Zé perdeu a cabeça
P- VC2	N_0 V (C de C)$_1$	O Zé salvou a honra do convento
P- VC3	N_0 V C_1 A N_1	O Zé deu carta-branca à Ana
P- VC4	N_0 V C_1 Prep N_2	O Zé passou uma esponja sobre o assunto
P- VC5	N_0 V Prep C_1	O Zé rema contra a maré
P- VC6	N_0 V Prep (C de N)$_1$	O Zé tocou na corda sensível da Ana
P- VC7	N_0 V N_1 Prep C_2	O Zé meteu a Ana num chinelo
P- VC8	N_0 V Prep C_1 C_2	O Zé fez das tripas coração
P- VC9	N_0 V C_1 Prep C_2	O Zé entregou a alma ao Criador
P- VC10	N_0 V Prep C_1 Prep C_2	A novidade voou de boca em boca
P- VC11	N_0 V que F Prep C_2	Ele soube isso de fonte segura
P- VC12	N_0 V C_1 Prep que F	O Zé pensa duas vezes antes de falar
P- VC13	N_0 V C_1 Prep C_2 Prep N_3	O Zé tirou as palavras da boca à Ana

Table 5: Classes of frozen expressions

Proverbs are popular multiword expressions and they have been studied in depth by authors such as [Mieder, 1993] [Mieder, 2001] [Mieder, 2004]. A proverb is described in [Mieder, 1993: 24] as "a short, generally known sentence of the folk which contains wisdom, truth, morals, and traditional views in a metaphorical, fixed and memorizable form and which is handed down from generation to generation." There are many types: proverbs, proverbial sayings, clichés, maxims, adages, aphorisms, platitudes, mottos, old saws, inanities, common sayings, etc. Normally they are non-compositional and non-distributionally productive and behave similarly to frozen expressions.

Semi-frozen expressions allow some variability. A support verb construction is a typical example of a semi-frozen expression. Support verb constructions are sentences where a verb without specific lexical content (support verb) appears with a noun or an adjective with verbal properties (with argumental structure), viz. *estar com atenção* (≠ **to be with attention*) and *estar atento* (≠ *to be attentive*), both meaning *to pay attention*. Support verb constructions cross more borders than the traditional compounds, but they are also different from other multiword expressions. Chapter 4 will be exclusively dedicated to support verb constructions.

3.3.3. **Lexical Bundles**

Lexical bundles constitute special types of multiword expression. They have been defined by [Biber et al., 1999a] as common recurrent sequences of words, lexical combinations that cross the classical syntactic borders. They are not complete structural units, they are not idiomatic in meaning, they are structurally complex – often composed of a matrix clause/phrase and the beginning of an embedded clause/phrase and they provide a discourse frame for the presentation of new information. There are three main classes of lexical bundles: **stance bundles**, viz. *I don't know what* (=> *não sabia que*), *if you want to* (=> *se quiseres*); **discourse organizing bundles** viz. *let's talk about* (=> *vamos falar de*), *if you look at* (=> *se vires*), and **referential bundles**, viz. *that's one of the* (=> *esse é um de*), *the nature of the* (=> *a natureza de*). A list of structural types of lexical bundles can be found in the work of the above mentioned authors.

3.4. **Idiomacity and Translation**

As stated, except for lexical bundles, most other multiword expressions are idiomatic, being either lexical units or frozen expressions. They are expressions whose meaning cannot be deduced from their literal definition, the expression has a specific meaning as a whole that is very different from the meaning of each individual word. They contradict the principle of compositionality. They often have a figurative meaning that is known only through common use, some become fixed (fossilized) over time.

Idiomacity is difficult to handle in translation because often the expressions are specific to a certain region or culture. If they are borrowed, they normally have to be adapted to the borrowing language. For example, many proverbs are similar in different countries because they have been borrowed from similar languages and cultures. In Europe, many proverbs come from universal texts, such as the Bible, and have roots in Hebrew and Aramaic texts. But each language uses a particular linguistic strategy. For example, languages use alliteration and rhyme to dispense wisdom, but because words are different, the strategy is also different. For instance, in Portuguese, the proverb *querer é poder* (≠ *to want is to be able to*), or in French *celui qui veut, peut* (≠ *he who wants, can*), correspond to the English *where there is a will, there's a way*, that means that when a person really wants to do something, he/she will find a way of doing it. In

Portuguese and French, verbs are being used while in English nouns are being used and the expression in each language has a very different structure altogether. In this case both the Portuguese and the French use rhyme and the English uses alliteration. But in all cases, it is a sound that is used to dramatize the meaning. Even though these expressions have the same root and could be understood if translated literally, a literal translation would not preserve the idiomacity of the expressions contained in each language.

Other expressions do not have a corresponding expression in the target language. Almost every culture has examples of its own figures of speech, proverbial wisdom or idiomacity, viz. the Portuguese expression *abrir o jogo* (Ŧ **to open the game*). A non-native speaker knowing only the meaning of *abrir* and *jogo*, would not be able to deduce the expression's meaning, which is to reveal details about some issue or to denounce something. The literal meaning would be used by a Portuguese native speaker only in specific contexts, such as in opening the box of a game in order to start playing it or unwrapping a new game given as a present on a special occasion. Similarly, in the English expression *to kick the bucket*, a non-native speaker who did not know the idiom, would understand that someone was in fact literally kicking the object bucket (a intentional aggressive act or a unintentional funny situation) and would not understand the seriousness of the expression's actual meaning, which is *to die*. What would be the most appropriate English translation for *abrir o jogo*? What would be the best Portuguese translation for *to kick the bucket*? If the translator wants to be faithful to the original meaning and is knowledgeable or skillful enough to find a corresponding idiomatic expression in the target language, he/she may opt to use it. But no bad judgment about the quality of his/her translation would be made if the translator opted for a less idiomatic expression. In a particular type of text, the translator decides the style of the text requires a non-idiomatic expression. Often the writer may not be as aware as the translator about writing conventions and the translator may help improve the quality of the text within the style intended by providing the appropriate stylistic choices.

Adaptations are common strategies to preserve the idiomacity of an expression that has no corresponding translation. For example, an idiomatic translation of the frozen expression *um osso duro de roer* (Ŧ **a hard bone to chew on*), would be the English expression *a bitter pill to swallow*.

Translation of idioms is difficult because they require some knowledge, information, or experience, in order to use them only within a culture where both parties have common reference. Idioms are colloquial metaphors. They are part of the culture. As cultures are typically localized, idioms are often not useful outside the local context. However some idioms can be more universally used than others, and they can be easily translated, metaphorical meaning can be more easily deduced. However, in general, idioms tend to confuse those not already familiar with them; students of a new language must learn its idiomatic expressions the way they learn its vocabulary. They are no less difficult for machines to understand.

3.5. **Conclusion**

Many studies have attempted to identify multiword expressions/collocations within a corpus, with the goal of including them in extended dictionaries. As electronic dictionaries provide enhanced meaning of single words, including contextual significance and valuable tagging data, the role of a bilingual dictionary is to include entries for multiword expressions when the lexicographer considers that the understanding and analysis of each type of multiword expression needs to be enlarged and refined. The more that is known about multiword expressions and their different categories, the more sophisticated the descriptions in the electronic dictionaries or the more accurately they are formalized in computer grammars, the better the quality of machine translation output and of natural languages applications in general. The research described here begins with support verb constructions and their paraphrases. The purpose is to identify paraphrasing capabilities between several multiword expressions (monolingual and bilingual), as part of the problem of identifying paraphrases. The ability to give the machine translation user multilingual paraphrasing ability constitutes an important step towards achieving better quality machine translation and that is where the focus will be.

Chapter 4

Support Verb Constructions

*

Chapter Four presents support verb construction-specific linguistic knowledge, including syntactic-semantic and transformational properties, paraphrases, and their relevance to machine translation. The most prominent models for interpretation of multiword predicates are briefly described. The syntactic-semantic and transformational properties of the support verb constructions are presented in the light of the Lexicon-Grammar theory. Predicate noun and predicate adjective constructions are defined; syntactic tests to distinguish support verbs are presented; issues such as the semantic weight and stylistic variants of the support verb are pointed out; predicate-argument structure and distributional properties, such as co-occurrence of modifiers and prepositions are discussed.

*

Support verb constructions are a pervasive feature in the linguistic system of many languages. A large amount of predicates are expressed by verbs, or can be converted into semantically analogous verbs. However, certain predicates can only be conveyed by means of support verb constructions, as they play an important role in the linguistic system of many languages. For example, Portuguese support verb constructions such as *fazer a cama* (*to make the bed*), *dar as mãos* (₸ **to give the hands => to hold hands*;), *fazer um galo na cabeça* (*to make a bump on one's head*) or *fazer uma mini-cirurgia* (*to do a small surgery*) have no corresponding lexical strong verbs or other verbal expressions to represent the same meaning. All these support verb constructions are made up with autonomous predicate nouns. Autonomous predicate nouns occur frequently in both technical and specialized expressions, and in common usage expressions. In Portuguese, as well as in other Romance languages and also in English, there are plenty of examples of support verb constructions that are mandatory and cannot be substituted by a lexical verb without losing meaning. However, support verb constructions can be used also for

social reasons where they convey meaning that can be conveyed using equivalent or similar expressions. They can be used for purely stylistic reasons, for linguistic preference, simplification or to make an idea sound more elaborated or sophisticated (viz. *realizar/efectuar uma viagem* - ₸ **to perform a trip*). But they may also be used deliberately to express vagueness, hesitancy, or other sentiments that form the linguistic etiquette systems found in many cultures. Some support verb constructions are used strategically to convey properties comparable to the ones conveyed by modals, etc. The next few sections will discuss support verb constructions from a linguistic and translation viewpoint.

4.1. **Models for Interpretation of Support Verb Constructions**

Multiword predicates such as support verb constructions have been extensively studied both within the French and the Anglo-Saxon schools of linguistics since the 1980's by authors like [Gross, 1981]; [Allerton, 1989]; [Giry-Schneider, 1978]; [Giry-Schneider, 1987]; [Dras, 1995]; [Mel'čuk, 2003] and [Fillmore et al, 2003], among others. In many studies, support verbs are known as *light* or *weak* verbs, mostly in the Anglo-Saxon school [Grimshaw & Mester, 1988] [Larson, 1988] [Hale & Keyser, 1993] [Chomsky, 1999] [Kearns, 2002] [Butt, 2003], among others. There are a few models for interpretation of support verb constructions. The most significant ones will now be presented.

The Meaning-Text Theory model (MTT), conceived by [Mel'čuk, 1988]; [Mel'čuk, 1996] formalizes and describes support verb constructions in terms of their lexical functions [Mel'čuk, 2003]. Lexical functions are deep lexical units that describe dependencies, such as argument sharing properties and other syntactic and lexical relations. Lexical functions specify which support verbs can occur with which nominal and verbal entries. They can be used as an interlingua to facilitate lexical transfer and are useful in machine translation projects for the establishment of relations across languages.

Besides MTT, there are other models for semantic interpretation of [Verb + Noun] combinations. The lexicographic FrameNet model [Fillmore et al., 2002] [Fillmore et al., 2003] is one of the models that compete with MTT. FrameNet records for a set of support verbs the information necessary for the representation of argument mapping relations between a support verb and a nominalization. FrameNet encodes where the nominalization occurs with respect to its support verb, the role of the 'shared argument'

(the argument of the support verb that is also an argument of the nominalization) and its syntactic realization.

The extended NOMLEX dictionary developed by the Proteus Project at NYU also covers support verb construction phenomena. They map syntactic positions in nominalizations to verbal arguments and identify the allowed complements for a nominalization, relating the nominal complements to the arguments of the corresponding verb, including information about support verbs. [Meyers et al., 2004b] consider nominal argument sharing phenomena, in which support verb constructions fit, as a subtype of multiword expression.

Finally, another well-known model is the Lexicon-Grammar [Gross and followers] whose emphasis is on determining the syntactic and transformational properties of the predicates of individual languages [Gross, 1981] [Gross, 1982a]. Nominal predicates have been studied thoroughly in this theoretical framework. Lexicon-Grammar assumes that the basic unit of meaning is the simple (or elementary) sentence rather than the word or the phrase. The simple sentence contains a predicate, which is the head of that sentence. The properties of that predicate (the sentence where it occurs) are stored in so-called lexicon-grammar tables or in the dictionary. The Lexicon-Grammar interest in studying the properties of the predicates extends to predicate nouns which are not nominalizations, but are verb-like in meaning. Consequently, the support verbs selected by each predicate are paid attention to. The Lexicon-Grammar model is the one adopted in the research work described in this dissertation.

4.2. **The Lexicon-Grammar Model**

The term support verb construction became common in the French school of linguistics after the designation of **support verb** for verbs carrying little or no semantic value. [Gross, 1975] and later developments of the Lexicon-Grammar Theory are based on the methodological principles established in the transformational operator grammar by [Harris, 1951] [Harris, 1957] [Harris, 1964]. Within the context of this work, a **support verb construction** is a multiword or complex predicate consisting of a semantically weak verb (the support verb), and a predicate noun, a predicate adjective, or a predicate adverb. Predicate nouns and predicate adjectives are also known as predicate nominals. *Fazer*, *dar* or *ter* may represent support verbs in Portuguese and *have*, *take* or *give* may

represent support verbs in English. For example, the support verb construction *dar apoio a* (*to give support to*) is made up of the support verb *dar* (*to give*) and the predicate noun *apoio* (*support*); the support verb construction *estar cansado* (*to be tired*) is made up of the support verb *estar* (*to be*) and the predicate adjective *cansado* (*tired*) and the support verb construction *ir depressa* (*to go fast => to speed up*) is made up of the support verb *ir* (*to go*) and the predicate adverb *depressa* (*fast*). Predicate adjective constructions, such as *estar atento a* (*to be attentive to*) will be illustrated only when these constructions have a predicate noun associated to them (*prestar atenção a => to pay attention to*) to illustrate paraphrasing, but focus will be given mostly to predicative noun constructions. Predicate adverb constructions are rare, and they will not be considered in this study.

Support verbs have been extensively and systematically studied within the Lexicon-Grammar Theory, from both theoretical and practical perspectives over a considerable period, by many authors and in several different languages. Researchers of the Laboratoire d'Automatique Documentaire et Linguistique (LADL) have worked on support verb constructions since the seventies, inspired by the work done on French by M. Gross. There are many detailed works on this topic in many different languages, including Portuguese [Ranchhod, 1983] [Ranchhod, 1990] [Baptista, 2005] [Chacoto, 2005]. Support verb constructions have also been taken into account in contrastive studies [Salkoff, 1990] [Salkoff, 1999].

Most studies on support verb constructions outline the representation of argument mapping relations between a support verb and a nominalization. But support verbs often combine with autonomous predicate nouns. Some studies focus on predicate adjective constructions [Casteleiro, 1981] [Carvalho, 2007]. Until now, we are not aware of any studies that contrast Portuguese support verb constructions with other languages, or of Portuguese-English and English-Portuguese machine readable resources containing paraphrases for support verb constructions.

4.3. **Predicate Noun and Predicate Adjective Constructions**

There are two main linguistic components in the support verb constructions presented in this study: the verbal component, which is the support verb, and the nominal component, which is either a predicate noun or a predicate adjective. The support verb expresses the grammatical function in the sentence, carrying the features of person, number, tense,

and aspectual value, and sharing one or more arguments with the predicate, but contributing with little or no meaning to the sentence. Predicate nominals hold the main lexical meaning of the sentence and select their support verbs. They denote events, states, qualities, properties, etc., the same way verb predicates do. For example, *estar doente* (*to be sick* – temporarily) is a state, and *ser doente* (*to be sick* – permanently) is a characteristic of the individual. Similarly to what happens in predicate noun constructions, in predicate adjective constructions, predicate adjectives are the elements that select the arguments and the support verb with which they co-occur.

Predicate nouns can be derived or autonomous. Derived predicate nouns are verbs "disguised" as nouns, i.e., nominalizations or nominalized verbs. For example, the noun *apresentação* (*presentation*) is a nominalization, morphologically derived from and semantically related to the verb *apresentar* (*to present*). Autonomous predicate nouns (non-nominalizations) are morphologically independent. An autonomous predicate noun does not have a verb that is morphologically related to it. For example, in *fazer um mestrado* (*to do a master's*), there is no such verb as **mestrar* (**to master*) or in English, the noun *fun*, as in *to have fun*, does not correspond to the verb **to fun*. Similarly, the predicate noun *boas-vindas* (*welcome*) is not a nominalization because it does not derive morphologically from any verb. There is no such verb as **boa-vindar*. However, *boas-vindas* is a verb-like noun that can take complements the same way a verb does. For example, in the sentence *Ele dá as boas-vindas a todos os participantes* (∓ **he gives the welcome to all participants => he welcomes all participants*), *todos os participantes* is a complement of the predicate noun *boas-vindas*. In Portuguese, the expression *dar as boas-vindas* could be paraphrased by a semantically-related verb, *saudar* (∓ **to salute => to greet*). The same way, the English noun *role*, as in the example: *His role in Iran in 1952* (*was crucial*) (example from [Gross, 1982a]), is a predicate noun for which there is no corresponding verb. There is no such verb as **to role*. According to Gross, non-nominalizations like *role* take place and time modification, just like verbs.

Usually *ser* or *estar* (*to be*) is called **copula** or **linking** verb in most grammars. Following [Gross, 1996], we use the term support verb when referring to the verb of the predicate noun constructions, such as *estar concentrado* (*to be focused*), where the predicate adjective *concentrado* (*focused*) functions as the predicate of the sentence where it

occurs. It is the element that plays the central role, and that selects its support verb, which is there only to support it, and to select its arguments.

Many support verb constructions are made of complex nominals. For example, complex predicate nouns are either compounds or modified nouns. In the support verb construction *fazer uma visita relâmpago* (∓ ?*to make a lightening visit*), the predicate nominal *visita relâmpago* is a compound lexical unit with the same meaning as *visita rápida* (*quick visit*). Support verb constructions with complex predicate nouns are difficult to recognize by natural language processing systems mainly because most systems do not contain a good coverage of compounds. Recognition of a support verb construction with a complex predicate noun implies that the compound lexical unit that forms the complex predicate is recognized adequately, before the whole expression is formalized. This is a strong enough reason to gather a comprehensive dictionary of compounds.

4.3.1. **Prepositional Predicate Noun Constructions**

A support verb can be attached to a predicate noun by means of a preposition. This preposition is selected by the predicate noun for that particular support verb. Examples of prepositional predicate noun constructions are: *estar com soluços* (∓ *to have hiccups* = *to hiccough*) (cf. (1)); *estar em dívida* (∓ ?*to be in debt* = *to owe*) (cf. (2)) or *estar de férias* (*to be on vacation* = *to vacation*) (cf. (3)). They can be compared to the English verb particles.

(1) Desculpem-me, devo estar com soluços (…) [Overmundo]
=> *Forgive me, I must have hiccups.*

(2) (…) tal incumprimento dá o direito à empresa financiadora de preencher a livrança, com o valor que entender ainda estar em dívida, acrescido da sanção por incumprimento (…) [DR-Advog]
=> (…) *such non-payment gives the right to the funding company of sending a written order of payment, with the value they understand to be owed, plus the penalty due to the non-payment* (…)

(3) Para os irrequietos nativos de Signo Carneiro, estar de férias não é sinónimo de estar parado. [MH-Astrol]
=> *For the restless born under Aries, to be on vacation doesn't mean to be quiet.*

Prepositional predicate noun constructions normally contain predicate nouns that belong to the same semantic class. For example, in the prepositional predicate noun constructions *entrar/estar em* (∓ *to get into* => *to be in*) as in *entrar/estar em conflito* (*to be in conflict*), *entrar/estar em guerra* (*to be at war*), *entrar/estar em zaragata* (*to be in a fight*), *entrar/estar em confusão*, *entrar/estar em desacordo* (*to be in disagreement*); *entrar/estar em combate* (*to be in a combat*), all predicate nouns denote adverse events, disagreement or conflict of ideas (cf. (4)). Similarly, in the prepositional predicate noun constructions *entrar em pânico, depressão, isolamento, desespero, ansiedade, coma* (∓ **to enter in panic, depression, isolation, despair, anxiety, coma*) all predicate nouns designate adverse events, in this case more related to health or psychological conditions (cf. (5)). However, there are some less common prepositional predicate noun constructions where the predicate nouns do not belong to a specific semantic group, viz. *entrar em contacto com* (∓ *to get in contact with* => *to contact*), or *entrar em vigor* (*to commence*) (cf. (6)).

(4) O Governador do Banco de Inglaterra admite que a Grã-Bretanha pode estar a entrar em recessão. [Not-RTP]
=> *The Governor of the Bank of England admits that Great-Britain may be entering a recession.*

(5) (…) ela sofreu parada cardiorrespiratória e chegou ao hospital com constrição pulmonar, antes de entrar em coma. [EstadoSP]
=> (…) *she suffered a cardiorespiratory arrest and arrived at the hospital with pulmonary constriction before going into a coma.*

(6) As regras anti-tabagistas vão entrar em vigor a partir do dia 1 de Janeiro do próximo ano, pelo que o objectivo passa por limitar o consumo do tabaco (…) [PortugalMail]
=> *The anti-tobacco rules will be enforced from 1st January next year, so the goal is to limit tobacco consumption* (…)

Similarly, the compound support verb *estar com* (*to be with* => *to have* or *to be*) combine with predicate nouns such as *frio* (*cold*), *calor* (*warm*), *fome* (*hunger*), *sede* (*thirst*), among others, to express a feeling or sensation (cf. (7)). Translation of these feelings can be done by using the English verb *to be* plus a noun (*to be cold* => *ter frio*) or plus an adjective (*to be hungry* => *ter fome*). In Portuguese, in most cases, *estar com* is a stylistic variant of *ter*.

(7) 42 milhões de crianças podem estar com fome em África por volta de 2025 [MongaBay]
≃ 42 milhões de crianças podem ter fome em África por volta de 2025
=> *42 million children may be hungry in Africa in 2025*

Prepositional predicate noun constructions have been studied systematically for Portuguese by [Ranchhod, 1983] [Ranchhod, 1990] and [Baptista, 2005]. In these works, support verb constructions are sometimes associated with related adjective constructions, such as *Ele é de uma grande idiotice* (∓ **he is of a great idiotness*) = *ele é idiota* (=> *he is an idiot*) or *Ele está com fome* (∓ **he is with hunger*) = *ele está esfomeado* (=> *he is hungry*), which represent paraphrases of each other.

4.3.2. **Syntactic-Semantic and Distributional Properties**

There are certain syntactic restrictions that permit us to distinguish between support verb constructions and free syntactic structures. Syntactic tests, such as the pronominalization test and others already presented in Chapter 3 help the identification of support verbs. For example, in the structure [V(dar) (Det) N], is *dar* (*to give*) a strong verb or a support verb? Is N a predicate noun or a regular noun? Consider examples (8) and (9).

[Pronominalization]

(8) Afrodite deu uma rosa ao seu filho Eros, o Deus do amor. [GuiaML]
∓ *Afrodite gave a rose to her son Eros, the god of love*
=> *Afrodite gave her son Eros, the god of love, a rose*
Afrodite deu-a ao seu filho Eros, o Deus do amor.
=> *Afrodite gave it to her son Eros, the god of love.*

(9) (…) o Nani até deu um abraço ao Ricardo Carvalho [Blog-MUGC]
∓ (…) *Nani even gave a hug to Ricardo Carvalho*
* (…) o Nani até o deu ao Ricardo Carvalho
∓ *(…) *Nani even gave it to Ricardo Carvalho*

In example (8) it is possible to replace the noun phrase *uma rosa* (*a rose*) by a pronoun, because *dar uma rosa* (*to give a rose*) is not a support verb construction. In example (9) such replacement is not allowed for the noun phrase *um abraço* (*a hug*). The expression *dar um abraço a* (*to give a hug to*) is a support verb construction.

Support verb constructions can have different meanings according to syntactic-semantic and distributional properties of the verb or verbal predicates. For example, in (10) the verb *dar* (*to give*) in the support verb construction *deu a mão* (*gave (his) hand*) means *estender a mão* (Ŧ **to stretch the hand*).

(10) O rapazinho deu a mão a Mariana e guiou-a até à corda. Começaram a saltar de mãos dadas. [Blog-HIST]
=> *The little boy gave Mariana his hand and guided her to the rope. They started jumping holding hands.*

The expression is non-compositional. It is not possible to pluralize the predicate noun and maintain the meaning (**o rapazinho deu as mãos a Mariana* - Ŧ **the little boy gave Mariana his hands*). Other syntactic tests, such as the pronominalization test, fail. **O rapazinho deu-a a Mariana* (Ŧ **the little boy gave it to Mariana*) are not acceptable sentences. However, while the support verb construction *dar a mão* in (10) has a more literal meaning, the support verb construction *dar uma mão* in (11) has a more idiomatic interpretation. It means *ajudar* (*to help*). The two constructions have different meanings which present different syntactic restrictions with respect to the the determiner that can co-occur with each noun (definite article *a* => *the*, in the first case, and indefinite article *uma* => *a*, in the second case).

(11) Jerónimo de Sousa deu uma mão na montagem da festa do Avante! [Not-RTP]
=> *Jerónimo de Sousa gave a hand setting up the festival of Avante!*
Л Jerónimo de Sousa ajudou na montagem da festa do Avante!
=> *Jerónimo de Sousa helped set up the festival of Avante!*

The syntactic representation for the literal meaning would be [dar] [a] [mão], such as in *o João deu a mão à Maria* (**João gaves his hand to Maria* => *João gaves Maria his hand*), while the syntactic representation for the idiomatic meaning would be [*dar uma mão*] Л *ajudar*; such as in *o João deu uma mão à Maria* (Ŧ *João gave Maria a hand* => *João helped Maria*). There are some limitations on the use of *give a hand* = *to help* in English. The expression can only use the NP NP frame, not the NP-to-NP frame, viz. *Mary gave John a hand* means that *Mary helped John*, but **Mary gave a hand to John* sounds odd and may favor the literal meaning.

In Portuguese, in specific contexts, *dar a mão* also appears with an idiomatic meaning. In (12), the expression *dar a mão* also means *ajudar* (*to help*).

(12) ?E os Tucanos cantavam: "Você pagou com traição a quem sempre te deu a mão" [DiariodoRio]
Ŧ **And the Tucans sang: "You betrayed the ones who always gave you a hand"*

Because of the very nature of support verbs, and their inherent lack of meaning, it is possible to find less common occurrences of support verb constructions in corpora, especially web corpora. Some may reflect the influence of "local" usage, i.e., differences between the Brazilian and the European Portuguese variants. Others may be examples produced by non-native speakers or by native speakers that use support verbs that are classified as borderline in terms of acceptability by the majority of the native speakers. We avoided less frequent occurrences of support verb constructions containing support verbs different from the one generally marked as 'the correct one' by language teachers.

In (13), the support verb construction *dar a mão* has a different meaning from (10), (11), and (12), because it consists of an extended structure [dar] [a] [mão] [de], which means *oferecer em casamento* (*Ŧ to offer in marriage*).

(13) A D. Raimundo deu a mão de sua filha Urraca e o Condado da Galiza e a D. Henrique deu a mão de sua filha D. Teresa e o Condado Portucalense [EB1-Porto]
Ŧ ?*He offered his daughter Urraca's hand in marriage and the Galiza county to D. Raimundo and his daughter D. Teresa's hand in marriage and the Portucalense county to D. Henrique.*
=> *He married his daughter Urraca to D. Raimundo and offered them the Galiza county and married his daughter D. Teresa to D. Henrique and offered them the Portucalense county.*

Oferecer em casamento is an old expression used mostly in historical texts. Culture has changed, and dating and marriage protocols are different nowadays, so this sentence would not appear in texts describing current Portuguese culture.

[Bacelar do Nascimento et al., 1993] presents a quantitative study on the distribution of the verb *dar* in corpora and a proposal for the processing of this verb.

Predicate nouns in support verb constructions often occur with pre- or post-modifiers. There are strong constraints between predicate nouns and modifiers, such as determiners

in predicate noun constructions. The presence or absence of a determiner may change the meaning and/or the grammaticality of a multiword predicate. The presence of an indefinite determiner normally forces another modifier (adjective or other) to be explicit regarding the predicate noun. For example, in *ter um sucesso enorme* (*to have/be a big success*) (cf. (14)), there are two types of modifier: a pre-modifier, the definite article *um* (*a*) and a post-modifier, the post-nominal adjective *enorme* (*big*). This adjective is an intensifier for the noun *sucesso* (*success*). If the adjective was not there to modify the noun, the expression would be *ter sucesso* (∓ **to have success*), without the determiner *um* (*a*), which would be used in a different circumstance and with a different meaning (cf. (15)). While the expression *ter* (*imenso*/*muito*) *sucesso* can be paraphrased by *ser (muito) bem sucedido* (∓ *to be (very) (well) successful*), the expression *ter um sucesso enorme* (*to have an enormous success*) would imply a bigger and more subjective involvement from the utterer in the expression, an impression caused to him/her. However, the difference between the meaning expressed by the quantifiers in the first and the second case is very subtle.

(14) Ballmer disse também que a consola está a ter um sucesso enorme e que está a vender muito bem, por todo o mundo, (…) [GameTuga]
=> *Ballmer also said that the console is having an enormous success and that it is selling very well, worldwide, (…)*

(15) Como já se constatou, *ter sucesso* é para muita gente sinónimo de ter dinheiro (...) [Expresso]
=> *As we have already seen, for many people being successful means the same as having money* (…)

Some predicate nouns are more likely to occur without a determiner than others. For example, in (16) (a newspaper headline – characterized by having no determiner at the beginning of the sentence), the autonomous predicate noun *barulho* (*noise*) occurs freely without a determiner, while in (17) the autonomous predicate noun *algazarra* (Л *noise*) is more resistant to the absence of a determiner.

(16) Manifestação fez barulho, mas foi pacífica. [Blog-AdP]
∓ *Manifestation made noise, but was peaceful.*
=> *Manifestation was noisy, but peaceful*

(17) Ele fugiu do quartinho e fez uma algazarra enorme! [Blog-CMF]
₸ *He escaped from the little bedroom and made a big noise*
*Ele fugiu do quartinho e fez algazarra enorme!
=> **He escaped from the little bedroom and made big noise*

While in (18) both predicate nouns *barulho* and *algazarra* combine well with an indefinite article (*um*, *uma* - *a*) and another pre-modifier (*grande* - *big*), in (19) the existence of the indefinite article without any other modifier makes the sentence non-grammatical or grammatically questionable. Normally, the occurrence of an idefinite article requires the presence of a modifier [Chacoto, 2005: 103].

(18) Ele fugiu do quartinho e fez um(a) grande barulho / algazarra [BLOG-DouA]
=> *He escaped from the little bedroom and made a big noise*

(19) ?* Ele fugiu do quartinho e fez um(a) (barulho / algazarra)
=> ?*He escaped from the little bedroom and made a noise*

In (20) and (21) both predicate nouns *barulho* and *algazarra* combine well with indefinite articles (*um*, *uma* - *a*) and post-modifiers (*dos diabos – a hell of a*; *terrível* - *terrible*).

(20) Ele fugiu do quartinho e fez um barulho dos diabos
=> *He escaped from the little bedroom and made a hell of a noise*

(21) Ele fugiu do quartinho e fez uma algazarra terrível
=> *He escaped from the little bedroom and made a terrible noise*

Support verb constructions can present several degrees of variability or be (semi) frozen. Variable support verb constructions are more or less flexible, they may take pre- and/or post-modifiers and predicate nouns can be in the singular or the plural forms. For example, the predicate noun *passeio* (*walk*) in the support verb construction *dar um passeio* (*to go for a walk*) is in the singular form, while the predicate noun *passeios* (*walks*) in the support verb construction *dar* (*uns*) *passeios* (*por*) (*to go for (several/a few) walks* (*on*)) is in the plural form. Both expressions mean basically the same, *passear* (*to walk*). The slight difference between the two expressions is related to the specification of

whether there was only one or more than one walk, but there is no semantic change in the meaning expressed by the predicate. The unspecified plural expression *dar passeios*, as in the sentence *Ele foi dar uns passeios pela praia* (≠ *He went for a few* (*some*) *walks on the beach*), can be modified with the presence of an intensifier, as in the expression *dar muitos passeios* (≠ *to give many walks*). While the definite singular expression *dar um passeio* and the indefinite plural *dar passeios*, mean *passear*; the intensified indefinite plural expression *dar muitos passeios* is similar to *passear muito/bastante* (*to walk a lot*). However, the meaning of a particular support verb construction containing an indefinite article plus a predicate noun (*uma conta*), can be related to the meaning of a support verb construction containing a predicate noun in the plural form (*contas*), but both can be very different from the meaning of a support verb construction with a morphologically and semantically similar singular (indefinite) predicate noun (*conta*). For example, *fazer uma conta* (≠ *to do a mathematical operation*) and *fazer contas* (≠ *to do mathematical operations*) mean *contar* (*to count*) or *calcular* (*to calculate*), but the singular indefinite support verb construction *fazer conta* (≠ **to do count*) is idiomatic, and it can mean: *fazer tenção de* (*to intend to*), *pretender* (*to intend to*), *supor* (*to know*), *esperar* (*to expect*) or *ter intenção de* (≠ *to have intention of*). If instead of a simple noun, the support verb construction contains a compound noun, such as *fazer uma conta de somar* (*to do an addition*); *fazer uma conta de subtrair* (*to do a subtraction*), *fazer uma conta de dividir* (*to do a division*), or *fazer uma conta de multiplicar* (*to do a multiplication*), the meaning is already different from *contar* (*to count*). Information about these constraints needs to be specified at the lexical level.

Support verb constructions such as *ter fome* (*to be hungry*) or *tomar uma bebida* (*to have a drink*) allow some inserts, such as intensifiers. Variations of these support verb constructions can be *ter muita fome* (*to be very hungry*), *ter uma fome dos diabos* (≠ *to have a hell of a hunger*), etc. or *tomar uns copos* (*to have some drinks*), among others.

Within the variable range of expression, some support verb constructions have syntactic patterns that can be quite regular. For example, in Portuguese [*fazer N Prep(de) NP*], corresponding to an equivalent construction [*V NP*], as in *fazer a apresentação do livro*, paraphrasable by *apresentar o livro*; in English, [*make N Prep(of) N*] = [*V NP*], as in *to make a presentation of the book*, paraphraseable by *to present the book*. Some support verb constructions have syntactic patterns which can be slightly more variable, like [*fazer*

de NP/PRO um escravo] corresponding to an equivalent [*escravizar NP/PRO*] (*to make a slave out of someone = to enslave someone*); or [*make (OWN/MOD) decisions*] = [*decide (for SELF/MOD)*] as in *to make my own decisions*, paraphraseable by *to decide for myself* or *to make very intelligent decisions*, paraphraseable by *to decide wisely/intelligently*.

Frozen support verb constructions are generally idiomatic and invariable. Except for the support verb, most of their elements are fixed or fossilized. For example, in the idiomatic support verb constructions *fazer sinal* (*to make a sign*), *fazer vista grossa* (∓ *to have thick sight - to ignore*) and *ter lugar* (*to take place*), the only variable forms of the expressions are the verbs. They have been fossilized and belong to the domain of phraseology. Among several types, a general perspective is contained in [Baptista et al., 2004] [Baptista, 2004] and [Fernandes & Baptista, 2007] and [Fernandes, 2007]. They have been working on representation and recognition of Portuguese fossilized expressions, such as *dar com os pés a Nhum* (∓ **to give with the feet to Nhum => to send Nhum away*). Some of these expressions may have started as support verb constructions and ended up as set expressions. Generally, they are not considered support verb constructions. Fossilization is a phenomenon that occurs with different types of syntactic structure and covers all levels of analysis in a language, both simple and complex sentences. There is no theoretical impossibility for a support verb construction to be the mould in the creation of frozen expressions. However, in these expressions there is no element to which the verb is a support verb. For example, in *N0 dar com os pés a N2*, the noun pés (*feet*) is not a predicate noun. Similarly, in *N0 abrir fogo sobre N2 = disparar* (*N0 to open fire over N2 => to shoot*) or *N0 dar fé de N2 = perceber* (∓ **N0 give faith of N2 => to understand*), the nouns *fogo* (*fire*), *fé* (*faith*) are not predicative elements. In idiomatic expressions, the meaning is expressed by the whole expression. For example, in *dar a mão à palmatória* (*to acknowledge being wrong*) or *dar o braço a torcer* (*to give up*) it is not possible to say that the nouns *mão* (*hand*) and *braço* (*arm*) carry the semantic weight of the expressions. They are frozen expressions, with unified meanings. It is not possible to insert any element in the expression. Only explicit information can permit adequate translation of fully idiomatic (set/frozen) expressions, viz. *fazer boa figura => to do well*; *fazer o sangue subir à cabeça => to make the blood rush to one's head*; *fazer o papel de pai => to play the father*.

4.3.3. **Semantic Weight of the Support Verb and Stylistic Variants**

The semantic weight in a nominal support verb construction is usually carried by the predicate noun or by the whole expression. For example, the support verb *dar* (*to give*) in the support verb construction *dar um abraço a* (*to give a hug to*) contributes with little meaning to the expression, while the predicate noun *abraço* (*hug*) carries the semantic weight of that expression. It is possible to insert a modifier such as *um* (*one*/*a*/*an*) or *muitos* (*many*) before the predicate noun, and keep the basic meaning.

The support verb can be more or less weak. Elementary support verbs tend to be weaker than their variants. For example, stylistic variants such as *cometer* in *cometer um erro* (≠ **to commit a mistake* => *to make a mistake*), *passar* in *passar sinal* (≠ **to pass a sign* => *to signal*), and *realizar* in *realizar uma pesquisa* (≠ *to perform an investigation*) are suitable replacements for the elementary support verb *fazer*, i.e., they are in complementary distribution with *fazer*. Even though they are not semantically strong, lexical-syntactic extensions of elementary support verbs are normally less neutral [Buvet, 2003] and perhaps more specific than these. For example, *fazer desporto* (≠ *to do sports*) might be slightly more informal than *praticar desporto* (*to practice sports*). Some occur exclusively with one specific predicate noun, establishing with that noun a privileged relationship, being considered appropriate to that noun, viz. *cometer um crime* (*to commit a crime*) or *expressar uma opinião* (*to express an opinion*). It is important to establish a link between all these possible paraphrasing capabilities because they help to improve translation quality.

Support verb constructions that use elementary support verbs such as *fazer* (*to make/do*), *dar* (*to give*), *tomar* (*to take*), or *ser de* (*to be of*), seem to occur more frequently in utterances, colloquial, everyday, and popular spoken language, viz.: *dar ênfase a* (*to put emphasis on*), *tomar nota de* (*take note of*) and *fazer uma apresentação* (*to make a presentation*). In scientific and technical domains or formal registers it is more common to find non-elementary support verbs that, combined with the predicate nouns, constitute stylistic variants or lexical-syntactic extensions of the support verb constructions with an elementary support verb. Several support verb stylistic variants are defined in the research works of [Ranchhod, 1990]; [Baptista, 2005]; [Chacoto, 2005]; and [Carvalho, 2007] for Portuguese. Support verb constructions that use non-elementary support verbs, such as *efectuar* (*to perform*), *criar* (*to create*), *realizar* (*to perform*),

desenvolver (*to develop*), or *expressar* (*to express*) are more frequent in a more formal style of text, such as in specialized and technical language. They are used in auditing and financial reports, viz.: *efectuar*/*realizar uma auditoria* (*to perform an audit*), in electrical engineering, viz.: *criar*/*gerar*/*causar um curto-circuito* (*to create*/*generate*/*cause a short circuit*), or in biomedical texts, viz.: *realizar uma neurocirurgia* (*to perform neurosurgery*), among other technical language expressions. In informal style, these stylistic variants are frequently replaced by elementary support verbs, even when co-occurring with technical nouns. It is common to find support verb constructions such as *fazer uma auditoria financeira* (∓ *to do a financial audit*), *fazer um curto-circuito* (∓ *to do a short circuit*), or *fazer uma broncoscopia* (∓ *to do a bronchoscopy*) in newspapers, internet documents and other less specialized or technical texts. They also occur in specialized or technical texts, but less frequently.

It is also common to find non-elementary support-verb constructions in more formal generic language texts or in specialized language, not referring to technical predicate nouns. They are used to create a formal, polished text. For example, *realizar*/*efectuar uma visita oficial a* (∓ *to perform an official visit* => *to make an official visit to*) and *expressar uma opinião* (*to express an opinion*), are expressions with non-elementary support verb constructions that appear in newspaper language or in technical auditing standards documents but not referring to technical terms.

4.3.4. **Predicate-Argument Structure**

In natural language sentences, events or relations are frequently expressed by verbs, but in the case of support verb constructions they are expressed by nominal predicates. Predicate-argument relations are labeled either with grammatical functions (subject, object, etc.), thematic roles (theme, agent, patient) or some other scheme (ARG1, ARG2, ARG3), depending on the theoretical framework. Predicate-argument structure establishes which items are predicates and which items are arguments. Independently of the framework, predicate-argument structure is used to describe levels of representation that neutralize some of the different ways the same ideas can be expressed. Certain syntactic operations, such as passivization, can change the syntactic argument valency, but the semantic arguments remain unchangeable. For example, the same predicate-argument representation may be able to represent both the active and passive. In the

sentence *the doctors operated on John*, the arguments are: ARG0=doctors, and ARG1=John, while in the sentence *John was operated on*, the only argument is: ARG0=John. Even though there was a change in the number of syntactic arguments and the grammatical function "subject" changed (*doctors* in the active and *John* in the passive), the semantic argument did not change. In both sentences *John* is the patient, independently of its grammatical function. If the syntactic argument is left unexpressed, as in the case of agentless passives, the semantic argument is not expressed.

Much work has been done to identify underlying argument structures in English, but less in other languages. Most of it is related to the lexicon because much of this information is specific to individual words, such as verbs or to small classes of words, such as support verb constructions. Some resources already consider these linguistic relationships. For example, PropBank [Palmer et al., 2005] corpus includes predicate-argument structure and applies it to the Penn Treebank [Santorini, 1990] [Marcus et al., 1993] as relation between nodes on the trees using predefined argument frames of verbs [Kingsbury et al., 2002]. Also, some works in the biomedical field discuss the nature of predicate-argument structures for event or relation information extraction purposes. Predicate-argument structure seems to be a useful intermediate structure for information extraction in domains such as molecular biology [Tateisi et al., 2004] and is suitable for representing aspects of the semantics of biomedical verbs [Cohen & Hunter, 2006] [Cohen et al., 2008]. Nevertheless, the analysis of verb phrases is not sufficient to establish the predicate-argument structure, because events and relations can be and often are expressed in nominal phrases or in verbal phrases with predicate nominals. So, it is important to analyze phrases involving non-verbal predicates. As pointed out earlier, works such as [Meyers et al., 2004a]; [Meyers et al., 2004b] have been focusing on predicate nominals, including the analysis and formalization of nominalizations, which occur frequently in support verb constructions. NOMLEX is a dictionary of English nominalizations that describes the allowed complements for a nominalization and relates the nominal complements to the arguments of the corresponding verb. Support verbs share arguments with the nominalizations, so the relationship between the two is captured at the dictionary level.

As demonstrated by [Mel'čuk, 1988] and [Mel'čuk, 1996], different support verbs have different argument-sharing properties. For example, in English the support verb

construction *to have a visit* is different from the support verb construction *to pay a visit* because the subject of the support verb *to have* is the object of *visit*, but the subject of the support verb *to pay* is the subject of *visit*. In *John had a visit* someone visits John, whereas in *John paid a visit*, John visits someone. Mel'čuk lists three lexical functions that apply specifically to support verb constructions containing nominalizations: $Oper_i$, $Func_i$, $Labor_{i,j}$. In this work we are only handling the cases where the predicate noun is the direct object of the support verb subject ($Oper_i$), which are by large the most frequent ones.

However, the analysis becomes more complex when the same support verb has two distinct predicate-argument structures. This is the case of the Portuguese support verb *fazer* (*to do*) that appears with a double function in many support verb constructions, especially biomedical-related support verb constructions. For example, the Portuguese support verb construction *fazer uma amputação* (≠ *to make an amputation*) is ambiguous as far as its semantic role is concerned. It can mean either that someone is doing it or that someone is having it done. So, depending on the arguments of the predicate in one reading or another, the expression can mean *to perform an amputation on someone* paraphraseable by *amputar alguém* (*amputate someone's + body part*) or *to have an amputation* paraphraseable by *ser amputado* (*body part + to be amputated*). In the first reading, the following arguments are involved: (1) ARG0, corresponding to "who", (2) ARG1 corresponding to "what", (3) ARG2 corresponding to "of what" and (4) ARG3 corresponding to "whom". In the second reading, the arguments involved are: (1) ARG0, corresponding to "what", (2) ARG2 corresponding to "of whom", (3) and argument ARG3 corresponds to "what". If the argument ARG0 or subject of the support verb is an agent, such as a doctor or another health care practitioner responsible for the procedure *amputation*, the construction means *to perform an amputation on* or *to amputate on*. On the other hand, if the argument ARG0 of the support verb is a patient, the construction means *to have/undergo an amputation* or (*body part + to be amputated*).

4.4. **Advantages of Paraphrasing**

We argue that if we substitute a support verb construction with a single verb or a stylistic variant, and include explicit arguments, in most cases it renders the text more comprehensible and more ready for machine translation. Verbal predicates demand well-defined arguments. For example, when technical writers use (strong) verbs, the text

opens up in a way that permits identification of the subject of the verb, and they are given the means to identify the object (patient). If paraphrases and predicate-argument knowledge are included as a pre-edit for submission to a machine translation engine, we find that the output is of a higher quality. The machine will not be able to translate well if it does not know who the subject of *fazer uma amputação* (≠ *to make an amputation*) is. The identification of subject and object of predications is one essential problem of natural language analysis. So it is important to interpret beyond the surface syntactic structure. Even for a simple declarative sentence, we need to know the complement structure of a verb in order to identify its arguments correctly. For example, the verb *amputar* (*to amputate*) takes a complement structure *amputar X de Y* (≠ *to amputate X of Y*) where *X* is part of the body and *Y* is the person from whom that part of the body was removed/amputated. Also, we need to know that *amputar* is a verb that has an argument ARG0, which refers to an agentive human, more specifically a medical professional, because only humans with specific qualifications can perform surgical operations. Only under very exceptional and restricted conditions would any other human perform such a procedure. Similarly, doctors and other medical professionals can be (and sometimes are) the patients of medical exams or surgical procedures, but the incidence of being so is smaller and only the context would help identify these cases.

The dual predicate-argument representation of the same support verb construction illustrates the importance of a deeper linguistic analysis of texts, including those related to the biomedical field and of the characteristics of language used in clinical trials to report medical events or surgical operations. It also demonstrates the need for a strategy of representing these relations and teaching them to the machine. We believe that paraphrasing support verb constructions is an efficient strategy to overcome this problem and a suitable solution for the improvement of machine translation. If we take into account that many texts containing support verb constructions are going to be translated, paraphrasing becomes even more important for the success of the results, especially when sentences lack well-defined predicate-argument relations. Ambiguity or misunderstandings resulting from semantically weak verbs and omission of arguments have a serious negative impact on the translation results.

Chapter 5

Paraphrasing and Translation of Support Verb Constructions

*

Chapter Five presents the typology of paraphrasing capabilities of support verb constructions that are relevant to the present study and discusses their usefulness. Support verb constructions are linked to verbal predicates (semantically strong verbs and other verbal phrases) or to non-elementary support verb constructions with equivalent or similar meaning. These established paraphrasing capabilities, lexical or stylistic variants are tested. Finally, the problems that support verb constructions pose to translation will be discussed in depth.

*

5.1. **Typology of Paraphrasing Capabilities**

Lexical-syntactic variability is a property of support verb constructions. In most cases, the replacement of support verb constructions with verbs or with other support verb constructions brings little or no change in meaning. Support verb constructions have certain syntactic-semantic and transformational properties that allow formal paraphrasing capabilities. Predicate noun constructions can often be paraphrased with equivalent predicate noun constructions or with morphologically related predicate adjective constructions and predicate verb constructions. Morphologically related support verb constructions normally select the same arguments. For example, the morpho-syntactic pair of predicates {*depressão*, *deprimido*} (*depression, depressed*) selects one argument, ARG0 filled by a human noun that occupies the syntactic position of subject of the sentence, as in *O João tem uma depressão* (*João has a depression*) and *O João está deprimido* (*João is depressed*). On the other hand, the morpho-syntactic triplet of predicates {*receio, receoso, recear*} (*fear, afraid, to fear*) requires an ARG0-subject and another argument, a complement of these constructions, as in *O João <u>tem receio</u> do pior / do que lhe vai acontecer* (∓ **João <u>has fear</u> of the worst / of what is going to happen to*

him), *O João está receoso do pior / do que lhe vai acontecer* (*João is afraid of the worst / of what is going to happen to him*) and *O João receia o pior / o que lhe vai acontecer* (*João fears the worst / what is going to happen to him*). The constraints imposed by the predicates on the structural nature of the complement (a noun phrase (*o pior – the worst*) or a completive (*o que lhe vai acontecer – what is going to happen to him*)), as well as the lexical specification of the preposition that introduces that complement (in the case of the predicate nominals, the preposition *de* (*of*)) are identical. Types of paraphrasing capabilities were established on the basis of empirical knowledge extracted from the support verb paraphrases found in corpora and analyzed linguistically. The proposed classification of support verb construction paraphrasing capabilities is presented next.

5.1.1. **SVC = V**

The first category corresponds to the paraphrasing capability between semantic and morpho-syntactically related support verb constructions and (strong) verbs. This category includes predicate noun constructions (1)-(2), prepositional predicate noun constructions (3)-(4), and predicate adjective constructions (5).

[Vsup N = V]

(1) Porque é que as pessoas estão a **fazer gestos** com as mãos? [DGIDC]
=> Why are people making gestures with their hands?
= Porque é que as pessoas estão a **gesticular** com as mãos?
=> Why are people gesticulating with their hands?

[Vsup N PrepN = V]

(2) Depois de Britney Spears, a rainha da pop deu um beijo a uma fã. [IOLDiário]
Ŧ *After Britney Spears, the pop queen gave a kiss to a fan.*
= Depois de Britney Spears, a rainha da pop beijou uma fã.
=> *After Britney Spears, the pop queen kissed a fan*

[Vsup Prep N = V]

(3) Alguém está de acordo com o aumento de impostos? [AntiTretas]
Ŧ ?*Is anyone in agreement with the tax raise?*
= Alguém concorda com o aumento de impostos?
=> *Does anyone agree with the tax raise?*

[Vsup Prep N PrepN = V]

(4) Inconformado, José Mourinho entrou em contacto com um dos seus colaboradores para se inteirar do que se estava a passar (…) [GForum]
=> *Dissatisfied, José Mourinho got in contact with one of his colaborators to know what was going on* (…)
Inconformado, José Mourinho contactou um dos seus colaboradores para se inteirar do que se estava a passar (…)
=> *Dissatisfied, José Mourinho contacted one of his colaborators to know what was going on* (…)

[Vsup A = V]

(5) Os pandas ficaram cansados. [OTempo]
=> *The pandas got tired.*
Os pandas cansaram-se.
∓ **The pandas tired themselves.*

5.1.2. **SVC = SVC**

The second category corresponds to the paraphrasing capability between different types of semantic and morpho-syntactically related support verb construction. This category covers paraphrasing capabilities between different predicate noun constructions (6)-(8), between predicate noun and predicate adjective constructions (9)-(11), and between different predicate adjective constructions (12).

[Vsup N = Vsup N]

(6) A Europa tem um papel importante na observância da igualdade de oportunidades entre homens e mulheres. [ECEuropa]
=> *Europe has an important role in observing equality of opportunity between men and women.*
≡ A Europa desempenha um papel importante na observância da igualdade de oportunidades entre homens e mulheres.
=> *Europe plays an important role in observing equality of opportunity between men and women.*

[Vsup Prep N = Vsup Prep N]

(7) O sector financeiro mundial está em crise [Not-Sapo]
=> *The world financial sector is in crisis*
≡ O sector financeiro mundial anda em crise
∓ **The world financial sector goes in crisis*

[Vsup N = Vsup Prep N]

(8) Temos dificuldades económicas grandes [SMS]
=> *We have great economic difficulties*
≡ Atravessamos dificuldades económicas grandes
∓ **We are going through great economic difficulties*
≡ Estamos com dificuldades económicas grandes
∓ **We are with great economic difficulties*

[Vsup N = Vsup N = Vsup A]

(9) Simão fez uma operação ao joelho [TVTuga]
=> *Simão made an operation on his knee*
≡ Simão efectuou / realizou uma operação ao joelho
∓ **Simão performed an operation on his knee*
≡ Simão foi operado ao joelho
=> *Simão was operated on his knee*

[Vsup Prep N = Vsup N = Vsup A]

(10) Você está com ansiedade e teve uma crise de pânico [e-FamilyNet] / você está com uma depressão
∓ **You are with anxiety and had a panic attack / *you are with a depression*
≡ Você tem ansiedade e teve uma crise de pânico / você tem uma depressão
∓ ?*You have anxiety and had a panic attack / you have a depression*
≡ Você está ansioso e teve uma crise de pânico / você está deprimido
You are anxious and had a panic attack / you are depressed

[Vsup Prep N = Vsup A]

(11) Aqui o Tomás também está com atenção *a* tudo que o envolve. [FotoBlog]
∓ ?*Tomás is also attentive to everything around him.*
≡ Aqui o Tomás também está atento a tudo que o envolve.
∓ ?*Tomás PM is paying attention to everything around him.*

[Vsup A = Vsup A]

(12) Em dia de Vitória ninguém fica cansado. [Cipsga]
∓ *?On a winning day nobody gets tired.*
≡ Em dia de Vitória ninguém está cansado.
∓ *?On a winning day nobody is tired.*

5.1.3. **SVC = SVC = V**

A third category corresponds to the paraphrasing capability between different types of semantic and morpho-syntactically related support verb construction and verbs. This

category covers paraphrasing capabilities between predicate noun constructions, predicate adjective constructions and verbs (cf. (13)).

[Vsup N = Vsup A = V]

(13) O Tribunal de Contas fez uma auditoria à empresa (…) [AG-FIN]
=> *The Court Account made an audit of the company (…)*
≡ O Tribunal de Contas efectuou uma auditoria à empresa
=> *The Court Account performed an audit of the company*
≡ O Tribunal de Contas realizou uma auditoria à empresa
=> *The Court Account performed an audit of the company*
≡ A empresa foi auditada (pelo Tribunal de Contas)
=> *The company was audited (by the Court Account)*
≡ O Tribunal de Contas auditou a empresa
=> *The Court Account audited the company*

5.1.4. **SVC = N**

This category consists of the elimination of the support verb after verbs, such as *recomendar* (*to recommend*) and *sugerir* (*to suggest*). This strategy is valuable in specialized fields, such as in the biomedical field (cf. (14)).

[Vsup N = N]

(14) O médico recomendou / sugeriu fazer fisioterapia [COMPARA]
∓ ?*The doctor recommended / suggested *to do physiotherapy*
≡ O médico recomendou / sugeriu fisioterapia
=> *The doctor recommended / suggested physiotherapy*

5.1.5. **SVC + Mod = V**

This category corresponds to the paraphrasing capability between support verb constructions that contain modifiers and verbs. When transforming the support verb construction with modifiers into a verb, some of the detail may be lost. However, the meaning is maintained. The resulting equivalent is called implied meaning paraphrase. In the example (15), the support verb constructions *dar vários passeios* (*to go for several walks*) and *dar um longo passeio* / *dar longos passeios* (*to go for a long walk* / *to go for long walks*) were transformed into the verb *passear* (*to walk*). This results in a loss of the determiner *vários* (*several*) and the modifier (intensifier) *um longo* / *longos* (*a long* / *long*). The verb *passear* (*to walk*) is more generic, and it does not specify how many walks

there were (one or more than one, many) or if the walk(s) was/were short or long. In many contexts that information might not be really necessary, or lack of it does not bring serious negative loss in translation.

[Vsup Mod N = V]

(15) Ele deu vários passeios pela praia [COMPARA]
≠ *He went for several walks on the beach*
≡ Ele deu um longo passeio / uns longos passeios pela praia
≠ *He went for a long walk on the beach / several long walks on the beach*
≡ Ele passeou pela praia
=> *He walked on the beach*

Similarly, in the example (16) the support verb construction *causar uma boa impressão a* (*to cause a good impression on*) has the modifier *uma boa* (*a good*). Transforming *causar uma boa impressão em* (*to cause a good impression on*) into the verb *impressionar* (*to impress*) results in a loss of the qualifying adjective *boa* (*good*). However, not much loss is taking place, because the verb *impressionar* normally already implies a positive sense. A more exact equivalent would be *impressionar pela positiva* (*to impress in a positive sense*), instead of the less probable meaning of causing a negative impression (*causar uma má impressão = impressionar pela negativa*).

[Vsup Mod N PrepN = V]

(16) Áustria causou uma boa impressão na Eurocopa [uClue]
=> *Austria made a good impression in the Euro 2008*
≡ Áustria impressionou na Eurocopa
=> *Austria impressed in the Euro 2008*

5.1.6. **SVC + Mod = V ADV**

The paraphrasing capability between a predicate noun construction, where the predicate is pre- or post-modified by an adjective and a verb plus an adverb morphologically related to that adjective is a straightforward establishment of paraphrases. This transformation requires a change in the part of speech of the adjective into an equivalent adverb. In (17), the pre-modifying adjective *boa* (*good*) is transformed into the adverb *bem* (*well*) and in (18), the post-modifying adjective *correcto* (*correct*) is transformed into the adverb *correctamente* (*correctly*).

[Vsup N A = V Adv]

(17) Governo não tomou uma boa decisão [TVnet]
=> *The Government did not make a good decision*
≡ Governo não decidiu bem
=> *The Government did not decide well*

[Vsup A N = V Adv]

(18) Izmailov tomou uma decisão correcta [DNonline]
=> *Izmailov made a correct decision*
≡ Izmailov decidiu correctamente
=> *Izmailov decided correctly*

However, some transformations of support verb constructions with modifiers into equivalent predicate verb constructions, such as a verb plus an adverb, are not so easy to obtain. Sometimes the paraphrasing capability cannot be established without the insertion of new words and a less-than-ordinary transformation. In (19), the transformation of the support verb construction *provocar uma onda de aplausos* (*to cause a round of applause*) into *aplaudir afincadamente* / *com entusiasmo* (*to applaud enthusiastically* / *with enthusiasm*) is a complex transformation.

[Vsup Mod N = V Adv]

(19) Tevez provocou uma onda de aplausos e assobios nas mais de mil pessoas que estavam presentes no evento. [Not-UOL]
=> ?*Tevez caused a round of applause and whistles from the over two thousand people who were present at the event.*
≡ As mais de mil pessoas que estavam presentes no evento aplaudiram e assobiaram afincadamente Tevez / *com entusiasmo*.
=> *The over two thousand people who were present at the event applauded and whistled Tevez enthusiastically / with enthusiasm.*

5.1.7. **SVC = V [*x*]**

Some paraphrasing capabilities are not possible without requiring additional information. For example, in (20), it is not possible to transform the support verb construction *fazer compras* (*to go shopping*) into the verb *comprar* (*to shop; to buy*) without specifying a complement, even if that complement is vague, such as the noun *coisas* (*things*). In Portuguese, the verb *comprar* (as in English, *to buy*) is transitive and it requires a direct

object. Similarly, in (21), it is not possible to transform the support verb construction *fazer uma descoberta* (*to make a discovery*) into the verb *descobrir* (*to discover*) by itself. In Portuguese the verb *descobrir* (*to discover*) require a direct object, such as *uma coisa* (*something* or [*a ADJ thing*])

[Vsup N = V [x]]

(20) (…) ela pode ir fazer compras o dia inteiro. [Scribd]
=> *she can go shopping the whole day.*
≡ (…) ela pode ir comprar coisas o dia inteiro.
∓ *?She can buy things the whole day.*

(21) Em 1729 o astrônomo francês Jean Jacques d'Ortous de Mairan fez uma descoberta importante em biologia. [Fapesp]
=> *In 1729, the French astronomist Jean Jacques d'Ortous de Mairan made an important discovery in biology.*
≡ Em 1729 o astrônomo francês Jean Jacques d'Ortous de Mairan descobriu uma coisa importante em biologia.
=> *In 1729, the French astronomist Jean Jacques d'Ortous de Mairan discovered something important in biology.*

5.1.8. **SVC = V(meaning)**

This category corresponds to the paraphrasing capability between support verb constructions and verbs which are not morphologically related, but have the same meaning and same syntactic properties. In (22), the support verb construction *dar aulas* (*to give classes*) is equivalent to verbs with the same meaning, such as *ensinar* (*to teach*) and *leccionar* (*to lecture*). This is such an easy type of correspondence to understand and implement that, even though it does not deal with morphological derivation, we have included it in this research.

[Vsup N = Vlex]

(22) Também gostava de dar aulas na universidade [COMPARA]
∓ *?I also would like to give classes at the university*
≡ Também gostava de leccionar na universidade
=> *I also would like to lecture at the university*
≡ Também gostava de ensinar na universidade
=> *I also would like to teach at the university*

5.1.9. **SVC = NP**

The last category corresponds to the paraphrasing capability between support verb constructions and noun phrases. Examples such as (23) and (24) show sentences that are semantically equivalent to noun phrases. In (23), *as dificuldades que passamos* = *as nossas dificuldades* (*the difficulties we have* = *our difficulties*) and in (24), *o papel que a Europa tem* = *o papel da Europa* (*the role that Europe has* = *Europe's role*).

[Npred que Vsup = PronPoss Npred]

(23) As dificuldades que passamos são para nos fortalecer... para podermos viver a vida de verdade... para aprendermos a ser um ser de verdade...humano. [Pensador]
=> *The difficulties we go through serve to makes us stronger… to be able to live life truly… to learn how to be real human beings*
≡ As nossas dificuldades são para nos fortalecer... para podermos viver a vida de verdade... para aprendermos a ser um ser de verdade...humano.
=> *Our difficulties serve to makes us stronger… to be able to live life truly… to learn how to be real human beings*

[Npred que N Vsup = Npred PronPoss]

(24) O papel que a Europa tem no futuro do Mundo é muito importante. [Blog-OlharD]
∓ **The role that Europe has in the world's future is very important.*
≡ O papel da Europa no futuro do Mundo é muito importante.
∓ *Europe's role in the world's future is very important.*

5.2. **Non-Equivalence to Verbs**

The typology of paraphrasing capabilities illustrated in § 5.1, included cases where support verb constructions were equivalent to morphologically and semantically related verbs. However, it is important to note that there is not always a morphologically related verb that confers the same meaning, nor is there always a replacement with a morphologically related verb that produces semantically equivalent results. For example, in sentence (25) the support verb construction *fazer uma procuração* (∓ *to give a power of attorney*) does not have a morphologically equivalent verb that has the same meaning as the expression.

(25) Para fazer uma procuração é necessário que o interessado se apresente pessoalmente no Consulado munido de um documento de identificação válido e o número do código fiscal italiano (…) [CGI-PA]

=> In order to give a power of attorney, it is necessary that the person interested in having it, go in person to a Consulate and bring a valid identification document and the Italian tax payer number (…)

There are morphologically related verbs in *fazer/apresentar uma queixa = queixar-se* (*to make/present a complaint = to complain*) and *fazer/emitir uma mandato = mandatar = tornar obrigatório* (*to make/emit a mandate = to mandate = to make obligatory*). However, in (26) and (27), the support verb constructions are made of compound predicate nouns *queixa-crime* (*crime complaint*) and *mandato de captura* (*capture mandate*). Whenever there is a compound predicate, the meaning changes. Verbs that can replace support verb constructions with simple predicate nouns are not suitable for replacing compound predicate nouns.

(26) Ricardo Sá Fernandes fez saber [...] que pretende (fazer + apresentar + ?interpor) uma queixa-crime contra Domingos [BLOG-Obsc]
*=> Ricardo Sá Fernandes informed […] that he intends (to make + to present + *to interpose) a criminal complaint against Domingos*
≠ Ricardo Sá Fernandes fez saber […] que pretende queixar-se contra Domingos
≠ Ricardo Sá Fernandes informed […] that he intends to complain about Domingos

(27) Apelou aos juízes do TPI para (fazerem + emitirem) um mandato de captura contra EL Béchir [PorDarfur]
=> He asked the TPI judges to (make + emit) an arrest warrant against EL Béchir
≠ ?Apelou aos juízes do TPI para mandatarem contra EL Béchir
≠ ?He asked the TPI judges to mandate against El Béchir

Paraphrasing support verb constructions with specific verbs does not always produce paraphrasing capability in communication and, in some situations they may not be properly understood without having and adding further information. Interesting cases that need further research and that can pose problematic translations are contrasts of meaning between support verb constructions which include nominalizations or other predicate nouns with similar roots to verbs. At times they express a slightly different meaning, involving reflexive forms, such as in *ter pressa ≠ apressar-se* (*to be in a hurry ≠ to hurry up*); other times they convey completely the opposite meaning, with active/passive implications, such as in *ter culpa ≠ culpar* (*to be someone's fault ≠ to blame*) or *ter uma surpresa ≠ surpreender* (*to be surprised ≠ to surprise*). Other cases exist,

having similar orthographic and sometimes semantic roots, but express a totally different meaning, such as in *dar afecto ≠ afectar* (*to give affection ≠ to affect*) or *ter pena ≠ penar* (*to be sorry ≠ to have a hard time*). We need to come back to these examples on another occasion.

5.3. **Usefulness of the Paraphrasing Capabilities**

Most of the categories described above are relatively straightforward and easy to implement in a machine translation environment. Except for the paraphrasing capabilities where new information is required or those requiring complex transformations, automatically converting support verb constructions into verbs can be helpful in producing a controlled language that is easier for the machine to translate. In most cases a substantial reduction of the number of nominalizations can help systems improve their results significantly. Some writers use nominalizations aiming for a formal writing style. However, often formal and technical writing is characterized by the clarity, precision and conciseness that are obtained by strong verbs whenever they are available. This is recommended practice in English style and writing aid manuals and they can also apply to Portuguese. When compliance requirements such as consistent use of strong verbs and avoiding nominalizations are included, controlled language compliant text produces clearer formal writing and better machine translation results. For example, in the Portuguese corpus sentence *Nenhum de nós fez menção a Proserpino ou a Tarciso* (*Neither of us mentioned Proserpinus or Tarcisis*), the support verb construction *fazer menção a* means *to mention*. There is a Portuguese strong verb *mencionar*, which can replace the three words of the support verb construction. In this particular sentence there would be a fourth word reduction because of the coordination. Language that is domain specific, written under controlled conditions, based on the needs of end-users permits machine translation to be almost completely automatic. Linguistic analysis can make machine translation better, even without being limited to a controlled language environment. Language used in our daily lives, even simple literary texts, can be transformed into a machine friendly language. The translator chooses between the support verb construction *fazer menção a* and the verb *mencionar*, according to the contextual need or stylistic preference. In certain contexts the support verb construction

may be preferred. This stylistic variance shows the language richness that is difficult to formalize and transfer to a machine.

5.4. **Importance to our Research**

The role of support verb constructions as paraphrases is central to the discussion of the linguistic phenomenon. First, they carry units of meaning, and their meaning is not compositional. Furthermore, they are abundant in many languages. Even though in this thesis we focus on extracting support verb constructions for the Portuguese-English language pair, the methods and conclusions are applicable to many languages and resulting data can be mapped to different languages that are at the center of our interest. Proper handling of support verb constructions can improve results. Lastly, and most importantly, support verb constructions often represent phrasal alternatives, i.e. paraphrases. In almost all cases, support verb constructions can be substituted by other phrases or words. In some cases, the only possible translation is a paraphrase. As already mentioned, support verb constructions often represent phrasal alternatives to *verbs*. For instance, the Portuguese phrase *dar um beijo* (*to give a kiss*) can be expressed by the semantically equivalent verb *beijar* and the English phrase *to give a kiss* can be expressed by the equivalent verb *to kiss*. Both paraphrases are good translations for either the support verb construction or the verb. Similarly, the support verb construction *ficar cansado* (*to get tired*) can be expressed by the verb *cansar-se* (∓ ?*to tire self*).

Many of these phrases are semi-idiomatic in the sense that one word, the noun or the adjective, retains its full meaning while the verb has a partially obscure meaning, understandable only in its relation to the first, dominant word, viz. *fazer desporto* (∓ ?*to do sports*) which means *praticar desporto* (*to practice sports*).

The ambiguity that attaches to support verbs such as the Portuguese verbs *estar*, *vir*, *fazer*, *dar*, *ter* or English *be*, *take*, *have*, *make*, among others, and their semantic contribution to the support verb construction ranges between no semantic contribution to the point where it is difficult to say which word contributes most to the support verb construction, the support verb, or the noun or noun phrase that occurs with it. In the cases where there is a semantically strong verb that is morphologically related to the noun in the support verb construction, this verb paraphrases the support verb construction. For instance, in Portuguese *dar um beijo* is paraphraseable by the verb

beijar (*give a kiss* = *kiss*) or in English *take a seat* is paraphraseable by the verb *sit*. Even in the cases where there is no corresponding verb which is morphologically related, there is one that is semantically related to the support verb construction. For instance, in Portuguese *dar as boas vindas* can be paraphrased by a semantically related verb *saudar* or in English, *make a trip* can be semantically related to the verb *travel*, since the verbs **boa-vindar* and **trip* do not exist. Again, even in the cases where there are no verbs which can relate either morphologically and semantically or just semantically to the support verb constructions, there is still usually a way of paraphrasing or simplifying the meaning or making it less idiomatic, if needed or desired.

5.5. **Translation of Support Verb Constructions**

Translation of problematic expressions, such as lexical gaps, "collocations" and idiomatic expressions has been attempted by [Santos, 1990]. Support verb constructions present two key problems to machine translation: semantic weakness and consequent ambiguity. In support verb constructions the predicate noun is the element that carries semantic information and not the verb. However, the same verb can be a support verb or a lexical verb (predicate). Therefore, it is often necessary to obstruct the interpretation of the support verb reading when it should take the lexical verb reading. For example, in *fazer uma casa* (∓ *to make a house*) (*cf.* (28)), *fazer* is a verb that has the meaning of *construir* (*to build*) and can be paraphrased as *construir uma casa* (*to build a house*). In *realizar um filme* (*to produce a film*), *realizar* is a verb equivalent to *produzir* (*to produce*) (*cf.* (29)). In both cases, *fazer* and *realizar* are not support verbs, they are the predicates themselves and, therefore, they do not occur with predicate nouns.

(28) Hoje você pode fazer uma casa no estilo que quiser. [CdP]
=> *Nowadays you can build a house in any style you want.*

(29) Spielberg parte de uma história emocionante para realizar um filme que os americanos, não sem uma boa dose de infantilidade, chamam desde 1.º de janeiro de o filme de o ano. [CP]
=> *Spielberg starts from a moving story to produce a film that Americans, rather childishly, have been calling the movie of the year since January 1st.*

The first problem that a machine is confronted with is whether verbs such as *fazer* and *realizar* are lexical or support verbs and formalize them appropriately. This is called automatic word sense disambiguation (WSD). Verbs are predicates, support verbs are not predicates. It makes all the difference to machine translation quality to identify these distinctions. If the machine already distinguishes support verbs, there is an additional problem. Support verbs are semantically weak. So, even when one ambiguity problem is solved, another one still remains. The weak verb itself can add confusion to the understanding required, where the corresponding weak verb may be different in the target language or when there is a corresponding semantically strong verb that is used instead of the support verb construction. For example, Portuguese and English support verbs differ in the support verb constructions *fazer uma visita* (∓ *to make/do a visit*) and *to pay a visit* (*cf.* (30)), paraphraseable by *to visit* and *colocar uma questão* (∓ *to put a question*) and *to ask a question* (*cf.* (31)).

(30) (…) partindo depois para Obidos onde, cerca das 16 horas deverá fazer uma visita à vila histórica, onde se encontrará com o candidato do partido, (…) [CdP]
=> (…) *leaving then for Óbidos where, around 4 o'clock he will probably pay a visit to the historical village, and meet the party's candidate, (…)*

(31) Registe-se para receber o seu nº de utente e password, que lhe permitirá colocar uma questão para consulta. [CA]
=> *Register in order to get a patient number and a password that will enable you to ask a question for your medical appointment.*

In some cases, support verb constructions in the source language have no corresponding expression in the target language, or if they have an equivalent, that expression is not as commonly used as the one in the source language. For example, the equivalent to the Portuguese support verb construction *pedir/apresentar desculpa* (∓ ?*to ask for/present an apology*) is the English verb *to apologize*, which sounds more natural (*cf.* (32) and (33)). Similarly, the support verb construction *pedir esmola* (∓ ?*to ask for alms*) is more appropriately translated into the verb *to beg* in English (*cf.* (34)).

(32) Virou-se para ela, não para pedir desculpa (Morris Zapp nunca pedia desculpa), mas para lhe lançar o famoso Olhar Zapp, infalível em deixar estarrecida qualquer criatura

humana num raio de vinte metros, desde reitores universitários a Panteras Negras. Mas dá com uma impenetrável cortina de cabelo loiro. [COMPARA]
=> *He turned towards her, not to apologize (Morris Zapp never apologized) but to give her the famous Zapp Stare, guaranteed to stop any human creature, from University Presidents to Black Panthers, dead in his tracks at a range of twenty yards, only to be confronted with an impenetrable curtain of blonde hair.* [COMPARA]

(33) Queremos também apresentar desculpa por só agora vos felicitarmos, de forma tão tardia. [TWICE-BLOG]
=> *We want to apologise for complimenting you only now, after so long.*

(34) Tu vais pedir esmola na cidade. [COMPARA]
=> *You'll be begging all over the town.* [COMPARA]

Support verb constructions are complex and often ambiguous. A support verb construction in the source language does not always translate to a similar support verb construction in the target language. Therefore, it is impossible to carry out a literal translation when translating these expressions. The support verb construction has meaning as a whole and it is necessary to look at the whole expression in order to be able to choose the best translation equivalent. This makes translation of support verb constructions difficult not only for machines, but also for non-native speakers. Support verb constructions that have no literal translation may not only bring additional problems to machine translation, but also to people working in a multinational workforce. Another challenge to the translation of support verb constructions is related to the argument structure of predicates, as seen in § 4.3.4. Machine translation of support verb constructions fails mostly because of the lack of proper identification/recognition of the predicate-argument structure relations by the system, namely the identification of the type of subject. Predicate-argument structure deals with the information of the type "*who does/did what to whom*". In order to disambiguate the meaning of support verb constructions, all these syntactic-semantic roles need to be identified. It happens that sentences with support verb constructions often do not have a clear subject specified and therefore, they are more ambiguous than if another type of predicate was used. It is important to identify subject, object and other arguments for support verb constructions. It is also important to identify the type of subject or object, whatever they are. In some cases, it is not enough to identify the subject as a human, it is necessary to identify it also

as an agent or a patient, a goal, a receiver, etc. For example, if the support verb construction *fazer uma cesariana* (*to perform*/*have a cesarean section*) appears in a text to be translated, it is necessary to teach the machine to translate it correctly into English, according to who is the subject, a health care professional or a health care patient. If the subject is a health care professional, the expression is translated into English as *to perform a cesarean section* (*cf.* (35)); if the subject is a health care patient, the expression is translated into English as *to have a cesarean section* (*cf.* (36)).

(35) Estive em período expulsivo cerca de 1 hora mas o meu filho não desceu e a obstetra de serviço resolveu fazer uma cesariana. [HumPar]
=> *I was in labour for about 1 hour, but my son didn't come down and the obstetrician in service decided to perform a cesarean section.*

(36) Eu adoro o meu GO, mas não quero fazer uma cesariana só porque lhe dá mais jeito. [PinkBlue]
=> *I love my gynecologist, but I don't want to have a cesarean section just because it is more convenient for him.*

A deeper and finer analysis of the support verb construction argument-structure helps natural language processing system understanding.

In addition to the aforementioned problems of support verb constructions for machine translation, there is another setback that needs to be mentioned: wordiness. Most machine translation engines have difficulties in handling words that serve no specific purpose. This is the reason why controlled technical language has been more suited for machine translation. Technical writers omit many irrelevant or unnecessary words. Support verb constructions also use words that add noise to the understanding of the linguistic information by the machine. For example, in *fazer a sessão de abertura* (≠ *to make*/*do the opening session*) (cf. (37)) and in *fazer a sessão de encerramento* (≠ *to make*/*do the closing session*) (cf. (38)) the words *a sessão de* add confusion to the machine, which probably well understands the words *abertura* (*opening*) and *encerramento* (*closing*) because they are listed in its dictionary as process nouns.

(37) Após a indicação dos nomes, a Casa ainda tem três dias para fazer a sessão de abertura. [MP-RGS]

=> *After the nominations have been given, Casa still has three days to do the opening session.*

(38) Renato Sampaio referiu que cabe a Jorge Coelho abrir a iniciativa, a 07 de Setembro, enquanto António Vitorino vai fazer a sessão de encerramento. [DD]
=> *Renato Sampaio mentioned that it was up to Jorge Coelho to open the event, on September 7th, while António Vitorino will do the closing session.*

When translating the support verb construction *fazer a sessão de abertura* and *fazer a sessão de encerramento* into English, a human translator may paraphrase them immediately using verbs instead of nouns, because in the human brain there is a mental and linguistic connection between them. They are both predicates expressing events, one lexically realized as a noun and the other as a verb. At a deeper level than the surface sentence, or what speakers produce, there is an underlying semantic association between the noun *abertura* and the verb *abrir* (*open*), and between the noun *encerramento* and the verb *encerrar* (*close*). These linguistic associations are natural and spontaneous to humans, but machines have to learn them.

Support verb constructions which contain modifiers such as determiners and adjectives for the predicate noun may be more difficult for the machine to understand, unless it has the linguistic knowledge about support verb constructions. For example, in the support verb construction *ir um bocado mais depressa* (*to go a bit faster*) (cf. (39)), the words *um bocado* (∓ **a bit*) are idiomatic and often used in less formal situations for reasons of politeness, but they may confuse the machine. They are not really necessary for the understanding of the expression. If they are removed from the expression, the resulting support verb construction is *ir mais depressa* (*to go faster*), which is shorter but basically the same in essential meaning (cf. (40)). Additionally, this support verb construction when in context can be even further reduced into a verb such as *acelerar* (*to acellerate*) (cf. (41)).

(39) És capaz de ir um bocado mais depressa. Assim vamos chegar atrasados! [Talvezumdia-BLOG]
=> *Can you go a little faster? Otherwise, we will arrive late!*

(40) És capaz de ir mais depressa. Assim vamos chegar atrasados!
=> *Can you go faster? Otherwise, we will arrive late!*

(41) És capaz de acelerar. Assim vamos chegar atrasados!
=> *Can you acellerate? Otherwise, we will arrive late!*

Like determiners and other modifiers, and idiomatic fragments, prepositions also occur frequently in transitive support verb constructions after the predicate noun. These prepositions belong to the support verb construction. They are needed to make a link between the predicate noun and the noun that fills the syntactic position of direct object. For example, in the support verb construction *fazer uma transferência de* (*to make a transfer of*), the preposition *de* (*of*) is necessary to make the link with the direct object complement *dinheiro* (*money*) (cf. (42)), if it is lexically realized. *Fazer uma transferência* (*to do a transfer*) when alone, i.e., without a complement, means normally to transfer money, but in special contexts, such as in the sports domain, it means to transfer players. If the expression is paraphrased by *transferir dinheiro* and translated into *to transfer money*, the predicate noun is converted into a verb with the consequent elimination of the determiner. The preposition is also eliminated because *transferir* (*transfer*) is a strong transitive verb, more specifically a di-transitive verb (cf. (43)). Di-transitive verbs take both direct and indirect objects and are distributed according to their argument structure. *Transferir* is similar to *mandar* ou *enviar* (*to send*) and the argument structure is [Nhum + V + DO + IO] where the direct object is *dinheiro* (*money*), any other mass financial type of noun among others and the indirect object is a person or a place, preceded by a preposition of a certain type, in this specific case, the preposition *para* (*to*).

(42) Preciso fazer uma transferência de dinheiro para um conhecido que vive na Inglaterra e ele pediu que fizesse uma transferência swift. O que é isso? [YahooR]
=> *I need to make a transfer of money to a relative who lives in England and he asked me to make a swift transfer. What is that?*

(43) Preciso transferir dinheiro para um conhecido que vive na Inglaterra e ele pediu que fizesse uma transferência swift. O que é isso?
=> *I need to transfer money to a relative who lives in England and he asked me to make a swift transfer. What is that?*

In sum, support verb constructions are difficult to translate for several reasons. Support verb constructions are not exclusivelly lexical, they exhibit syntactic-semantic behavior

that distinguishes them from common lexical phrases. They exceed lexical units in complexity, by presenting multiple levels of variability. Some of them can be literally translated, viz. *fazer dinheiro* => *to make money*. But, most have no literal translation, as in Figure 2.

fazer adiamentos => to give extensions
fazer perguntas => to ask questions
fazer o exame => to take the exam
fazer uma visita => to pay a visit
fazer uma operação/cirurgia => to have an operation/surgery

Figure 2: Support verb constructions where the support verb cannot be translated literally

Because support verbs in source and target languages differ, it is very difficult to know in each particular case which is the adequate support verb. Some Portuguese support verb constructions can only be translated into English by a verb that conveys the same idea. Others can be translated literally, but the corresponding verb sounds more natural. Figure 3 illustrates some of the examples where a Portuguese support verb construction is translated into an English verb.

fazer publicidade => to advertise
fazer um esboço => to draft
fazer compras => to shop
fazer a barba => to shave
estar presente/ausente = attend/not attend

Figure 3: Support verb constructions translated naturally into verbs

When teaching the machine how to translate support verb constructions, it is important to program the machine to use the right support verb construction in the right circumstances in the target language. It is the job of the computational linguist to help the machine make the appropriate choices. Linguistic rules may help specify in which circumstances *tomar* is equivalent to *have* or *fazer* is equivalent to *take*, in English. For example, the Portuguese support verb *tomar* is translated into the English support verb

have when followed by any kind of drink (beverages, alcoholic drinks, tea, coffee, etc.) and some kinds of food, as the syntactic-semantic rules of Figure 4 illustrate.

PT: <tomar> N(MA+liqu) = <beber> N(MA+liqu)

EN: <have> N(MA+liqu) = <drink> N(MA+liqu)

PT: <fazer> comida

EN: <make> food = <prepare> food

Figure 4: Syntactic-semantic rules

If these phrases represent challenges to human translators, they are no less of a problem for machine translation. Support verb constructions can offer substitutions that improve translation precision and quality. They are valuable to automatic evaluation of machine translation. Linguistic knowledge of support verb construction paraphrases is important for detail and precision reasons. It is the *raison d'être* of a refined language application, where effective paraphrasing is a synonym for intelligent paraphrasing, that eliminates what is unnecessary, and augments the possibilities of success. In our work, we seek to obtain effective paraphrasing by application of linguistic resources that generate paraphrases with equivalent meaning permitting acceptable replacement of the source phrase.

PART TWO

Natural Language Processing Applications:

State of the Art

Our Resources, Methodology, and Paraphrasing Software Tools

Chapter 6

Approaches to Machine Translation and Paraphrasing

*

Chapter Six presents the main approaches to machine translation and paraphrasing in natural language processing. The grammar-based, example-based, statistical and hybrid approaches are described. Similar approaches are used to understand paraphrasing and the characteristics of three important methods for automatic acquisition of paraphrases are discussed. The ideal machine translation is also debated.

*

6.1. **Machine Translation Approaches**

The main approaches to machine translation are the rule or grammar-based, the example-based and the statistical approaches. Until recently, the prevailing trend in machine translation research was statistical (data-driven approach). Data-driven machine translation uses several types of corpora. Some systems use non-parallel monolingual corpora to generate dictionaries. Others use parallel monolingual corpora, i.e., aligned translations of the same source texts, to create lists of synonyms and paraphrases. Others use parallel bilingual corpora to achieve translation equivalents. However, corpora dependency of data-driven machine translation systems makes machine translation accuracy reliant on the quantity, appropriateness, availability and quality of the selected corpora. For languages like English, or which corpora are easy to find, data-driven systems have succeeded in finding good corpora and have a consistent performance. However, many languages lack the kind of corpora resources that English has.

Rule-based machine translation, which is not corpora dependent, was a leader until the nineties, but after that it has become less widespread. In an attempt to move faster, human resources and investments have been directed towards the development and improvement of machine learning techniques, and probabilistic calculations, and there has been less focus on the development of enriched dictionaries, grammars and other linguistic tools or linguistic-driven methodologies. Researchers and developers now

realize that contributions from all quarters are required to advance and make good machine translation a reality. The synergy of combining approaches is becoming productive. Linguistically driven or rule-based systems are now using parallel corpora and automated tools (n-grams, etc.) to enlarge their dictionaries and grammars [Wu & Wang, 2004], and creating methods to accelerate the time-consuming disambiguation process or to help with the post-editing [Dugast et al., 2007]. Hybrid approaches to machine translation are being used, and this is believed to be the most productive way to progress [Imamura et al. 2004] [Eisele et al., 2008] [Crego & Habash, 2008]. There were endeavors to promote mixed approaches at the MATMT 2008 workshop. The goal of the event was to gather resources and algorithms from the three major approaches. Sections § 6.1.1 to § 6.1.4 describe the different methods of the machine translation approaches mentioned above.

6.1.1. **Rule and Grammar-based Machine Translation**

Rule-based machine translation (RBMT) comprises paradigms such as transfer-based machine translation, interlingual machine translation and dictionary-based machine translation. Grammar based machine translation has been built on syntactic theories that have been often supported by a less than extensive lexicon. Once the system is faced with unfamiliar words, or "real" sentences and texts, it makes mistakes because it fails to build a comprehensive enough system to draw on Halliday's basic notion of the interrelationship and interdependence of all the levels of language. Halliday's theory [Halliday, 1985a] [Halliday, 1985b] [Halliday, 1985c], which developed out of Firth's work [Firth, 1957] [Firth, 1968], runs parallel to the structural and more traditional syntactic theories that have proved inadequate or insufficient to represent language when translating. The central idea in Halliday's approach is the "context of situation" which obtains "through a systematic relationship between the social environment on the one hand, and the functional organization of language on the other" [Halliday, 1985c: 11].

One criticism that has been made is that "Linguistic theories have rarely addressed questions of contrastive linguistics, i.e., the way in which different languages use different means to express similar meanings and intentions. Such questions are of course at the heart of MT." [Hutchins & Somers, 1992: 82]. It is important to emphasize contrastive linguistic analysis in machine translation, and, on the other hand, it is necessary to look at

empirical results coming from existing machine translation systems, to look at the errors and learn from them.

Grammar-based systems learn from the addition of formalized rules and can apply those rules independently of the input text domain. In grammar-based systems, the development cycle is as follows: the linguist analyses the output, diagnoses rules requirements, writes new rules, determines the impact of changes, decides on rules to implement, implements rules, tests and deploys them.

6.1.2. **Example-based Machine Translation**

Example-based machine translation (EBMT) was suggested by [Nagao, 1984], and captivated the attention of many researchers in natural language processing. It uses bilingual parallel corpora as the basis to acquire knowledge of the process of translation. It is a machine learning approach that translates by analogy – translation of expressions is learned by looking at previously translated texts (case-based reasoning). The system is trained using fragments of text that have been previously translated. Corpus-based extraction techniques integrate sentence alignment and mapping tools, such as Giza++, for word and phrase mapping. These tools can align source text originals with their translations to create example-based machine translation. This approach to translation uses concrete examples of language usage in both the source text and in the translations. They may consist of a bilingual dictionary, lists of expressions or simple structures, whose paraphrasing capabilities in the other language are registered in a parallel text.

Example-based machine translation can integrate translation memories. Similar to bilingual parallel corpora (but never annotated), translation memories consist of text segments (sentences or sentence parts) and their translations. Translation memory tools are used by translators to speed up translation and check translation consistency automatically. Any elements that have already been translated can be reused.

There are shortcomings to the example-based approach. Because translated texts are likely to contain errors, extracting paraphrases from parallel corpora results in variance that may go from identical paraphrases to quasi-paraphrases or even for non-paraphrase pairs. We see that, for instance the predicate verb *to see* occurs in COMPARA parallel corpora [Frankenberg-Garcia & Santos, 2003] [Santos & Inácio, 2006] as a paraphrase of the predicate verb construction *to have a look at*, even though it is not the same as the

verb *to look* or, in Portuguese, as the verb *ver*. While *to see* is not usually a voluntary action, the other expressions are.

The disadvantages of parallel corpora (either monolingual or bilingual) are that they are difficult to find, and when they exist, they may be specific to a certain subject matter and may not be of good quality. Even if they are a scarce resource, researchers use them for cross-language retrieval, mining terms for translation and machine translation, among other applications.

Corpus-based methods for developing language tools are useful. Language applications have gained value from adopting techniques which use corpora to extract linguistic knowledge. However, corpora have limitations. Particularly for machine translation purposes, parallel literary corpora may be inappropriate if not used properly. In literary texts, translators often use free translation, which is a translation that preserves the meaning of the original, but involves frequent re-structuring of the syntax and a more flexible use of the lexicon to deliver a natural articulation of the forms of the target language. The result is a text that sounds "smooth". Free translation is often idiomatic. When paraphrasing to another language, even professional human translators can make errors, *errare humanum est*. In addition, there are texts too difficult for a particular translator. In the worst case, the phrase *tradutore traditore est* (the translator is a traitor) is applicable. Parallel corpora contain mistakes resulting from lexical variation or inappropriate use of the lexicon, which converts into different semantics and unsuitable translations. Can machine translation rely on inadequate corpora to provide translations? Which corpora (if any) are trustworthy for good translation? Not all corpora are suitable and sometimes a revelant or representative corpus is unavailable. In addition, a 'representative' corpus has to be representative of a particular genre, style, domain, variety, or whatever, of language. The scope of language is too vast to be encompassed by a single corpus. Not even the Internet with its almost limitless body of text can be representative of language usage.

6.1.3. **Statistical Machine Translation**

Statistical machine translation (SMT) is a machine translation approach where translations are generated on the basis of statistical models. The first ideas of statistical machine translation were introduced by Warren Weaver in 1949 [Weaver, 1955], and re-

introduced in 1991 by researchers at IBM's Thomas J. Watson Research Center. There has been a recent, significant resurgence in interest in machine translation, using the widely studied machine translation paradigm and located where there is substantial funding.

As with the example-based machine translation, statistical machine translation also learns translation from bilingual text corpora. Statistical-based machine translation is a trial-and-error based interpretation of translation, driven by knowledge gained from the text input. Statistical machine translation systems check the frequency with which a certain translation equivalent is used. In statistical based systems, the statistician looks for correlation between input and output, determines correlation, determines learning parameters, performs impact analysis of change, and implements the change. Statistical machine translation divides the task of translation into two steps: a word-level translation and word reordering during the translation process.

There are different types of statistical approach system. Most statistical machine translation models are trained on parallel corpora. Some systems use non-parallel corpora [Rapp, 1995] [Fung, 1995] [Koehn & Knight, 2002] [Nazar et al., 2008] to automatically generate one-to-one mapping of words in different languages. The systems that use monolingual corpora identify cognates in two corpora of two different languages and create a kernel (seed) dictionary. Beginning with such a kernel, they use clues such as context and frequency to identify potential translation pairs, multiword expressions, etc. Alternative systems have bilingual dictionaries and non-parallel corpora to learn translations of unknown words and thus generate parallel corpora. For instance, [Schulz et al., 2004], created an automated Spanish medical lexicon from an already existing Portuguese seed lexicon. They use a cognate mapping method based on heuristics to find Spanish lexeme candidates as well as Spanish and Portuguese non-parallel monolingual corpora and validate the translation hypotheses by "determining the similarity of fixed-window context vectors on the basis of Portuguese and Spanish text corpora".

Statistical based machine translation systems need substantial amounts of domain-specific text on which to learn, i.e. they need to train the statistical model's parameters. Furthermore, they need good translators to determine the best translations of the texts used to train the statistical models' parameters. Nowadays, there are statistical machine translation models augmented with morphology, syntax or semantics. However, statistical methods perform poorly when they translate text that is different in genre than

the genre they were trained on. They are not successful at long distance dependencies, obscure words, obscure linguistic phenomena, and co-reference.

6.1.4. **Hybrid Approaches**

The "preservation of meaning" remains the first and most difficult challenge for any machine translation model. An interesting issue in comparing systems is to examine how each model treats meaning. Some machine translation approaches are more suitable for the translation of lexical units. They produce literal translations based on direct lexical correspondence. The more lexical approaches need to handle complementation with a sound linguistic basis, mainly syntactic-semantic analysis. These approaches need syntactic modules that specify rules, which integrate words in adequate syntactic structures and embrace coherent meaning. The models need to be analyzed to determine the effective end-state for the fully developed performance of the system. When the evaluation uses documents from a different domain, the grammar-based machine translation system translates far better than the statistical machine translation system, which would need training data using domain based texts to bring it up to speed.

Grammar-based and statistical approaches can be (and in some cases already are) combined to provide both an optimized solution to speed up translation and to save time by optimizing the system. Mathematical algorithms and linguistic knowledge such as morphology, syntax and semantics are being combined for use by statistical machine translation models [Melamed, 2004] [Quirk & Corston-Oliver, 2006] [Menezes & Quirk, 2008] [Cherry & Quirk, 2008]. For example, [Melamed, 2004] introduces an automated MT evaluation method that accounts for paraphrases which he calls "meaning-preserving syntactic alternations" or "translational equivalence". He applies specialized algorithms that estimate n-gram language models to function as generalized parsers to perform syntax-aware statistical machine translation (translation by parsing).

It is possible that synergistic effects may be obtained from the combination of grammar-based machine translation models with various other models, as for example the lexical priming theory. [Hoey, 2005] argues that humans internalize certain lexical blocks and a certain semantic prosody and that this allows us to process language as fast as we do - that and the fact that our listener shares much of this information. If our brains had to process everything syntactically first, our speech production would be much

slower. So, by analogy, we aim to create machine translation systems which integrate such lexical blocks.

6.2. **Techniques of Paraphrasing in Natural Language Processing**

Given the relevance of paraphrases, some of today's systems use automated methods to collect paraphrases. In recent years, paraphrasing has become an area of increasing interest in natural language processing. The benefits of paraphrasal knowledge to natural language processing have been quantified in areas such as summarization [McKeown et al., 2002], [Barzilay, 2001], [Barzilay, 2003], [Hirao et al., 2004] [Zhou et al., 2006b], question answering [Paşca, 2003], [Duboué & Chu-Carroll, 2006], information extraction [Ibrahim et al., 2003], [Shinyama et al., 2002] [Shinyama & Sekine, 2003] [Sekine, 2005], and machine translation [Zhou et al., 2006], [Callison-Burch et al., 2006a], [Callison-Burch et al., 2006b], [Callison-Burch, 2007], [Callison-Burch, 2008], among others. Recent workshops dedicated exclusively to paraphrasing reveal the growth in this field of knowledge. Paraphrases were discussed at NLPRS-2001, ACL-2003 and IJCNLP-2005 workshops, at the workshop on Empirical Modeling of Semantic Equivalence and Entailment (at ACL-2005), and at 2005, 2006, and 2007 PASCAL Recognizing Textual Entailment Challenges.

Paraphrases are being used in many natural language processing applications for a variety of purposes. Paraphrasal knowledge plays a very important role in interpretation and generation of natural language. In *natural language interpretation*, dynamic semantics and identical parses resulting from paraphrases are important to successful applications. In *natural language generation*, the generation of paraphrases allows more varied and fluent text to be produced [Iordanskaja et al. 1991]. In *multi-document summarization*, the identification of paraphrases allows information across documents to be condensed [McKeown et al., 2002] and helps improve the quality of the generated summaries [Hirao et al., 2004]. In *question answering*, discovering paraphrased answers may provide additional evidence that an answer is correct [Ibrahim et al., 2003], and paraphrases can be useful in text mining, preventing a passage being discarded due to the inability to match a question phrase deemed as very important [Paşca, 2005] [Paşca & Dienes, 2005]. In *information extraction*, paraphrases help text categorization tasks or

mapping to texts with similar characteristics, lessening the disparity in the trigger word or the applicable extraction pattern [Shinyama & Sekine, 2005].

Phrasal and sentence transformation can be achieved by means of different types of linguistic resources. Paraphrases can be obtained manually, semi-automatically, and statistically, often by using corpora resources. Lexical resources, general language dictionaries and ontologies, such as WordNet [Miller et al., 1990] [Miller, 1995] [Miller et al., 1995] [Fellbaum, 1998] are relevant sources of knowledge for paraphrasing. There are three generally accepted methods for paraphrase acquisition: corpora-based, statistical-based and dictionary-based. We will describe different paraphrase resources and techniques in § 6.2. The works we present below show that the study of paraphrases plays an important role in natural language processing in general. Paraphrases are a valuable resource for many applications.

6.2.1. **Corpora-based Paraphrases**

Until now, the most popular way of obtaining paraphrases was by using the corpus-based method, where corpora are used to train statistical models for paraphrasing. For instance, [Barzilay, 2001], [Bannard & Callison-Burch, 2005], and [Pang et al., 2003] use parallel corpora and statistical-based methods to achieve paraphrases while [Lin & Pantel, 2001] and [Bhagat & Ravichandran, 2008] use non-parallel corpora. [Shinyama & Sekine, 2003] and [Shinyama & Sekine, 2005] use comparable corpora. [Ibrahim et al., 2003] extract paraphrases from aligned monolingual corpora. [Barzilay & McKeown, 2001] introduce a corpus-based method to extract paraphrases automatically by using multiple parallel English translations of novels. [Dolan et al., 2004] explore techniques for automated acquisition of monolingual paraphrases by using techniques such as distance and heuristic strategy of paralleling, evaluating them through a word alignment algorithm and metrics used in machine translation. [Poibeau, 2004] presents an algorithm to collect paraphrasal constructions from the corpus semi-automatically, in conjunction with a semantic net. [Pang et al., 2003] carried out similar work, using finite state automata to extract lexical and syntactic paraphrase pairs and to generate new, unseen sentences that express the same meaning as the sentences in the input sets. They also used them to predict the correctness of alternative semantic renderings. These finite state automata are built automatically from a syntax-based algorithm containing semantically equivalent

translation sets and may be used to evaluate the quality of translations. In the same line of research, [Barzilay & Lee, 2003] address the text-to-text generation problem of sentence-level paraphrasing, applying multiple-sequence alignment to sentences gathered from unannotated comparable corpora. Their algorithm learns a set of paraphrasing patterns represented by word lattice pairs and automatically determines how to apply these patterns to write new sentences.

6.2.2. **Statistical-based Paraphrases**

Paraphrases can also be found by means of statistical methods, i.e., based on word co-occurrences and word combinations. The statistical methods to acquire paraphrases have little or no linguistic knowledge. They use sophisticated algorithms and apply these algorithms to corpora. In most statistical methods, monolingual paraphrase acquisition is the most common process. The mapping of the monolingual paraphrases to the bilingual ones is executed separately and afterwards. However, other statistical methods may be used.

Statistical methods use tools such as n-grams to find combinatorial sequences of words. Tools that analyze and extract multiword expressions based on linguistic knowledge are still primitive [Maia et al., 2007]. For instance, BACO [Sarmento, 2006] is a database of Portuguese text and co-occurrences that uses n-gram tables. It claims to find semantically related words. Tools like these can be used for automated generation of paraphrases. Extraction of lexical units, unless idiomatic, can be complicated.

At present, one problem with statistical methods of finding paraphrases is that they do not produce clean data. N-grams and other statistical tools produce errors and therefore they cannot be reliable. They are also not robust, since they need a training corpus that is very close to each text. Linguists recognize that they may be helpful in finding empirical data to validate and help create better and more comprehensive linguistic rules. They may be useful to help speed up the process of linguistic annotation. They may work for some frozen expressions, but in general, they do not work for multiword expressions with a more flexible structure. They do not work to obtain other complex lexical information, such as disambiguation of a transitive versus an intransitive verb structure.

6.2.3. **Dictionary-based Paraphrases**

Dictionary-based methods of finding paraphrases are less popular because they require more linguistic knowledge and they take longer to process and retrieve results. WordNet [Fellbaum, 1998] [Green et al., 2004] and NOMLEX [Macleod et al., 2000] are examples of dictionary projects for the English language, now being used as a basis for new linguistic and natural language processing tools. Dictionary-based methods referenced above are only monolingual. They relate more to language analysis than to language generation.

We argue that dictionary-based methods work exceptionally well for lexically related paraphrases. For example, many support verb constructions match with verb counterparts, expressing exactly the same meaning. The support verb construction *dar um abraço a* (*to give a hug to*) is equivalent to the verb *abraçar* (*to hug*). However, many paraphrases use words that do not have the same lexical root as the ones in the original sentence. Chapter 7 describes how support verb constructions that are not morphologically derived can be linked to semantically related verbs.

6.2.4. **Related Techniques**

The following techniques are those which are closest in nature to those used in this research. Paraphrasing techniques such as those presented in [Yamamoto, 2002] and [Fujita et al., 2004] and source-side reordering strategies described, among others, by [Xia & McCord, 2004] and [Collins et al., 2005] proved extremely useful to natural language processing. The idea of using paraphrasing to improve the quality of machine translation has also been discussed. The notion is motivated by the fact that the source sentence represents only one of several possibilities in which the meaning could have been expressed, and some of those other possibilities might be easier to translate. For example, [Madnani et al., 2007] and [Dyer et al., 2008] represent source language alternatives in lattices and let a decoder decipher which choice is better. In a similar approach to what [Fujita et al., 2004] did for Japanese, in our research, the dictionary was manually enhanced with information to improve the paraphrasing process. Many support verb constructions were transformed into verbs, leading to the possibility of the latter being more straightforwardly translated. Similar work was done for Chinese-English

translation [Wang et al., 2007] by using manually constructed rewriting rules operating at the syntactic level so as to make the translations better.

6.3. **The Ideal Machine Translation**

Machine translation has made some progress recently, but machine translation trends have been focusing on the power of statistics more than on the power of linguistics. The value of statistical tools is real; it has helped natural language processing to advance, it has helped create resources faster and more efficiently, comprising elements like coverage, speed and low cost. The fact that a software machine can learn and follow instructions quickly and objectively makes machine translation a valuable asset and with certain advantages in comparison to human translation, especially in machine translation compatible texts. However, we argue, and prove with support verb construction phenomena, that linguistics can add significant contributions to the improvement of machine translation quality and anticipate that future models will include more linguistic knowledge.

The ideal machine translation system will resolve ambiguity, complexity issues and their interaction. The Logos system was among the systems that incorporated solutions to these problems based on models of human sentence processing [Scott, 1989] [Scott, 2003]. The result was a machine translation model that was commercially successful. Adopting a similar approach, experiments on Chapter 9 prove our hypothesis that using paraphrase to simplify the source text makes machine translation more likely to work. There are two reasons why paraphrase helps: (i) shorter passages are less likely to introduce errors; (ii) paraphrase alleviates the sparse data problem because it simulates a situation in which there are fewer distinct cases and each case is more frequent - this helps statistical systems in particular. We argue that the ideal path is the interpretation of complex and sometimes ambiguous syntactic-semantic expressions into simple linguistic units. Humans can produce and understand long messages, but sometimes they also use or need to perform mental interpretative tasks (strategies of paraphrasing, reduction, etc.) when they do not understand other people's language. It is more difficult to program this type of language into a machine. However, if the machine is programed with the capability of linking semantically-related linguistic objects, similarly to the operations performed by human translators, the machine translation system is able to provide more

than one correct translation for the same expression. Less sophisticated machine translation software have problems with the longer and more complex sentences containing support verb constructions, but more sophisticated machines will understand and handle more than one possibility (the support verb constructions and their paraphrases).

In order to judge machine translation fairly, machine translation should be assessed according to where it lies on the graph of Figure 5.

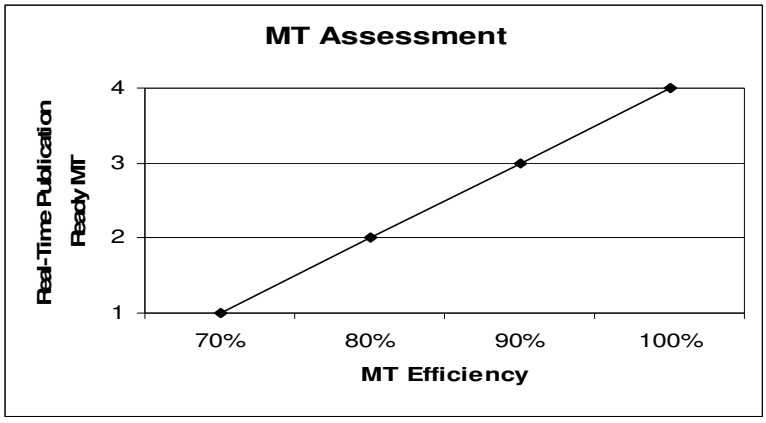

Figure 5: Graph for machine translation assessment

In this graph the numbers are arbitrary and we are working on assumptions only. Let us assume that 70% corresponds to the ability to do gisting, 80% to unedited translation for controlled language, 90% to a situation where the user is able to complete the job easily by post-editing, and 100% stands for 'perfect' machine translation. It remains to be seen if 100% machine translation efficiency will ever be possible, i.e., if the end-state of a perfectly, fully developed system is able to perform translation as well as a human translator can. It is too early to expect or believe it is. However, it is almost certain that in the future, machine translation models will have to combine the best features of each existing model, and focus more on grammar and grammar formalisms and linguistic rule representation.

Chapter 7

Automated Processing of Support Verb Constructions

*

Chapter Seven presents the linguistic resources, tools and methodology used in processing support verb constructions automatically, in generating paraphrases and in implementing them. A brief description of the original sources is presented. Lexical resources were converted from the OpenLogos system, and the NooJ linguistic environment was the linguistic platform that combined with Port4NooJ dictionaries and grammars created the paraphrases of support verb constructions. The enhanced electronic dictionary development process is described and the new features demonstrated. Identification and tagging of elements, such as predicate nouns and predicate adjectives and the application of the paraphrase resources to texts are included. The advantage of support verb construction paraphrases for machine translation is also demonstrated.

*

7.1. **Original Sources**

The Port4NooJ natural language processing system was developed using two original sources: the *NooJ* linguistic environment and *OpenLogos* lexical resources. The NooJ platform has been available for a number of languages with extensive resources, including large coverage dictionaries and grammars for each individual language. Since there were Portuguese resources missing in this platform, we needed to build a Portuguese system from the ground up. *OpenLogos*, an open source derivative of the *Logos Machine Translation System* [Scott, 1989] [Scott, 2003], became available at the beginning of 2006 (http://logos-os.dfki.de/). The Logos system's strength resides in its ontology (SAL) and lexical resources, the bilingual dictionaries and the semantic-syntactic rules (Semtab). The technology itself is dated, but the object-oriented character of Logos rules lends itself to NooJ, and the data from *OpenLogos* is useful in launching development of new language pairs.

As described in [Barreiro, 2008b], knowledge of the Logos system and experience in creating dictionaries and grammars was used to extract and select data from *OpenLogos* for innovative research. A new logical linguistic model was created based on a synergistic approach, where the components of the Logos system that offered valuable, functional abilities are maintained and integrated in a flexible platform for further development. NooJ is linguist friendly and robust enough to absorb the Logos system and make use of its best features. The major issue in re-using the Logos data concerns the grammar rule base, namely RES-PARSE, TRAN, and target generation rules. RES and PARSE are two sets of parsing rules. RES concerns the macro-parse of the input sentence and PARSE is about the micro-parse. TRAN rules are the transfer rules. Target generation rules are related to the output in the target language. In contrast to the lexical resources and the Semtab rules, grammar rules are inaccessible, due to a lack of comprehensive documentation. Grammar rules need to be built from the beginning. NooJ's *modus operandi* is suitable to leverage the linguistic phenomena that Logos grammar rules held, such as word order, syntactic transformations, inserts, etc. that are indispensable to the translation process and generation of the target languages.

7.1.1. **NooJ**

NooJ [Silberztein, 2004] is a freeware development environment for linguistic research and development. NooJ contains several modules that include large coverage lexical resources, dictionaries for specific purposes and local grammars for a dozen languages, and it is being extended to several other languages. Its tools support the development, testing, debugging, maintenance and gathering of other different types of linguistic resources, namely local grammars, and they assist the development of natural language processing applications. NooJ tools are also used to parse corpora, build sophisticated concordances, and apply their linguistic resources to texts for distinct purposes. Local grammars are language descriptions in the form of graphs containing an input entry (with linguistic information) and output entry (with linguistic constraints to the output, or simply the binary information of the recognized or unrecognized sequence). In NooJ, these local grammars are represented by finite-state transducers, called (Extended) Recursive Transition Networks (RTNs), and are widely applied to texts. They are used for identification and analysis of local linguistic phenomena, such as the:

(i) location and annotation of morphological, lexical and syntactic-semantic patterns;
(ii) identification and extraction of semantic units from texts, such as dates, named entities and terminological expressions;
(iii) recognition and tagging of words or multiword expressions;
(iv) identification of syntactic constituents such as noun phrases and other syntactic constituents;
(v) extraction of semantic relations, and disambiguation.

Among these feasible applications, specific local grammars were created to recognize, paraphrase and translate support verb constructions, such as *tomar uma decisão* > *make a decision*. In future developments, the transformational aspects of translation will need to be explored and the capability of the local grammars will need to be expanded to larger linguistic operations, indispensible for machine translation.

7.1.2. **OpenLogos**

OpenLogos source data, dictionary and rules employ a classification based on the so-called SAL ontology. SAL stands for Semantic-syntactic Abstract Language, a representation language, embodying both meaning (semantics), and structure (syntax). It is an interlingual-style hierarchical taxonomy comprising over 1,000 elements, distributed in supersets, sets and subsets, which are embedded in the dictionary. It was designed in a way so that developers would expand and add to its capabilities (extensible system). It was initially developed for the English language, but most of its elements are universal and therefore applicable to Portuguese and other languages. Unlike other ontologies, it places semantics and syntax on a continuum. It may be not totally original, but it is eclectic in the categories included in the representation schema. Notwithstanding acceptable shortcomings, this ontology was designed to work in combination with other linguistic resources, namely lexical resources and a diverse set of linguistic rules, and it has already been used successfully for several decades in commercial machine translation. This is reason enough for it to be used by other systems. Furthermore, the abstraction echelon makes the ontology applicable at several levels and useful for applications other than

machine translation. SAL is a good basis from which to work, and it can be enhanced and enlarged and be used similarly to the way other researchers have used Wordnet.

OpenLogos system represents an immense original investment and a serious work effort expended over thirty years and it has much to offer. The opportunity to use the linguistic knowledge and intellectual hard work contained in OpenLogos should not be wasted. However, for the open source ontology to work, there needs to be a standardization process, so that the cooperative project will succeed. The toolset and the linguistic resources to enable the creation of more effective and coherent machine translation systems are available. By using open source technology, we hope to improve the system used in the present work cooperatively, to improve and extend the SAL ontology and further develop grammar strategies not only for the language pairs already available in this system, but for new language pairs or simply for single language analysis.

The augmented linguistic resources to identify support verb constructions and automatically paraphrase them will be described in the next section.

7.2. **Augmented Linguistic Resources**

In order to process support verb constructions, Port4NooJ dictionary was enhanced with extended features and lexicon grammar annotations were added. Beyond the commonly used part-of-speech and inflectional paradigm, each dictionary entry includes a description of the syntactic and semantic attributes (*SynSem*), as well as the associated distributional and transformational properties, such as predicate arguments, support verbs, aspectual verbs, stylistic variants of elementary support verbs, information about which determiners and prepositions occur with predicate nouns in "less variable" expressions, and derivational descriptions. Derivation is a very important issue, because it has implications not only at the lexical level, but also at the syntactic level. Derivational suffixes often apply to words of one syntactic category and change them into words of another syntactic category, while semantically they maintain their integrity. For example, the affix *–ção* changes the verb *adaptar* (*to adapt*) into the noun *adaptação* (*adaptation*) and the affix *-mente* changes the adjective *literal* (*literal*) into the adverb *literalmente* (*literally*). This is extremely important for support verb constructions because it permits the establishment of paraphrasing capability grammars that map (i) support verb constructions such as *fazer uma adaptação (de)* (*to make an adaptation (of)*) to the verb

adaptar (*to adapt*), where the predicate noun *adaptação* (*adaptation*) has a semantic and morpho-syntactic relationship with the verb *adaptar* (*to adapt*) or (ii) support verb constructions such as *ter uma dilatação rápida* (Ŧ ?*to have a quick dilation*) to the verbal expression *dilatar rapidamente* (*to dilate quickly*), where the autonomous predicate noun *dilatação* (*dilation*) has a semantic relationship with the verb *dilatar* (*to dilate*), and the adverb *rapidamente* (*quickly*) has a semantic and morpho-syntactic relationship with the adjective *rápida* (*quick*). Thus, our verb entries contain the identification of derivational paradigms for nominalizations (annotation *NDRV*) and a link to the derived noun's support verbs (annotation *VSUP*), as in Figure 6 below. Nominalizations are followed by their inflectional paradigm properties. For example, for the Portuguese verb entry *adaptar* (*to adapt*), derivation table [*DRV=DRV00*] produces the noun/nominalization *adaptação* (*adaptation*) that inflects according to the inflectional paradigm [*:CANÇÃO*] and can be combined with the support verb *fazer* (*to make*) [VSUP=fazer] in the support verb construction *fazer uma adaptação* (*to make an adaptation*). Other lexical constraints, such as prepositions, determiners, specific arguments, etc., will be added. Autonomous predicate nouns (non-nominalizations), such as *favor* (*favor*) are lemmatized and classified with the annotation *Npred* and have associated with them support verb and other lexical constraints, such as a preposition (*NPrep*), and a verb (*VRB*) with the same semantics. For example, the autonomous predicate noun *favor* (*favor*) is linked to the verb *ajudar* (*to help*), to allow the paraphrasing capability between [*Nhum fazer um favor a Nhum*] and [*Nhum ajudar Nhum*] (*Nhum to do a favor to NHum = Nhum to help Nhum*). Predicate adjectives were classified and the link between them and the corresponding verbs (*ADRV*) was established, such as between the verb *adoçar* (*to sweeten*) and the adjective *doce* (*sweet*). The assignment of corresponding support verbs to these adjectives has been started, as well as the identification of derivational paradigms for adverbializations (annotation *AVDRV*). For example, derivation table [DRV=AVDRV05] in the entry for *literal* (*literal*) means that the adverb *literalmente* derives from the adjective *literal* (derivation rule number 05) and inflects according to the inflectional paradigm [:RAPIDAMENTE]. Stylistic variants of the support verb constructions are annotated as *VSTYLE*. For example, *realizar* and *efectuar* (*to perform*) are stylistic variants of the support verb fazer (*to do/make*) in the entry for *transplantar* (*to transplant*). Aspectual variants are annotated as *VASP*. For example, *iniciar* (*to begin*), *prosseguir* (*to continue*)

and *concluir* (*to finish*) are aspectual variants of *fazer* (*to do/make*) in the support verb *fazer um transplante* (*to do a transplantation*). The insertion of aspectual verbs was based on similar research carried out for the English Proteus Project group at New York University, described in [Macleod et al., 2000] and [Meyers et al., 2004a], among others. However, the assignment of the aspectual properties was random and experimental. Motivated by the Proteus Project dictionary work, syntactic and semantic arguments of a predicate started to be assigned to each verb entry. For example, in the entry for the verb *transplantar* (*to transplant*), the property *SUBJ=AG* means that a verb selects an agent as its semantic argument in the syntactic position of the subject. *SUBJ=PAT* means that a verb selects a patient as its semantic argument in the syntactic position of the subject. Syntactic argument *DO=ORG* means that the predicate selects a direct object that is an organ (subclass of body part). *IO=PAT* means that the predicate selects an indirect object that is a patient. *PrepN=de* means that the support verb plus predicate noun construction selects the preposition *de* (*fazer um transplante de – to do a transplant of*). Nouns are classified semantically. For example, the noun *médico* (*doctor/physician*) is classified as an animate being, denoting a profession or other human designation (*AN+des*), belonging to the medical field (*med*).

adaptar,V+FLX=FALAR+Aux=1+INOP57+Subset132+EN=adapt+*VSUP=fazer*+*DRV*=NDRV00:CANÇÃO +*NPrep*=de

favor,N+FLX=MAR+Npred+AB+state+EN=favor+*VSUP*=fazer+*NPrep*=a+*VRB*=ajudar

literal,A+FLX=IGUAL+IN+symb+EN=literal+*DRV*=AVDRV05:RAPIDAMENTE

adoçar,V+FLX=COMEÇAR+Aux=1+OBJTRundif75+Subset604+EN=sweeten+*DRV*=ADRV11:VERDE+*VCOP*=tornar

transplantar,V+FLX=FALAR+Aux=1+RECTR26+Subset=504+BioMed+EN=transplant+SUBJ=AG+VSUP=fazer
+DRV=NDRV79:ANO+NPrep=de+DO=BP+IO=PAT+*VSTYLE=sofrer*+*VSTYLE=realizar*+*VSTYLE=efectuar*
+*VASP=iniciar*+*VASP=prosseguir*+*VASP=concluir*

médico,N+FLX=ANO+*AN+des+med*+EN=doctor

médico,N+FLX=ANO+*AN+des+med*+EN=physician

Figure 6: Sample of the broad coverage dictionary

Figure 7 illustrates dictionary verb entries that are linked to support verb constructions with the support verb *passar*. For example, *passar revista a* (∓ **to pass a search*; *to make a search*) is equivalent to *revistar* (*to search*), *passar a escova por* (∓ **to pass the brush in*) to *escovar* (to brush), *passar o pente em* (∓ **to pass the comb in*) to *pentear* (*to comb*), *passar graxa em* (∓ **to pass polish in*) to *engraxar* (*to polish*), *passar tinta em* (∓ **to pass*

paint in) to *pintar* (*to paint*) and *passar a ferro* or more sporadic, *passar o ferro em/sobre* (∓ **to pass the iron in/over*) to *engomar* (*to iron*). In all these expressions *passar* has a different translation than *passar* as in [*passar* N(*exame,teste*,etc.) = *pass* N], where the verb is translated literally: *pass an exam*; *pass a test*.

revistar,V+FLX=FALAR+Aux=1+OBJTRundif75+Subset=319+EN=**search**+VSUP=passar+DRV=NDRV16:CASA+PrepN=a
escovar,V+FLX=FALAR+Aux=1+OBJTR19+Subset=654+EN=**brush**+VSUP=passar+DRV=NDRV332:CASA+PrepN=em
+PrepN=por
pentear,V+FLX=PASSEAR+Aux=1+OBJTRundif92+Subset=184+EN=**comb**+VSUP=passar+DRV=NDRV73:ANO
+PrepN=em
engraxar,V+FLX=FALAR+Aux=1+INOPwith+Subset=237+EN=**polish**+VSUP=dar+DRV=NDRV553:CASA+PrepN=a
+VSUP=passar+PrepN=em
pintar,V+FLX=FALAR+Aux=1+INOPwith+Subset=201+EN=**paint**+VSUP=fazer+DRV=NDRV109:ANO+VSUP=passar
+DRV=NDRV571:CASA+PrepN=em
engomar,V+FLX=FALAR+Aux=1+INOPmisc29+Subset=125+EN=**iron**+VSUP=passar+Prep=a+DRV=NDRV572:ANO
+PrepN=em+PrepN=sobre

Figure 7: Link between verbs and support verb constructions with the support verb passar

According to these linguistic constraints, we created relationship properties at the dictionary level and then applied those properties in local grammars to recognize support verb constructions in corpora and generate paraphrasing capabilities for them to be used in applications such as technical language writing and machine translation.

Our strategy to formalize idiomatic expressions and distinguish them from expressions with a more complex syntactic behavior was to lexicalize them. Therefore, semi-frozen expressions, where the verb is the only variable word in the whole expression, were listed in the dictionary of multiword expressions. For example, in *dar a mão à palmatória* (*to acknowledge being wrong*) or *fazer vista grossa* (*to ignore*), the verbs *dar* (*to give*) and *fazer* (*to make*) were assigned an inflectional paradigm and the rest of the words in the expression remain invariable.

As our electronic dictionaries provide enhanced meaning of single words, including contextual significance and increasingly more valuable tagging data, we also intend to enlarge and refine the role of a bilingual dictionary to include entries for multiword expressions that consider the understanding and analysis of each type of multiword expression, by beginning with support verb constructions and their paraphrases. The

ability to give the machine translation user multilingual paraphrasing ability constitutes an important step towards achieving better quality machine translation.

7.3. **Methodology**

To recognize and generate paraphrases for support verb constructions, morpho-syntactically and semantically related words are systemically linked in the electronic dictionary and annotated with derivational and distributional properties. Then, the properties formalized in the dictionary are combined with local grammars. For example, the lexical information in a local (syntactic) grammar can be used to identify the predicate in a support verb construction and subsequently this grammar can be applied to corpora.

Figure 8 illustrates a very simple grammar, which recognizes and annotates support verb constructions and their predicates. The grammar checks for a support verb (*VSUP*), followed by any left modifier (LeftMod) and a nominalization (*N+Nom*), and annotates it as a support verb construction, identifying the contents of the variable *N* (in parentheses) as the predicate of that construction.

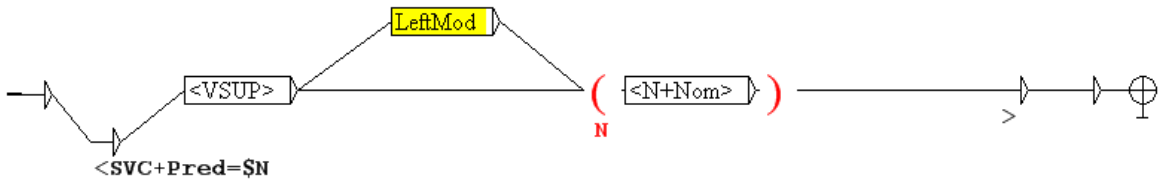

Figure 8: Grammar for recognizing and annotating support verb constructions and their predicates

To establish relations of equivalent morpho-syntactic predicates in the same language (Portuguese) or between two languages (particularly between Portuguese and English), as mentioned earlier, all predicate nouns in the dictionary have been classified as [NPred]. This lexical information can be used in a syntactic grammar to identify the predicate in a support verb construction and apply this grammar in corpora. Given the information in the lexicon, it is possible to identify and tag support verb constructions in a Portuguese text and identify the predicate noun for each support verb construction. After identifying the predicate noun, it is associated to a corresponding lexical strong verb, if it exists, and monolingual paraphrases are obtained. This is possible because there is a link in the

dictionary between the nominalization and the support verb, with specification of the verb that corresponds to the combination of those elements. For example, in Portuguese the multiword expression *dar um beijo* (*to give a kiss*) is recognized as a support verb construction, whose predicate noun is *beijo* (*kiss*). Figure 9 illustrates a concordance that identifies and tags support verb constructions in text, and identifies the predicate noun for each support verb construction.

no bordado para lhe	dar um beijo/<SVC+Pred=beijo>	na cara e os nossos olhos se cr
uer sair. -- Está a	dar uma festa/<SVC+Pred=festa>	? -- perguntou. Talvez fosse do
a uma agulha me faz	dar um salto/<SVC+Pred=salto>	; e, quando não consegue encor
à primeira e tem de	fazer várias tentativas/<SVC+Pred=tentativas>	-- o que é muito raro --, fica m
o ponto de me fazer	dar um grito/<SVC+Pred=grito>	. Toma conta da loja quando o
fazer dar um grito.	Toma conta/<SVC+Pred=conta>	da loja quando o pai está a dar
lições a iniciados.	Toma conta/<SVC+Pred=conta>	da loja quando o pai está a dar
regados, alunos que	fazem gazeta/<SVC+Pred=gazeta>	e mães com crianças de colo, c
ianças de colo, que	dão graças/<SVC+Pred=graças>	por terem este local acolhedor
que dão graças por	terem este local acolhedor/<SVC+Pred=local>	e alegre para passarem as tarde
primeira pedra, que	dá a outra/<SVC+Pred=a>	face, e por fora, mas não me c
jogador encantador,	dá gosto/<SVC+Pred=gosto>	vê-lo a defender, como se tives
a defender, como se	tivesse a bola atada/<SVC+Pred=bola>	aos pés, desviando os adversár
a maneira de evitar	fazer o pino/<SVC+Pred=pino>	, pôr-se a dar murros no peito c
er o pino, pôr-se a	dar murros/<SVC+Pred=murros>	no peito ou gritar de excitação.

Figure 9: Annotation of support verb constructions and identification of the predicate noun

Figure 10 shows the support verb construction *fez um esforço* (*made an effort*) in text before the application of the grammar in Figure 8 above. Figure 11 shows the same support verb construction already identified and annotated.

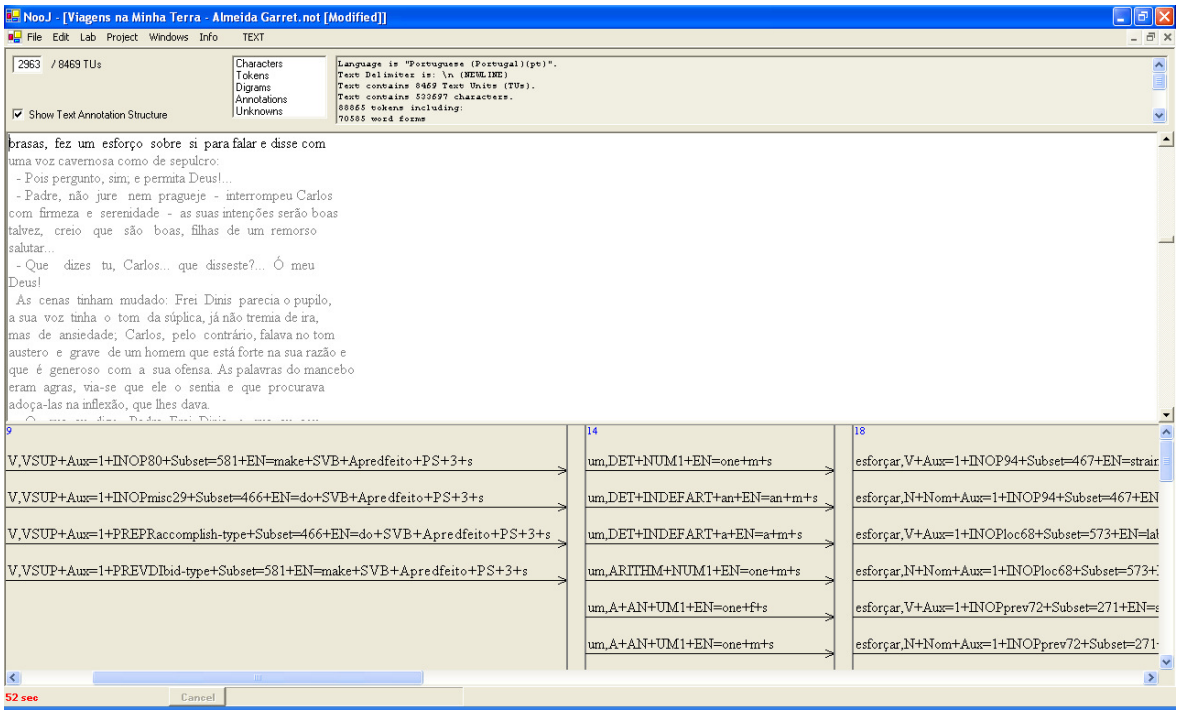

Figure 10: Annotation for the support verb construction *fez um esforço* before the application of a support verb construction grammar

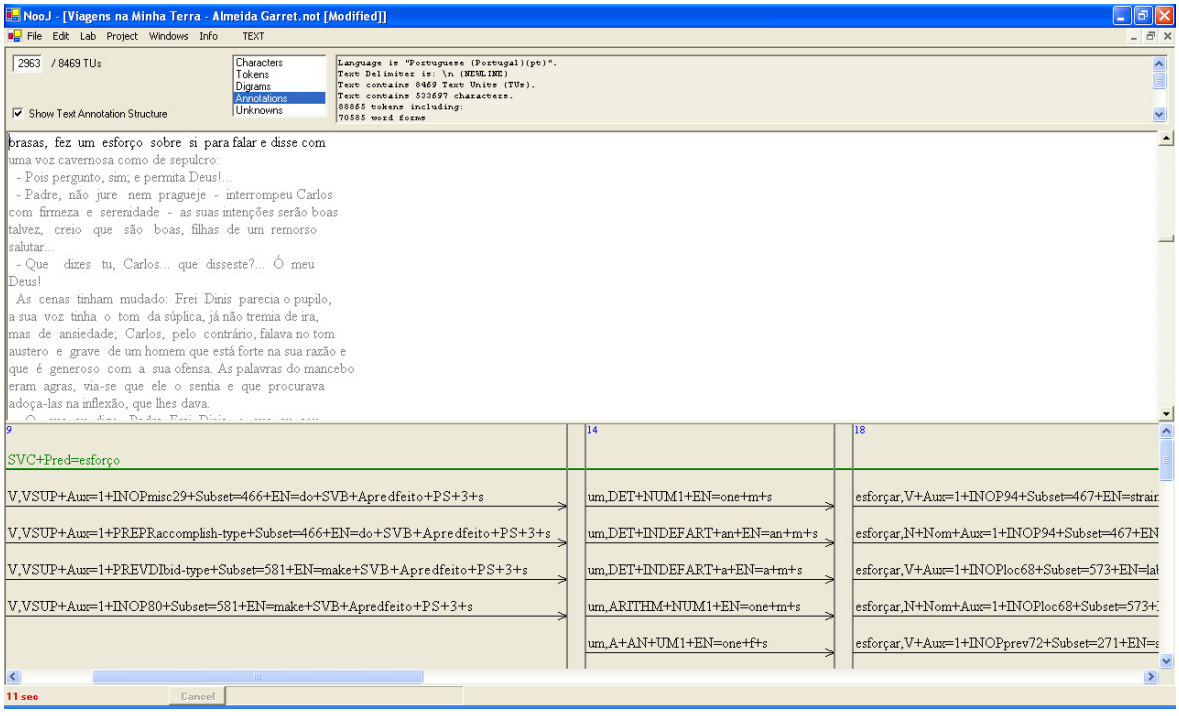

Figure 11: Annotation for the support verb construction *fez um esforço* after the application of a support verb construction grammar

Grammars that facilitate the recognition and annotation of multiword expressions, such as support verb constructions, may enable the whole expression to be paraphrased using another equivalent construction or replacing it by a strong verb, and for this reason they are extremely important for translation. For example, the support verb construction *fez um esforço* in Figure 10 and Figure 11 could be paraphrased by the verb *esforçou-se* and translated into English as one single word, *tried*.

7.3.1. **Paraphrasing**

Monolingual paraphrases can be obtained by means of local grammars that use the properties formalized in the dictionaries and make associations between words. Figure 12, extracted from [Barreiro, 2008a] represents a "naïve" local grammar used to recognize and generate support verb constructions and transforms them into their verbal paraphrases.

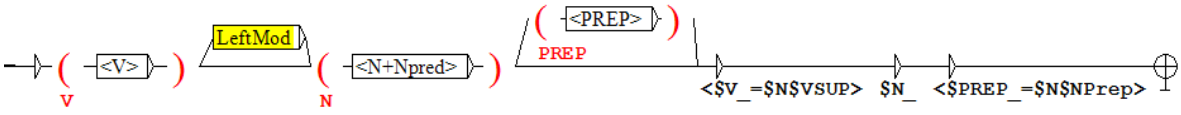

Figure 12: Grammar to recognize and paraphrase support verb constructions

This grammar matches verbs, which are marked in the dictionary as support verbs that are followed by a left modifier (determiner, adjective or adverb or other quantifiers), a predicate noun and optionally a preposition. Since we have classified all predicate nouns in the dictionary as [NPred], we can now use this lexical information in a syntactic grammar to identify the predicate in a support verb construction and apply this grammar in corpora. The elements in parentheses ( ) are stored in variables V, N or PREP. If a dictionary entry has a lexical constraint, such as NPrep=*a* in the phrase [*dar um grande abraço a*] (*to give a big hug to*], the support verb construction will be recognized by the grammar and mapped to the verb *abraçar* (*to hug*), the lemma of the noun specified in the variable $N_. The elements in bold <$V_=$N$VSUP>, and $PREP_=$N$NPrep> represent lexical constraints that are displayed in the output, such as specification of the support verb or the preposition that belongs to a specific support verb construction. The

predicate noun is identified, mapped to its deriver and displayed as a verb, the other elements of the phrase are eliminated. Figure 13 and Figure 14 show a concordance where Portuguese support verb constructions are recognized and paraphrased as lexical strong verbs, when applying these linguistic resources.

gosto de ver o comboio a	fazer corridas /correr	à velocidade máxima ao lon
o de cheque especial para	fazer doações /doar	às entidades que escolher. A
res e, quando é preciso ir	fazer filmagens/filmar	fora do estúdio, às vezes fic
e queria trocar de pares e	fazer um jogo /jogar	ao melhor de três sets , mas
dra deu-me um papel para	fazer uma lista de/listar	todas as coisas boas que ex
res foram à caracterização	fazer uns retoques/retocar	, outros estão a descansar n

Figure 13: Recognition and monolingual paraphrasing of support verb constructions (support verb/corresponding strong verb)

Reset Display: 5 characters word forms before, and 5 after. Display: Matches Results

Text	Before	Seq.	After
	militar, para que Edward pudesse	dar início à/iniciar	carreira de clínico geral. Quando
	militar, para que Edward pudesse	dar início à/começar	carreira de clínico geral. Quando
	elegante, se me é permitido	dar uma opinião/ter uma opinião	. Bernard aproveitou para dar uma
	elegante, se me é permitido	dar uma opinião/dar uma opinião	. Bernard aproveitou para dar uma
	elegante, se me é permitido	dar uma opinião/opinar	. Bernard aproveitou para dar uma
	Mr. Walsh. -- Quanto se deve	dar de gorjeta/gratificar	? -- O que quis dizer foi
	sinto que, se puder, devo	dar uma ajuda ao/ajudar	Council. Um deles abanou a
	servia isso mandaram-no ir	dar uma volta/passear	. Uma mulher com a saia
	sugestão de Persse, começa a	dar exemplos/exemplificar	: O Espelho da Sinceridade (Péricles
	parecesse irracional impedia-a de	dar voz ao/ouvir	medo. Ainda não eram cinco
	parecesse irracional impedia-a de	dar voz ao/escutar	medo. Ainda não eram cinco
	e sentia-se capaz de	dar um murro em/esmurrar	quem tentasse detê-lo. Mas
	seu carrossel. Gostávamos de lhe	dar uma palavrinha/falar	. -- Então vamos dar um aperto
	dar uma palavrinha. -- Então vam...	dar um aperto de/apertar	mão. «Deixar que o seu
	múltiplos relatórios, «poderia não...	dar lugar a uma/originar	colisão qualquer, mas quem sabe
	múltiplos relatórios, «poderia não...	dar lugar a uma/causar	colisão qualquer, mas quem sabe
	putrefacto? Às vezes ouvia-a	dar instruções a/instruir	Julie acerca das compras ou
	o problema do altruísmo, podia	dar sempre origem a/originar	confusões. O Oliver gosta de
	ligado a um tractor. -- Posso	dar uma olhadela/olhar	? -- perguntou Zoe. E diz ao
	pai que gostava de lhe	dar uma palavrinha/falar	, está bem? Harriet levantou um
	quando eu não estiver a	dar aulas/leccionar	, e nos fins-de-semana
	quando eu não estiver a	dar aulas/ensinar	, e nos fins-de-semana
	acenou com o charuto para	dar ênfase às/enfatizar	suas palavras. «Gostaria primeiro de
	Safie, Agatha e Félix foram	dar um passeio pelo/passear	campo e deixaram o velho
	Safie, Agatha e Félix foram	dar um passeio pelo/andar	campo e deixaram o velho
	e fui incapaz de lhe	dar resposta/responder	. Apenas me preocupava com um
	declarou-lhe: Não precisa de	dar explicações/explicar	. Reza fita-me como se
	brancos. -- Se me permite, posso	dar uma sugestão/fazer uma sugestão	.
	brancos. -- Se me permite, posso	dar uma sugestão/dar uma sugestão	.
	brancos. -- Se me permite, posso	dar uma sugestão/sugerir	.

Figure 14: Recognition and monolingual paraphrasing of support verb constructions (support verb construction/corresponding strong verb)

7.3.2. **Translation**

Machine translation using NooJ is implemented by translation grammars, relating two languages by means of variables and a translation operator (*TRANS*). This operator makes NooJ suitable for machine translation and capable of providing several appropriate translations of the same sentence. A grammar similar to the one used for paraphrasing is used to generate English translations. The only difference is that the output is specified to be in English. Figure 15 and Figure 16 show concordances where Portuguese support verb constructions are recognized in texts and converted automatically into English verbs. Even though we have not yet fully addressed the prepositions and in some cases they are missing, the translation is considered of a good quality.

a fazer um estágio para	dar aulas de/teach	religião, mas não se impor
m -- os filhos -- juntos e	fizeram a mudança para/change	Johannesburg, e ensinaram
. Necessitava apenas de	ter a certeza de/know	que não escapara à sua
ente hipotética. -- Deves	ter alguma ideia/know	. Dorothy andava a fazer u
não podemos deixar de	ter cautela/beware	. Pobre Caro, pensou Lyn
ra dos chinelos, antes de	ter chance de/can	mudar de idéia. Como pos
ope a Jean, esta pareceu	ter dificuldade em/avoid	olhá-lo nos olhos. Deixou
ao Kiss dela. Apesar de	ter falta de/lack	amor-próprio, isso não sig
igos e imprensa estava a	ter lugar /occur	numa longa galeria com ca
uiu ter filhos. -- Tens de	ter mão /control	nessa confusão toda. Sam
spondi, minha mãe deve	ter medo de/fear	cobras. Eu disse no Gabin
da loja antes de ele	ter tempo de/could	chamar a brigada de narcó
a triste aventura havia de	ter um fim/finish	.
Ela ouvira a tia Velma	ter uma discussão com/argue	Jack acerca de mostarda
de olhos fechados para	ter uma ideia de/know	como seria ser cego e
ter paciência.» «Voltei a	ter uma imensa vontade de/want	viver. A conversa parecia

Figure 15: Recognition and translation of support verb constructions (Portuguese support verb construction/corresponding English verb)

Reset Display: 5 characters word forms before, and 5 after. Display: Matches Results

Text	Before	Seq.	After
	a fazer um estágio para	dar aulas de/tutor	religião, mas não se importava
	a fazer um estágio para	dar aulas de/lecture	religião, mas não se importava
	a fazer um estágio para	dar aulas de/teach	religião, mas não se importava
	militar, para que Edward pudesse	dar início /start	à carreira de clínico geral. Quando
	militar, para que Edward pudesse	dar início /initiate	à carreira de clínico geral. Quando
	militar, para que Edward pudesse	dar início /commence	à carreira de clínico geral. Quando
	militar, para que Edward pudesse	dar início /begin	à carreira de clínico geral. Quando
	«Casamos no Natal.» -- Estou a	dar aulas para/tutor	me livrar da tropa -- explicou
	«Casamos no Natal.» -- Estou a	dar aulas para/lecture	me livrar da tropa -- explicou
	uma opinião. Bernard aproveitou para	dar uma volta /walk	pela cozinha. Ainda há-de dar
	Mas quem é que vai	dar um subsidio/subsidize	capaz para estudar os Oof
	Mas quem é que vai	dar um subsidio/subsidise	capaz para estudar os Oof
	a dizer-te que pode	dar a impressão de/seem	que não confio em ti
	sinto que, se puder, devo	dar uma ajuda /help	ao Council. Um deles abanou a
	sinto que, se puder, devo	dar uma ajuda /aid	ao Council. Um deles abanou a
	sinto que, se puder, devo	dar uma ajuda /assist	ao Council. Um deles abanou a
	sugestão de Persse, começa a	dar exemplos/exemplify	: O Espelho da Sinceridade (Péricles
	parecesse irracional impedia-a de	dar voz /hear	ao medo. Ainda não eram cinco
	parecesse irracional impedia-a de	dar voz /listen	ao medo. Ainda não eram cinco
	e sentia-se capaz de	dar um murro em/punch	quem tentasse detê-lo. Mas
	seu carrossel. Gostávamos de lhe	dar uma palavrinha/speak	. -- Então vamos dar um aperto
	dar uma palavrinha. -- Então vamos	dar um aperto de/shake	mão. «Deixar que o seu

Figure 16: Recognition and translation of support verb constructions (Portuguese support verb construction/corresponding English verb)

The concordance illustrated in Figure 17 shows that the output produces several different English verbs for each support verb construction. For instance, *fazer várias tentativas* (*make several attempts*) can be translated into five different verbs *try*, *endeavour*, *attempt*, and *intend*. This ambiguity is related to the fact that there are five English dictionary transfers for Portuguese verb *tentar*, the verbs *try*, *endeavour*, *attempt*, and *intend*. The support verb construction *fazer várias tentativas* could be further translated into *strive*, *aim*, *seek*, or *undertake*, if these verbs were dictionary transfers for the Portuguese verb.

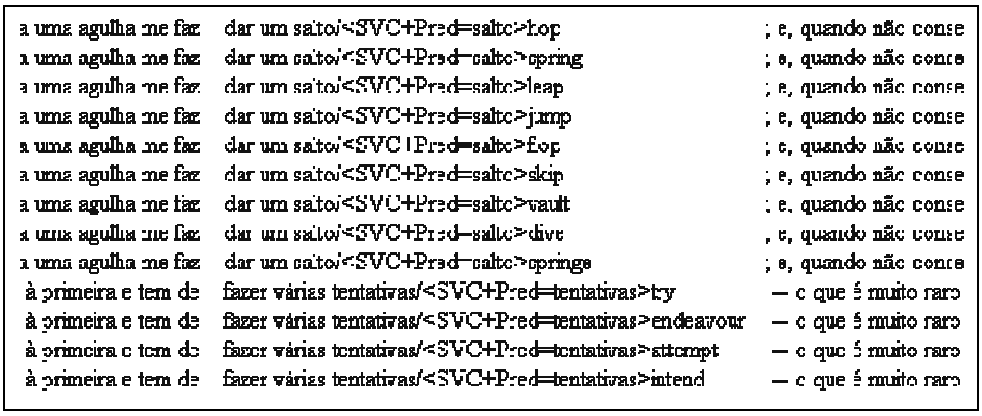

Figure 17: Annotation of support verb constructions, identification of the predicate noun, entailed paraphrase into a single verb and translation into English

Like humans, machines can also produce more than one "acceptable" translation for the same sentence. Parallel translations of the same sentence are paraphrases. Like other machine translation systems, NooJ enables multiple translation, with the advantage that the *+UNAMB* feature makes it possible to establish a default transfer to be used in general translations. This provides one translation, eliminating all other possibilities. As a result, the system can be customized according to the needs of the linguist or the user, provide multiple translations or one specific translation.

Semantic constraints can be used for meaning disambiguation and refinement, if necessary or preferable as long as we define those constraints in the grammars, as in Figure 18. Also, it is possible to establish a default transfer to be used in general translations so that the output does not show several possibilities. However, the interesting aspect here is to show that there are several valid possibilities for translating the same word or phrase, i.e., there are several bilingual paraphrases, which can be used either to simplify text before translation or to use in comparing legitimate outputs automatically. The grammar corresponding to this concordance has a constraint to tell NooJ that the Portuguese noun "*tentativa*" occurs with the support verb *fazer*. In other words, any graph that deals with support verb constructions indicates that it is not any support verb that can be used with any noun, but only the one specified in the dictionary. Compound variables are used so that the noun in the recognized sentence actually corresponds to the support verb in the dictionary for that noun.

Figure 18 shows how the syntactic-semantic properties in the dictionaries are used in local grammars to paraphrase Portuguese support verb constructions for *fazer barulho*

(*make a noise*) and translate them into English. This grammar recognizes the sequence of a support verb with a predicate noun of the type [measure + abstract + noise] with any pre- or post-modifiers and translates it into *make a noise*. The grammar filters support verbs or support-verbs-like, such as *fazer* (*make*), *produzir* (*produce*) or *criar* (*create*) and predicate nouns such as *barulho*, *ruído*, *barulheira*, *chinfrineira*, *chinfrim*, etc. as long as they are classified in the dictionary with the semantic property "noise". This is the type of paraphrase from one support verb construction into another support verb construction in different languages. This grammar shows how to recognize very specific support verb constructions in corpora with exact pattern recognition.

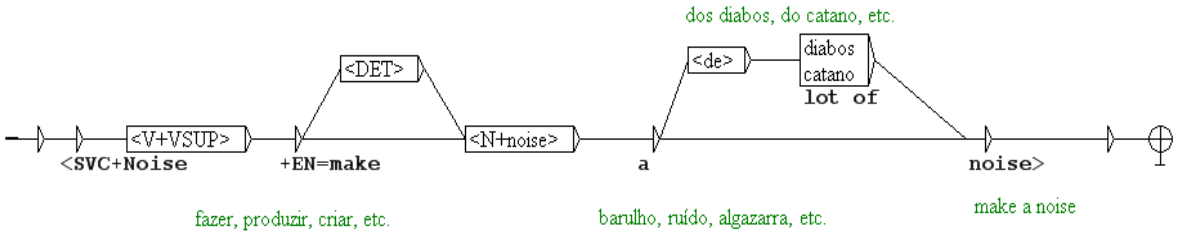

Figure 18: Local grammar to analyze, paraphrase and translate support verb constructions for Portuguese *fazer barulho* (*make a noise*)

Figure 19 shows the application of the previous grammar to text; i.e., the recognition and paraphrasing of Portuguese support verb constructions in text and their translation into English.

lo. Eles estão a	fazer barulho/<SVC+Noise+EN=make a noise>	. A criança fazia
a. A passarada	fazia uma algazarra/<SVC+Noise+EN=make a noise>	enorme. O carr
ıorme. O carro	fazia uma barulheira/<SVC+Noise+EN=make a noise>	enorme. A máq
arulho. O bebé	faz uma berraria/<SVC+Noise+EN=make a noise>	enorme. O bebı
ıno. A multidão	fará um barulho/<SVC+Noise+EN=make a noise>	dos diabos. Os

Figure 19: Application of previous local grammar to text

Figure 20 shows how to use semantic constraints to filter out undesired translations. For example, the verb *pregar* is translated differently into English depending on the noun that follows it. If it is a noun type information (IN), instructional/legal, ritual, such as *missa* (*mass*) or *sermão* (*sermon*), the expression translates into the verb *preach* or into *say* plus

the noun transfer *mass* or *sermon*. If it is a noun type abstract (AB), general concept (gen), such as *ideia* (*idea*), *virtude* (*virtue*), *religião* (*religion*) or *verdade* (*truth*), the expression translates into the verb *proclaim* or *advocate*. If it is a noun type concrete (CO), fastener (fast), such as *prego* (*nail*), the expression translates into the verb *hammer* plus the noun transfer. If it is the noun type abstract (AB) concept, negative cause (negc), such as *susto* (*fright*), the expression translates into the verb *scare*. In this case, an argument N1 is required, corresponding to the indirect object *a N* (preposition + N).

```
pregar Det N(missa,sermão) > preach = say N
pregar Det N(ideia,virtude,religião,verdade) > proclaim N > advocate N
pregar Det N(prego ,etc.) > hammer N(nail,etc.)
pregar N(susto) Prep(a) N > scare N
estabelecer Det N(negócio,empresa,loja,etc.) > open Det N
estabelecer N(regras,princípios) > lay down N
apresentar N(desculpa) > apologize
apresentar Det N(opinião,sugestão) > give Det N(opinion,suggestion)
apresentar Det N(moção,censura) >  bring N forward
prestar Det N(serviço) > offer Det N(service)
prestar N(atenção) > pay N(attention)
perseguir N(objectivo,propósito,etc.) > follow N
perseguir N(pessoa) > chase N = hunt after/down N
pedir N(desculpa,perdão) > apologize = say sorry
pedir Det N(esmola) > beg
observar Det N(lua) > observe/see N
observar Det N(lei) > obey N(law)
representar Det N(papel) > play Det N(role)
representar Det N(contributo,etc.) > represent N
```

Figure 20: Sample of Portuguese-English translation rules

Translation grammars assign precision to the translation of expressions that cannot be translated literally. They also help improve meaning disambiguation and provide semantic refinement to the source language. The few Portuguese-English translation rules of Port4NooJ were adapted (inverted) from the Logos English-Portuguese Semtab rules available at Linguateca website: http://www.linguateca.pt/ - *Repositório*.

This chapter has described the linguistic resources that were used to start a new system. Several grammars have been described and the manner in which they interact with the dictionary and with the morphological rules were exemplified. The plan is to improve and enlarge the system to create new resources and expand functionalities, as

well as fine-tuning its operation. However, despite the resources already developed, we regard these as a foundation on which to build further resources for enhancing machine translation capabilities.

Chapter 8

New Resources and Applications

*

Chapter Eight describes the new resources and applications created in this research. Port4NooJ is an ontology-based open source natural language processing system which includes bilingual resources for machine translation. DicTUM is a dictionary of multiword expressions. ReWriter and ParaMT are two new automated software tools to recognize multiword expressions and generate paraphrases of them. ReWriter is a monolingual paraphraser used as a writing aid, and ParaMT is a bilingual/multilingual paraphraser applied in machine translation. The interface for ReEscreve, the Portuguese version of ReWriter, is presented. We illustrate the application of *ReWriter* and its resources to controlled language of general and technical language and its extensibility to larger and more complex linguistic phenomena than multiword expressions.

*

The research conducted here is intended to find a place in the ideal machine translation tool. In the process of achieving such a goal, new applications and linguistic resources were created. In any knowledge-based language processing application, the linguistic resources represent the foundation. Good linguistic descriptions lead to sophisticated resources that help improve systems. In machine translation especially, the linguistic resources are the driving force that boosts the translation process. Associated with the resources, ontologies also play a very important role as descriptors of entities or realities. Ontologies have been discussed intensely from the philosophical and lexical perspectives and what seems to be the best way of representing entities and realities for some authors may not be the best for others. In addition, what may work for English (the language for which most work has been developed) may not work for other languages [Santos, 1999]. There is a long history of attempts to create ontologies. Greek philosophers thought about words describing entities, events, objects, etc. Since then, plenty of other attempts have been made, from thesauri [Roget's thesaurus] to the Wordnet [Fellbaum, 1998]. In

the computational area, the most modern application of ontologies is the semantic web [Antoniou et al., 2005] [Antoniou & Bikakis, 2007] [Davies et al., 2006]. Since 2000, OntoLex workshops have been an incentive to the interdisciplinary community of lexicographers, ontologists and computational linguists and has been stimulating the integration of ontologies with lexical resources (see call for papers for LREC 2008). The Ontolex concept is not totally new though. It has been used for decades in the Logos system and it worked efficiently in their machine translation. The new lexical and ontological resources of the present study adopted Logos SAL ontology.

The new lexical-ontological resources were used for paraphrasing and pre-editing of texts to support controlled language writing and experiments were made to provide evidence for the impact of this pre-editing on machine translation. As a result, we present two automated paraphrasing systems used for these two different purposes. ReWriter is used as a standalone paraphraser to generate monolingual paraphrases that help simplify source text, reduce ambiguity and the number of words. ParaMT is used to generate bilingual/multilingual paraphrases (or translations) and it operates as an integrated function for machine translation. It can be used as an on-line linguistic aid to translators so they can determine the best translation for a certain expression or sentence. In the next sections, the new system, resources and tools will be described.

8.1. **Port4NooJ: Ontology-driven Resources**

Port4NooJ is a set of linguistic resources developed in the NooJ linguistic environment for the automated processing of Portuguese language. They integrate a bilingual extension and can also be used in Portuguese to English machine translation. As mentioned in Chapter 7, Port4NooJ uses OpenLogos English-Portuguese lexical resources (which were reversed and converted). Corpora resources and other data and tools were also used, namely Linguateca's COMPARA [Frankenberg-Garcia & Santos, 2003] [Santos & Inácio, 2006] and METRA [Sarmento, 2007].

Port4NooJ resources are publicly available at Port4NooJ website: http://www.linguateca.pt/Repositorio/Port4Nooj/ and can also be downloaded from the NooJ website at http://www.nooj4nlp.net (Resources > Portuguese). They have been integrated in Corpógrafo [Sarmento et al., 2004], [Sarmento et al., 2006], [Maia & Matos, 2008] and are being used in this tool, thus far, to perform queries and to obtain

concordances (see http://www.linguateca.pt/corpografo/).

The Port4NooJ dictionary format is different from the one in the original resources. There are several different dictionaries, which can operate independently. The system has a completely new inflectional and derivational system, new parsing, translation, and generation components. Additional new resources for monolingual and bilingual/multilingual paraphrasing, and for the development of new machine translation systems are also available. The system is described in [Barreiro, 2008b] and more information can be obtained from the Port4NooJ Tutorial and the resources overview document on the Port4NooJ website. These resources are also available on the OpenLogos website.

8.2. **DicTUM: a Dictionary of Multiword Expressions**

DicTUM stands for "**Dic**ionário de **T**ermos e **U**nidades **M**ultipalavra" and it is an electronic dictionary of (non-specialized) terms and multiword expressions. It comprises compounds of general language, some lexical bundles and other expressions. These cover nominal expressions such as *cabo de vassoura* (*broomstick*) or *luz solar* (*sunlight*); verbal expressions, such as *marcar pontos* (*score*) or *piscar o olho* (*wink*); adjectival expressions such as *fraco de espírito* (*feeble-minded*), *cor-de-rosa* (*pink*); and adverbial expressions such as *com entusiasmo* (*enthusiastically*) or *de parte* (*aside*). This dictionary is soon to be significantly expanded by the incorporation of several thousand nominal compounds (predicate nouns), which appear frequently in support verb constructions, such as *juízo de valor* (*judgment*) as in *fazer um juízo de valor* / *fazer juízos de valor* (*make a judgment* / *make judgements*) or *chamada telefónica* (*phone call*) as in *fazer uma chamada telefónica* / *fazer chamadas telefónicas* (*make a phone call* / *make phone calls*). Figure 21 shows some of the entries that can be found in the multiword expressions' dictionary. The annotation [PL+encl] stands for enclosed spaces; [CO+tool] stands for concrete, functional tools/devices; [MA+liqu] stands for mass, liquids; [NAV+Apred+col] stands for non-adverbial, predicate, color; [AN+des] stands for animate, designations or professions; [LocTime+TEMP] stands for locative, time, temporal; [STAT+phr] stands for stative, phrase; [LocTime+TEMP+puncpast] stands for locative, time, temporal, punctual past; [COOR] is an annotation for a coordinating conjunction; [SUB] is an annotation for a subordinating conjunction; [ASSOC] stands for an associative preposition; [Loc+AT] stands for a locative,

at-type preposition; [ALOG] stands for analogical preposition. The compound '*bebida alcoólica*' appears twice. One entry is translated by the neutral expression '*alcoholic drink*', which can be used to produce a more neutral translation (less marked); another entry is translated as '*booze*', marked as a slang word. By default, the machine translation system translates the expression '*bebida alcoólica*' as '*alcoholic drink*', but in some texts the translation '*booze*', could be the most adequate.

```
adro da igreja,N+FLX=NPN00+PL+encl+EN=churchyard
cabo de vassoura,N+FLX=NPN00+COtool+EN=broomstick
bebida alcoólica,N+FLX=NA02+MA+liqu+EN=alcoholic drink+UNAMB
bebida alcoólica,N+FLX=NA02+MA+liqu+EN=booze+slang
cor de laranja,A+NAV+Apred+EN=orange
sul-americano,A+FLX=AA03+AN+des+EN=South American
a curto prazo,ADV+LocTime+TEMP+EN=in the short run
fora de serviço,ADV+STAT+phr+EN=out of order
há muito tempo,ADV+LocTime+TEMP+puncpast+EN=a long time ago
isto é,CONJ+COOR+EN=i.e.
já não,CONJ+COOR+EN=no longer
mesmo assim,CONJ+SUB+EN=even so
juntamente com,PREP+ASSOC+EN=along with
à direita de,PREP+Loc+AT+EN=at the right of
em conformidade com,PREP+ALOG+EN=in congruence with
```

Figure 21: Sample of the dictionary of multiword expressions – entries for compounds and lexical bundles

Only words of general vocabulary that have a less variable character are stored in the dictionary. Frozen expressions are also stored in this dictionary, such as fully idiomatic expressions as *dar a mão à palmatória* (*acknowledge being wrong*), *fazer vista grossa* (*to ignore*) or [*dar cabo dos nervos a NP*] (*to irritate NP*), in Figure 22.

```
dar a mão à palmatória,V+FLX=PHRDAR+EN=acknowledge being wrong
fazer o sangue subir à cabeça,V+FLX=PHRFAZER+EN=ficar tonto
ter o sangue nas guelras,V+FLX=PHRTER+EN=be alive
fazer vista grossa,V+FLX=PHRFAZER+EN=ignore
dar parte de fraco,V+FLX=PHRDAR+EN=give up
```

Figure 22: Sample of the dictionary of multiword expressions – entries for fully idiomatic expressions

Multiword expressions with a more variable behaviour are formalized in local grammars such as those described in § 7.3.1. That is the case of support verb constructions that

maintain a certain flexibility in respect to determiners and prepositions, and permissibility of inserts, such as [*dar um passeio*] (*to go for a walk*) or [*dar vários passeios por/em NP*] (*to go for several walks to NP*), among many others.

8.3. **ReWriter: a Standalone Paraphraser**

ReWriter is a writing tool based on a monolingual paraphraser that helps users to change, simplify or clarify their texts. It is a useful authoring aid in word processing and a useful tool to help write technical language. ReWriter can also be used as a pre-editor for machine translation. The Portuguese version, ReEscreve, is publicly available on the Internet at http://poloclup.linguateca.pt/Reescreve/ and described in [Barreiro, 2008c].

At the current stage, ReWriter recognizes support verb constructions and converts them into verbs or similar expressions. It follows three steps. The first step of the ReWriter tool is to recognize a support verb construction in a text and extract it. It is sent to the paraphrase database to be matched with an equivalent verb or another equivalent expression. This verb or expression receives the same inflectional features as the original support verb inflection features. At the end, one or more suggestions for rewriting the original support verb construction is provided.

ReWriter is designed to operate in an interactive way or automatically, but at the moment only the interactive mode is available. In interactive mode, ReWriter can be applied in word processors similarly to the way synonyms are applied. However, ReWriter has a more sophisticated replacement system. It recognizes inflected forms and retrieves the equivalent expression with the same inflectional features. For example, if the recognized form of the support verb is the third person plural, simple past, such as *deram um passeio* (*they went for a walk*), the equivalent paraphrase will be the third person plural, simple past of a corresponding verb, as in the example, *passearam* (*they walked*) or the same features for any possible equivalent. The paraphrase database contains only lemmas of the support verb constructions. The inflection forms are obtained by using the inflectional system described in [Barreiro, 2008b]. More flexible/variable support verb constructions with different types of modifiers are being successfully paraphrased.

Figure 23 describes the home page of the interface for ReEscreve. A brief description of the service is displayed on the right hand side. And the user can click on a hyperlink below the summary description to know more about the service. The window box allows

the user to insert text that needs to be rewritten. The possibility of submitting a file is also contemplated in the interface but not implemented yet.

Figure 23: Home page for the public web service of the writing aid tool, ReEscreve

The user can try the service with an example text, if he/she clicks on the hyperlink word "exemplo". A text is displayed in the text window box, such as it is illustrated in Figure 24.

Figure 24: Sample text with support verb constructions to illustrate the functionalities of ReEscreve

If the user clicks in the command button "Obter resultados", the text is retrieved with suggestions for the editing. The suggestions presented by ReEscreve appear in blue next to the original expressions, in green, as can be seen in Figure 25.

Figure 25: Paraphrasing capabilities retrieved by ReEscreve

The user can choose which expression is most suitable for his/her needs in that particular context, either the original one or one of the suggestions presented by the system. Then the user can rewrite the text by clicking on his/her favorite expression. If the user selects the suggestion presented by the system, the original expression disappears. If the user selects the original expression, the suggested expression gets hidden. Figure 26 shows the result of a rewritten text. The expressions in blue are new words introduced in the text. These words were chosen to replace the original expressions. The expression in red was a new expression that the user was allowed to insert in the system.

Figure 26: Text rewritten after interactive use of ReEscreve

The right-hand side menu shows the original expression in green for which suggestions of paraphrasing capabilities were provided for user approval. When the user clicks on each of these expressions, a box opens underneath the sentence that contains the pair original/suggestion. The box then shows the different possible sentences. These possible sentences are paraphrases of each other. The user can contribute to the enlargement of the number of paraphrases through the insertion of one or more equivalent expressions that are suitable in the same context and do not exist yet in the system database, i.e., not presented as suggestions by the program. He/she can also send a comment about the quality of the suggestions provided by the program. Figure 27 shows the interactive box before and after the user submitted a suggestion to the system. The suggestion appears in red, after the user clicks on the button "Sugerir" (suggest).

When the ReEscreve is used in an interactive mode to edit text, the choice for paraphrase is made and a usage counter for that choice can be incremented, thus building statistical data to further assist the tool in helping making the right choice. If this resource is deployed on the Internet, accumulation of usage data can "grow" this resource naturally.

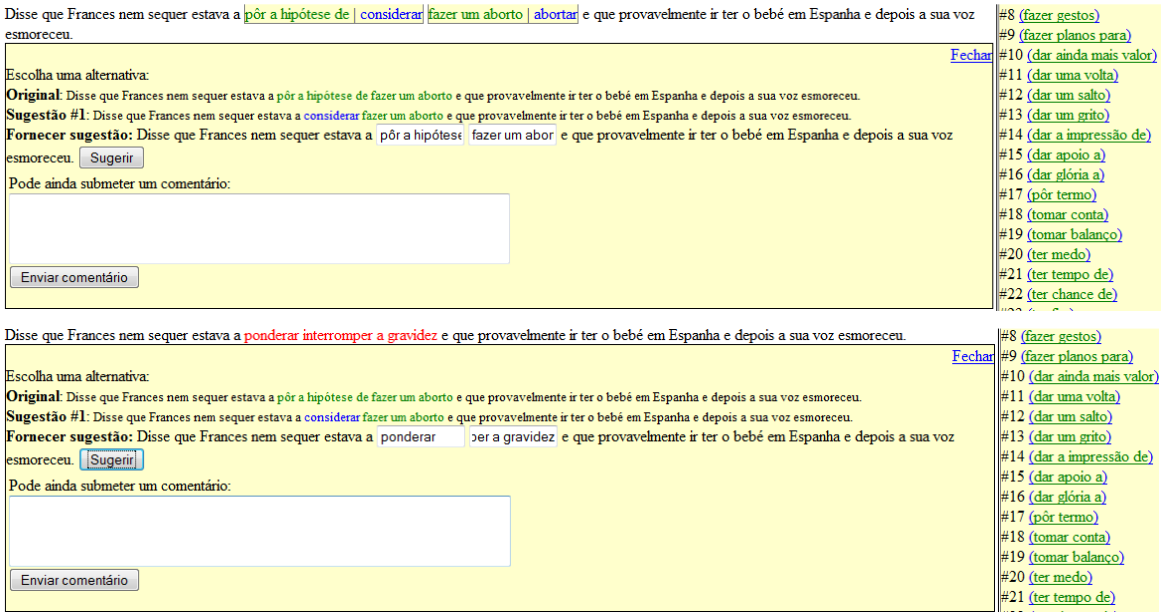

Figure 27: Box for user consultation and interaction/suggestion

8.4. **ParaMT: a Paraphraser for Machine Translation**

Bilingual and multilingual paraphrases are extremely useful, even indispensable, in machine translation, and any advancing machine translation system should require a paraphrasing system. ParaMT is an automated bilingual/multilingual paraphraser that could be integrated in machine translation systems. At the current stage, it uses the properties of the dictionary entries combined with local grammars to recognize, paraphrase and translate multiword expressions. ReWriter is currently working only with one linguistic phenomenon: support verb constructions. When translating a text automatically, support verb constructions are recognized and the system allows the user to translate them literally or transform them automatically into verbs in other languages. First, ParaMT uses local grammars to recognize support verb constructions. It then takes the predicate noun of the support verb construction and maps it to its corresponding verb link from the dictionary. Lastly, it translates the verb automatically. If there is no corresponding verb in the dictionary, the support verb construction is translated literally. Figure 28 illustrates a machine translation system architecture with the option of converting support verb constructions into their paraphrases (verbs).

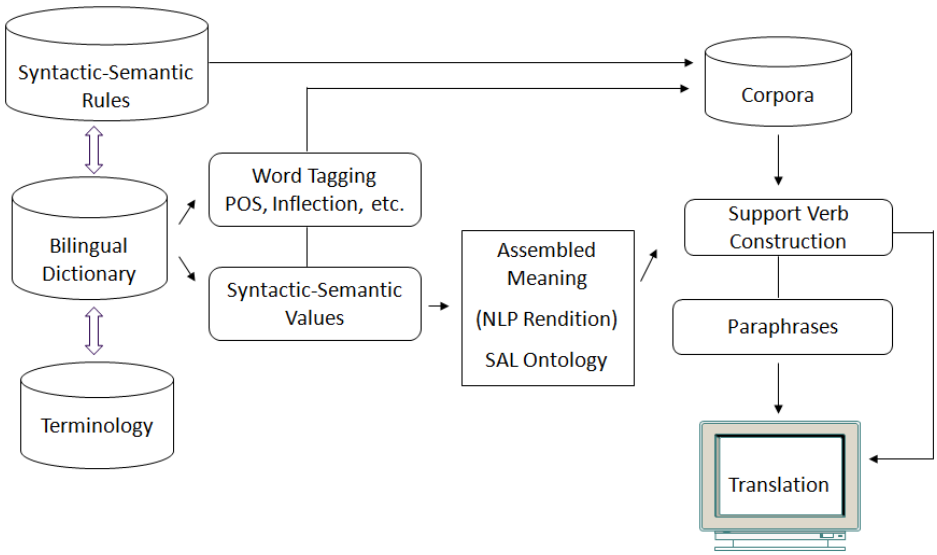

Figure 28: Machine translation with paraphrases

Paraphrases are used in the machine translation system to improve translation results. The prototype machine translation system contains three main input sources of data: (i) syntactic-semantic bilingual grammar rules (ii) bilingual dictionary and (iii) terminology databases. Each one of these data sources interacts directly with the dictionary properties (word tagging), grammatical category (PoS), inflectional system, and with the syntactic-semantic values represented as a SAL-based ontology. When all this linguistic information is applied to corpora, the support verb constructions are identified, paraphrased, and translated. In cases where there is no corresponding or suitable paraphrase, the support verb construction is translated literally.

8.5. **Controlled Language**

There are two distinct reasons for using controlled language, namely to improve readability and to enable automatic analysis and processing of the text. Controlled language is efficient and can improve the quality of the language to be written and/or translated automatically. Eliminating ambiguity, redundancy, vagueness, density and (sometimes) linguistic creativity and complexity has a positive impact on the quality of the machine translation results. It also brings an economic impact to the machine translation process by reducing the time and cost of the post-editing. This is the reason why machine translation developers recommend the use of technical and more controlled language to

their clients. Nevertheless, the rules that drive controlled language compliance are very simple and, to date, are limited to simple character string matching and replacement. Our research work looks into the advantages of augmenting existing controlled language with usable linguistic knowledge to provide clearer texts. Machine translation of support verb constructions cannot be successful unless it is disambiguated by using syntactic and semantic knowledge in processing them. However, conversion of support verb constructions to verbs has a positive impact on translation cost, if word count is a sensitive issue. Controlled language can apply to both general language and technical language, as can be seen in the following (cf. § 8.5.1 and § 8.5.2).

8.5.1. **Applied to General Language**

ReWriter can use our linguistic resources to paraphrase and pre-edit texts to support controlled language in general domains. Using controlled language brings several advantages: consistency, maintainability and control, and translatability. People often write without paying enough attention to consistency. It is common to find texts that are inconsistent regarding syntactic structures, voice, tense, style, etc. For example, consistently converting all the support verb constructions to verbs, makes the style of the text more direct, often brings improvement in meaning precision and changes the mood in the text from vague to instructive. In today's information society, maintainability of Internet multilingual sites, marketing information, and user manuals, is complex and important. Nowadays, texts must be written so that they can easily be updated. Controlled language helps maintain texts more easily and also permits easier and better control of language. And finally, one of the main characteristics of controlled language is its translatability, i.e., the writing of more easily translatable text. Nowadays, translatability is often a key requirement of text. There are many instances where the writer uses words that do not add to the meaning. Any process that reduces the number of words without sacrificing the meaning will reduce noise in the message. This facilitates machine translation. Presenting simplified paraphrases for certain linguistic phenomena, tends to render the text easier to understand and closer to "machine translation ready", that is, easier and more accurately translatable.

Controlled language that includes paraphrasal knowledge also provides significantly more coverage in the control of style. If used to pre-edit, it can produce superior quality machine translation output that meets the needs of the linguistic resource user.

Support verb constructions can be paraphrased and translated by other support verb constructions or by verbs. In the cases where the support verb construction can be replaced by a full verb, such as in *dar um passeio* (*to take/go for a walk*) (*cf.* (1)), morpho-syntactically and semantically equivalent to the verb '*passear => to walk*' (*cf.* (2)), paraphrasal knowledge can be very useful. Comparing the two types of expression, we notice that the two extra words, the support verb and the determiner, add little or no meaning to the phrase, to the point that some translators decide not to even use them.

(1) E depois vai dar um passeio por aí, (…) [CdP]
=> And then (s)he is going to go for a walk around, (…)

(2) Quando você acabar, vamos passear por aí. [CdP]
=> When you finish, we will go for a walk.

A machine translation system needs to understand support verb constructions so that it processes them with quality, precision and minimum loss of meaning. This research focused on support verb constructions, however, this phenomenon holds true for additional types of expressions that can be paraphrased and simplified. We expect that it will be possible to repeat the present study, using other categories of paraphrase, so that coverage of paraphrases can become more complete. The methodology adopted in the current research can straightforwardly extend to relative clauses, 'if' clauses, some passives and passive negation clauses, among others. A few examples of different types of linguistic phenomena where paraphrasing is useful and ReWriter could help to control the writing are listed below. Some types of corrections are standard practice in (human) revision.

[Paraphrasing adverbials]

à volta da órbita ≡ periorbital (popular versus technical)
around the orbit of the eye ≡ periorbital

[Paraphrasing relative clauses - into adjectival past participles]

N0 que têm sido escritos ≡ N0 que foram descritos ≡N0 escritos
N0 that have been written ≡ N0 that were described ≡ N0 written

[Paraphrasing if clauses]

se for necessário ≡ se necessário
if it is necessary ≡ if necessary

[Paraphrasing coordinated noun phrases - conjoining or disjoining]

recursos linguísticos para o ensino e para a investigação
Ŧ ?linguistic resources for teaching and for research
≡ recursos linguísticos para o ensino e a investigação
Ŧ linguistic resources for teaching and research

[Paraphrasing subjunctive clauses - into infinitives]

pedimos o favor que confirme a sua participação
*Ŧ *we ask the favor that you confirm your attendance*
≡ pedimos o favor de confirmar a sua participação
*Ŧ *we ask the favor of confirming your attendance*

[Paraphrasing marked-up constructions]

se a necessidade do utilizador é criar um texto em linguagem controlada
Ŧ ?if the end-user need is to create controlled language text
≡ se o utilizador necessita de criar um texto em linguagem controlada
Ŧ if the end-user needs to create controlled language text

[Paraphrasing of vague and undefined or null subject sentences] (whenever the real subject/actor is known)

[-] houve um grito na rua ≡ [N-PRON]/alguém gritou na rua
Ŧ there was shouting in the street ≡ [N-PRON]/someone shouted in the street

[Paraphrasing passives - whenever suitable]

Esse livro foi escrito por Saramago em 2008 ≡ Saramago escreveu esse livro em 2008
That book was written by Saramago in 2008 ≡ Saramago wrote that book in 2008
Florida foi atingida por um tornado ≡ Um tornado atingiu a Florida
Florida was hit by a tornado ≡ A tornado hit Florida
O carro foi roubado ≡ Alguém roubou o carro
The car was stolen ≡ Someone stole the car

A formal linguistic study of paraphrases of this kind in particular represents a significant contribution to natural language processing in general, and to machine translation in particular, and benefits verification of translation quality overall.

The concept of interchangeability of text has only become a reality since people understood the mechanism of text editing on a personal computer. Ease of editing has brought a new way of looking at text, as flexible and editable. With this prior concept firmly embedded in the mind, the ability to select suitable paraphrases helps improve the skill of the editor.

8.5.2. **Applications to Technical Language**

The applicability of the linguistic resources developed in the current research goes beyond controlled language in general domains. We believe that paraphrasing tools will be useful in specific semantic domains too. Even though in technical language, texts are more controlled and carefully written by professionals who know the terminologies and writing conventions of their field of knowledge, high-quality editing tools with good paraphrasing capabilities can help speed up the writing task.

Specialized and scientific domains already use controlled language to create less ambiguous and clearer and more precise texts. Reducing ambiguity and complexity is achieved by "controlling" grammar and vocabulary, and results in a better quality of technical documentation.

In technical texts, many predicates are nouns. We experimented by paraphrasing a few support verb constructions which use domain specific nouns or terms (most technical nouns are predicative), such as from the financial/auditing field (viz. *fazer uma auditoria financeira – to perform a financial audit*), from the legal field (viz. *dar o poder paternal – to give parental rights*) or from the biomedical field (viz. *fazer uma operação – to do an operation*). We verified that, contrary to what happens with support verb constructions that use non technical terms, and where it is very common to find corresponding verbs to replace support verb constructions, viz. *tomar uma decisão* (*to make a decision*), and *decidir* (*to decide*), in scientific and technical fields, there are often no corresponding verbs for the terms that name relations, procedures, etc. However, the paraphrasing of these expressions is done by employing a different linguistic strategy, the replacement of the elementary support verb with more sophisticated stylistic variants.

As mentioned in the preview presented in Chapter 1, support verb constructions, such as *fazer uma operação* (*to do an operation*) can often be disambiguated by being paraphrased with corresponding correct paraphrasing capabilities. The paraphrasing capabilities can be verbs (*operar* > *to operate on*) or lexical-syntactic extensions (*realizar uma operação* > *to perform an operation* or *submeter-se a uma operação* > *to undergo/have an operation* = *to be operated on*), depending on the context, and on the semantic classes of the arguments. These lexical-syntactic extensions represent stylistic variants of support verb constructions. For example, [Chacoto, 2005] presents a sample list of about 200 health related Portuguese predicate nouns that co-occur with the support verb *fazer* (*to do*). These nouns refer to clinical exams and medical treatments or surgical interventions, such as *fazer uma radiografia* (*to perform/have diagnostic X-ray examination*) or *fazer uma lobotomia* (*to perform/have a lobotomy*). Doctors and other medical professionals use these expressions in their daily language when communicating with their colleagues or with their patients. For example, obstetricians order their patients to have medical exams using expressions such as *fazer uma ecografia* (*to have an ultrasonography*), *fazer uma amniocentese* (*to have an amniocentesis or amniotic fluid test*), *fazer análises à urina* (*to have an urine test* or *urinalysis*), *fazer análises ao sangue* (*to have blood tests*), *fazer o parto* (*to deliver a baby*), etc. These expressions appear in Portuguese or Brazilian blogs, including doctor's blogs, in e-mails, in websites describing medical exams or surgical procedures, in online newspapers, etc.

Our linguistic resources can be used and the monolingual paraphraser ReWriter employed to re-write these support verb constructions. Figure 29 shows a concordance where Portuguese biomedical-related support verb constructions are recognized and paraphrased as lexical strong verbs or as stylistic variants. Stylistic variants *sujeitar-se a* and *submeter-se a* (∓ *to be submitted to*) are only allowed when the subject is a patient. Some lexical strong verbs are only allowed with agentive subjects. There is a strong connection between predicate-argument structure knowledge and the use of a particular stylistic variant.

nça, o cirurgião Faivre, ao	fazer uma amputação/amputar
nça, o cirurgião Faivre, ao	fazer uma amputação/efectuar uma amputação
nça, o cirurgião Faivre, ao	fazer uma amputação/realizar uma amputação
ı ser interrogadas antes de	fazer um aborto/submeter-se a um aborto
ı ser interrogadas antes de	fazer um aborto/abortar
ı ser interrogadas antes de	fazer um aborto/efectuar um aborto
ı ser interrogadas antes de	fazer um aborto/realizar um aborto
o público de saúde recusa	fazer uma operação cirúrgica/realizar uma operação cirúrgica
o público de saúde recusa	fazer uma operação cirúrgica/efectuar uma operação cirúrgica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/sujeitar-se a uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/submeter-se a uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/realizar uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/efectuar uma operação plástica
ıber se o doente consegue	fazer uma prova de esforço/sujeitar-se a uma prova de esforço
ıber se o doente consegue	fazer uma prova de esforço/submeter-se a uma prova de esforço
ıber se o doente consegue	fazer uma prova de esforço/realizar uma prova de esforço
ıber se o doente consegue	fazer uma prova de esforço/efectuar uma prova de esforço
ı médico também lhe pode	fazer uma prova de esforço para/realizar uma prova de esforço
ı médico também lhe pode	fazer uma prova de esforço para/efectuar uma prova de esforço
médico sempre vai querer	fazer um transplante de/realizar um transplante
médico sempre vai querer	fazer um transplante de/efectuar um transplante
ɔmista britânico, conseguiu	fazer uma transfusão de sangue/realizar uma transfusão de sangue
ɔmista britânico, conseguiu	fazer uma transfusão de sangue/efectuar uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/sujeitar-se a uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/submeter-se a uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/realizar uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/efectuar uma transfusão de sangue

Figure 29: Recognition and monolingual paraphrasing of biomedical-related support verb constructions (support verb construction / corresponding verb or stylistic variant)

Technical controlled languages limit language so it is easier to translate. We foresee the increasing ability to create controlled language in specific domains, which will clarify writing, making it more precise and meaningful, within the domain of the user, whether they are in linguistics, computer science, medicine or sports. In the business world, how the company conducts its business is also determined by how their documents are written.

To sum up, in the particular case of support verb constructions, the most important aspect of paraphrasing is word reduction, in particular nouns, but also prepositions, determiners and sometimes even adjectives and other grammatical words and replacement of a semantically weak verb by a strong verb. General language, but specially scientific and technical fields are rich in nouns, which correspond to the subject matter or domain terminology. So the strategy of reducing the number of support verbs and consequently the number of nominalizations and increasing the amount of morpho-syntactically related verbs by paraphrasing balances the text style. The ideas expressed in

the support verb construction can be replaced in most cases by the verb, which is a strong and simple way of moving ideas along saving on sentence length and complexity and making it easier to understand. This is also useful for people with fewer linguistic skills. Taken together, reduction of grammatical words such as determiners and prepositions is a machine friendly technique, since grammatical words create noise and give rise to bad output by many machine translation. It is important to note that, reducing the support verb constructions to single verbs is a way of tightening style in written language and making it more accessible to machine translation, but spoken language would often sound too formal if it were constructed this way. The support verb constructions form part of the complex web of less formal, deliberately vague elements that are part of the interpersonal or 'politeness' strategies that we use in spoken communication. In many cases the support verb constructions are used in spoken or colloquial situations, with the simpler verb version being more formal. Naturally, machine translation cannot take the textual element into account, and the fact that support verb constructions appear in less formal use brings with it the fact that informal, loosely structured language is more difficult for both humans and machines to translate than more formal texts using carefully constructed sentences.

Chapter 9

Evaluation

*

Chapter Nine is dedicated to evaluation. It refers to some methods for evaluating the quality of text which has been translated using machine translation. It briefly presents the pioneer work performed by Linguateca. Then it focusses on the machine translation problems presented by the support verb constructions and presents two experiments which show evidence of correction of these problems. A paraphrase suitability index and other evaluation measures are suggested.

*

9.1. **Evaluation of Machine Translation**

Evaluation of machine translation and translation tools is a complex task. For this reason, there has been lack of objective evaluation and comparative studies between systems. Human evaluation is reliable but time consuming. Automated assessment of machine translation is not as reliable but it is faster. In the last years, some automated methods of evaluation such as BLEU [Papineni et al., 2002] [Denoul & Lepage, 2005] [Callison-Burch et al., 2006b], NIST [Lee & Przybocki, 2005] and METEOR [Banerjee & Lavie, 2005] have gained popularity on the subject of the correlation between metrics and human judgements of quality of text. These metrics work by measuring the statistical closeness, in terms of n-gram co-occurrence, between a translation and the set of high quality human reference translations. However, there are a few setbacks to these methods. First, they are designed to approximate human judgement on a corpus level and do not perform well if used to measure the quality of machine translation in real situations. Second, these methods do not evaluate all the factors that need to be taken into consideration in machine translation assessment. They do not consider translation intelligibility or grammatical correctness, usefulness of the system in terms of productivity, consistency or accuracy, fidelity, appropriateness of style and register, adjustment to different domains, evaluation of textual output quality, and easing post

editing procedures. A thorough discussion of machine translation evaluation must also consider linguistic resources and tools to support the machine translation process. While these resources and tools do not perform machine translation, they do contribute to understanding of what can be done to improve machine translation. The present study tries to answer to the demand for resources and tools, by developing and enhancing dictionaries and building authoring aid software that allow controlling and pre-editing of the source language texts.

9.1.1. **The Linguateca Effort**

Little relevant machine translation evaluation was performed for Portuguese until 2002 (if any was done, there is no documentation reporting it). At the beginning of 2002, Linguateca started a process for the evaluation of the various areas of the computational processing of Portuguese, including machine translation. An evaluation group for machine translation, called ARTUR was presented in *AvalON'2003*. This group worked in the development of automated tools for machine translation evaluation. At Linguateca PoloCLUP evaluation experiments were carried out, which resulted in the creation of some useful tools: TrAVA, METRA, EVAL, and Boomerang. The experiments are described in [Santos et al. 2004] [Maia & Barreiro 2007] [Sarmento et al. 2007] [Sarmento, 2007], and most of these tools are available online.

In the last few years, our interest contributed to an understanding of the problems associated with preserving meaning throughout the translation process, and also of the understanding colloquialisms, metaphorical usage and disambiguation. Our first approach was to review and document some of the most critical linguistic deficiencies for the English-Portuguese machine translation language pair. A selection of some machine translation gray areas were described in [Barreiro & Ranchhod, 2005] and in [Maia & Barreiro 2007].

Machine translation results are far from perfect. A paraphrasal relevance assessment test was run to appraise the relevance of paraphrases of support verb construction to machine translation quality improvement. In the first place, this test confirmed our hypothesis that machine translation results for support verb constructions are poor. It then showed that paraphrases of support verb constructions help improve machine translation results, and confirmed that a linguistically based paraphrasal tool helps

control source language so as to facilitate the translation of texts submitted to machine translation.

The next three sections will present details on the assessment of support verb construction paraphrases. Section 9.1.2 presents empirical evidence that proves that machine translation of support verb constructions need to be properly addressed. Throughout this dissertation, we have argued that it is possible to improve machine translation results if support verb constructions are replaced with strong verbs. Sections 9.1.3 and 9.1.4 present two experiments that confirm our initial hypothesis that pre-processing of support verb constructions by verbs or verbal expressions helps systems to get better results from machine translation. In the first experiment the paraphrasing was manual. In the second experiment, the paraphrasing was automated. Both experiments follow a series of common, repeatable procedures:

(1) Identification of support verb constructions in corpora
(2) Creation of a corpus of support verb construction sentences
(3) Translation of the corpus sentences with machine translation systems available online
(4) Collection of the results for evaluation
(5) Replacement of the support verb construction sentences in the support verb construction corpus with equivalent verbs or verbal expressions in the human-controlled experiment, and with equivalent verbs only in the automatic experiment as a pre-processing step
(6) Translation of the pre-processed version of the corpus with machine translation systems available online
(7) Collection of the results
(8) Comparison of the results of both translation sets, the support verb construction translations and their paraphrased sentence translations

9.1.2. **Machine Translation Problems with Support Verb Constructions**

A brief analysis of the results for freely available machine translation systems on the Internet demonstrates that the translation of multiword expressions is currently a problem area for any machine translation system, whether statistical or rule-based. The

ambiguity of these expressions can lead to loss of meaning, mistakes, unclear or unsuitable translations, when there is no linguistic (syntactic-semantic) knowledge associated with them. Translation results extracted from METRA prove that machine translation engines are unsuccessful at handling the translation of support verb constructions. A literal and unnatural translation is provided by some machines. Table 6 below, illustrates how machine translation engines handle translation of the Portuguese support verb construction *tomar uma decisão* (*make a decision*) in the sentence *I can't make a decision about anything these days* [COMPARA]. For example, for some translation engines, the English support verb construction *make a decision* is translated into Portuguese as *fazer uma decisão* instead of *tomar uma decisão* or even as the strong verb *decidir*, which represent its optimal paraphrase. This inaccuracy means that the English support verb *make* is directly translated into the Portuguese support verb *fazer* (default translation), instead of being recognized as part of the support verb construction which embeds semantic meaning as a whole.

Translation Engine	Translation Results
Amikai	que eu não posso **fazer para uma decisão** sobre qualquer coisa estes dias.
FreeTranslation	Eu não posso **tomar uma decisão** sobre algo hoje em dia.
WorldLingo	Eu no posso **fazer a uma deciso** sobre qualquer coisa estes dias.
E-Translation Server	Não posso **tomar uma decisão** sobre qualquer coisa estes dias.

Table 6: METRA search results for machine translation of the English support verb construction *make a decision* into Portuguese (01/11/2008)

Similar problems are shown in the translations of Portuguese support verb construction *tomar uma decisão* into English, where the support verb *tomar* is being translated literally as *take*. This is not an incorrect translation, as the British National Corpus (BNC) presents about 6% of the examples with *a/the decision* occurring with *take*. However, the 40% results with *make* show that *take* is not the preferred support verb to go with the predicate noun *decision*. Machine translation is not addressing the appropriateness or suitability issue effectively.

We have tried to replace some support verb constructions with verbs and verified that overall machine translation engines showed significantly better results. For example,

machine translation engines are unanimous in choosing the Portuguese verb *decidir* as the correct translation for the English verb *decide*, as Table 7 shows.

Translation Engine	Translation Results
Amikai	que eu não posso **decidir** sobre nada estes dias.
FreeTranslation	Eu não posso **decidir** sobre algo hoje em dia.
WorldLingo	Eu no posso **decidir-se** sobre qualquer coisa estes dias.
E-Translation Server	Não posso **decidir** sobre nada estes dias.

Table 7: METRA search results for machine translation of the English verb decide into Portuguese (01/11/2008)

This pre-editing, or controlled language writing by paraphrasing, improves translation results and makes output sentences more comprehensible overall. This proves that, if we consider pre-editing of the input sentences where support verb constructions occur, changing each instance into a verb (whenever that is linguistically suitable), we are not changing the meaning of the source sentence and we are giving the machine translation engine a distinctly better chance of improving the output result, by filtering out some noise, i.e., the weak verb. The support verb construction *make a decision* is a stylistic alternative to the verb *decide*, where neither the support verb *make* nor the determiner *a* adds any meaning to the expression. In fact, in support verb constructions, the support verb is often void of meaning. Trying to translate them brings additional difficulties to machine translation systems, which is unnecessary until/unless they become more sophisticated. Our paraphrasing system allows several possibilities. However, paraphrasing by simplification proved to be the most suitable for publicly available machine translation systems. The ideal framework is that machine translation systems are not limited to one possibility, as long as they translate with precision. Besides, we believe that it is pointless to challenge one limited machine translation system with structures that we know *a priori* this system cannot translate well. For equivalent paraphrasing, the support verb must be recognized as part of a support verb construction which must be considered as a single semantic unit. The default assumption of all machine translation systems which cannot discern whether a word, in this case a support verb, adds semantic meaning to a phrase, is to assign equal semantic value to each word individually, unless otherwise instructed. The system fails by incorrectly assigning semantic value to a support

verb, resulting in a loss in paraphrasing capability of the output sentence. This is the problem of direct translation.

The ambiguity that support verb constructions can carry can lead to loss of meaning, mistakes and unclear translations, when there is no syntactic-semantic information associated to them. Again, we used METRA to translate unedited sentences extracted from a corpus of support verb plus biomedical-related term combinations found on the Internet and analyzed the outputs. After that, we pre-edited those sentences including subject and object whenever needed, paraphrased them accordingly and compared results. Table 8 presents an example of a sentence extracted from a less formal, oral-type of text, namely from a doctor's blog [Blog-WWH]. The results of the machine translation engines were the following for the original corpus sentence and for the paraphrased sentences:

*Por falta de condições técnicas, ele foi removido para o Hospital das Clínicas, onde **se fez uma amputação** a nível de ombro.*	
FreeTranslation	For lack of technical conditions, he was removed for the Hospital of the Clinics, where **was done an amputation** in terms of shoulder.
WorldLingo	Due to conditions techniques, it it was removed for the Hospital of the Clinics, where if the shoulder level **made an amputation**.
*Por falta de condições técnicas, ele foi transportado para o Hospital das Clínicas, onde **os médicos amputaram o seu braço** ao nível do ombro.*	
FreeTranslation	Due to conditions techniques, it it was carried to the Hospital of the Clinics, where **the doctors had amputated its arm** to the level of the shoulder.
WorldLingo	For lack of technical conditions, he was transported for the Hospital of the Clinics, where **the doctors amputated his arm** level with the shoulder.
*Por falta de condições técnicas, ele foi transportado para o Hospital das Clínicas, onde **o seu braço foi amputado** ao nível do ombro.*	
FreeTranslation	for lack of technical conditions, he was transported for the Hospital of the Clinics, where **arm was amputated** level with the shoulder.
WorldLingo	Due to conditions techniques, it it was carried to the Hospital of the Clinics, where its **arm was amputated** to the level of the shoulder.

Table 8: Machine translation results for sentences with support verb constructions and results for pre-edited paraphrases with predicate-argument relation knowledge (26/05/2008)

In Table 8, the translation of the support verb constructions result in corresponding English support verb constructions, where the support verb was translated literally. The sentences are confusing and unclear. WordLingo assigns the event *amputation* to the wrong subject. This error illustrates that it is risky to translate support verb constructions

by using machine translation. On the contrary, the paraphrasing applied to the sentence presented more comprehensible translated text when we replaced the support verb construction by a verb and added a subject and an object to the sentence. With a semantically expressed agent (*médicos > doctors*) and the part of the body that was affected as the result of the procedure (*o seu braço > his arm*), the sentence *os médicos amputaram o seu braço* (*the doctors amputated his arm*) became clearer and easier to translate. Similar good results were obtained when we replaced the support verb construction with a passive construction and specified the object (*o seu braço foi amputado > his arm was amputated*).

These simple word replacement operations show that a linguistic analysis of the syntactic and semantic arguments of the predicate allows the disambiguation of the support verb construction. Translating these types of support verb constructions is very difficult unless a linguistic analysis of the predicate-argument structure is performed. These results show enough evidence to argue that paraphrasing support verb constructions with other verbal constructions and a well-defined predicate-argument structure help produce better and "*safer*" machine translation results.

To sum up, empirical evidence shows that application of linguistic knowledge to proper handling of support verb constructions by machine translation systems or natural language processing applications is effective. We believe that our methodology leads to attainable paraphrasing translation solutions. This work demonstrates that we can create an instrument of some utility to the research community, which has good applicability in machine translation.

9.1.3. **Experiment on Machine Translation of Human-made Paraphrases**

This experiment consisted in choosing between the best translation of a sentence containing a support verb construction and the best translation of the same sentence where the support verb construction had been paraphrased from the translations retrieved by machine translation systems. Five distinct English and five distinct Portuguese support verbs were tested. The English support verbs selected were *to give*, *to do*, *to take*, *to make*, and *to have*. The Portuguese support verbs selected were *fazer* (*to do*/*make*), *ter* (*to have*), *dar* (*to give*), *ir* (*to go*), and *tomar* (*to take*). A total of one hundred sentences were selected for each language from COMPARA's parallel corpora,

twenty support verb construction sentences for each support verb. Table 9 illustrates the sentences selected for each English support verb. 1 to 5 a) represent the sentences with support verb constructions and 1 to 5 b) represent the paraphrases of those sentences, resulting from a human-controlled replacement process, the manual pre-edit of the support verb construction with a verb or verbal expression. Table 10 illustrates the kind of sentences selected for each Portuguese support verb. 1 to 5 a) represent the input sentences with support verb constructions and 1 to 5 b) represent the paraphrases of those sentences.

1. a) It was up to me to **make a decision**.
b) It was up to me to **decide**.
2. a) I decided it would be fair **to give him notice** after a month that I didn't want him in my house any longer.
b) I decided it would be fair **to notify him** after a month that I didn't want him in my house any longer.
3. a) The inquest is to **take place** this afternoon.
b) The inquest is to **occur** this afternoon.
4. a) I have a dictionary and **do my best** to write correctly.
b) I have a dictionary and **try** to write correctly.
5. a) Let's **have a look** at this wound,[…]
b) Let's **look** at this wound,[…]

Table 9: Sample of sentences with support verb constructions and their paraphrases for 5 selected English support verbs

1. a) Dirigiu-se ansiosamente à porta para **dar as boas-vindas ao** visitante e ao mesmo tempo para acender as luzes,
b) Dirigiu-se ansiosamente à porta para **saudar o** visitante e ao mesmo tempo para acender as luzes,
2. a) Catherine corou porque se sentiu como se estivessem a **fazer troça** dela.
b) Catherine corou porque se sentiu como se estivessem a **troçar** dela.
3. a) Era uma grande maçada ter de vestir-me outra vez para **ir à rua**, mas estava determinado a levar a minha experiência até ao fim.
b) Era uma grande maçada ter de vestir-me outra vez para **sair**, mas estava determinado a levar a minha experiência até ao fim.
4. a) Trabalhara principalmente nas urgências, mas começara a **ter medo do** efeito que esse trabalho tinha nela.
b) Trabalhara principalmente nas urgências, mas começara a **recear o** efeito que esse trabalho tinha nela.
5. a) Quis **tomar nota de** tudo, antes que sobreviesse o esquecimento.
b) Quis **anotar** tudo, antes que sobreviesse o esquecimento.

Table 10: Sample of sentences with support verb constructions and their paraphrases for 5 selected Portuguese support verbs

Ten human testers were asked to translate both the sentences with the support verb constructions and the sentences with the support verb construction paraphrases with METRA, the online meta-translator that retrieves translation results from different

systems. Testers were presented with a package made of an instructional file and a spreadsheet file with the input sentences. In the input sentences file, there were two columns, containing the selected twenty sentences on each column. The first column contained the support verb construction sentences and the second column contained the sentences with the support verb construction paraphrases, which testers needed to submit to machine translation (a total of forty sentences for each tester to submit to machine translation and to evaluate: twenty sentences with support verb constructions and twenty paraphrases). The expressions that testers needed to focus on in the evaluation task were marked in bold. Sentences in both columns differed only on those bolded expressions and only those expressions were being evaluated, not the whole sentence. The instructional file included all the information to be followed by the testers, to be read before starting the evaluation work. According to the instructions, a screenshot of each METRA translation results for each sentence had to be stored in a third file created by the testers for auditing purposes, i.e., screen shots were kept to show that the correct METRA result was stored. Testers could also add feedback or text comments considered relevant in that file. From the METRA window, testers would select the best translation output regarding the bolded expression and paste it into the spreadsheet file results sheet. The same procedure was repeated for each paraphrased sentence. Finally, testers were asked to choose the best sentences from the two best results, the support verb construction best result, and the support verb construction paraphrase best result. If they considered that the quality of the translations was similar, they were instructed to write "SAME" in the corresponding cell for the "Best result of all" column, as can be seen in Table 11.

Tester	**Input** **Pair of Sentences**	**Best result for support verb construction sentence**	**Best result for support verb construction paraphrase**	**Best result of all (#1 translation)**
Tester 1	A: Quer ir **dar uma vista de olhos** ao museu ou prefere descer já? B: Quer ir **ver** o museu ou prefere descer já?	Want to go **give a sight of eyes** to the museum or is going to come down already?	It wants to go **to see** the museum or it prefers to go down already?	It wants to go **to see** the museum or it prefers to go down already?
Tester 2	A: Sim, deves **ter razão**. B: Sim, deves **estar certo**.	Yes, deves **be right**.	Yes, you **should be right**.	Yes, you **should be right**.
Tester 3	A: resolvi tentar ficar um bocado de olhos fechados para **ter uma ideia de** como seria ser cego B: resolvi tentar ficar um bocado de olhos fechados para **imaginar** como seria ser cego	I decided to try to be a bit of closed eyes **to have an idea of** as it would be to be blind	I decided to try to be a bit of closed eyes **to imagine** as it would be to be blind	I decided to try to be a bit of closed eyes **to have an idea of** as it would be to be blind

Tester 4	A: Deixámo-los a **tomar conta da** casa, com a minha mãe. B: Deixámos-los a **cuidar da** casa, com a minha mãe.	Deixámo-los **to take care of** the house, with my mãe.	Deixámo-os **to take care of** the house, with my mãe	SAME

Table 11: Sample of the Portuguese-English human evaluation results sheet

Figure 30 and Figure 31 show METRA results for the translation of paraphrases from Portuguese into English. Figure 30 illustrates the translation results for a sentence with a support verb construction and Figure 31 illustrates the translation results for a sentence with a verb.

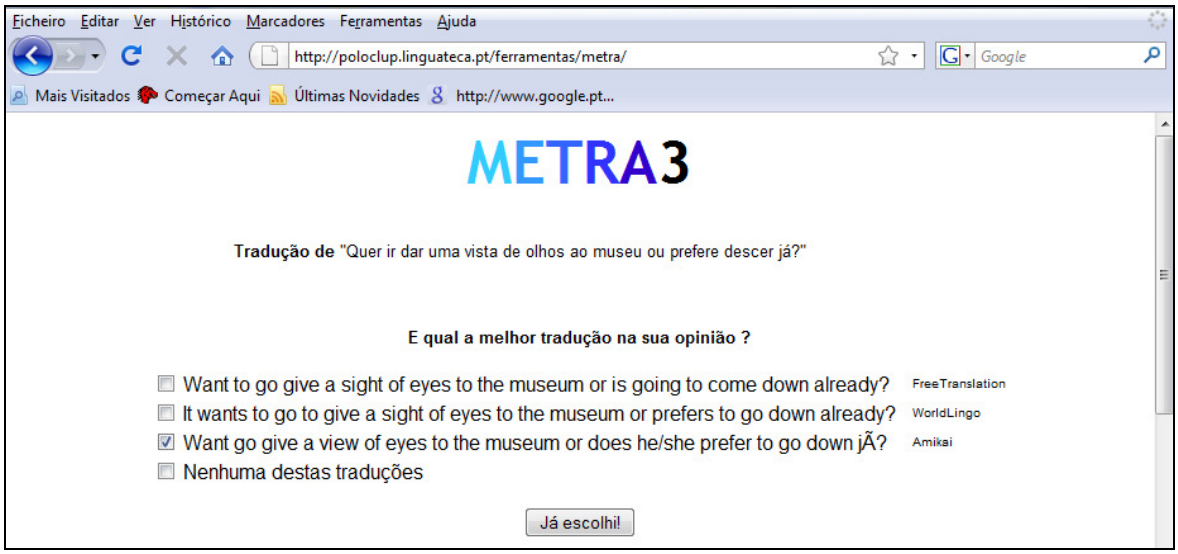

Figure 30: METRA results for Portuguese-English translation of a sentence with a support verb construction

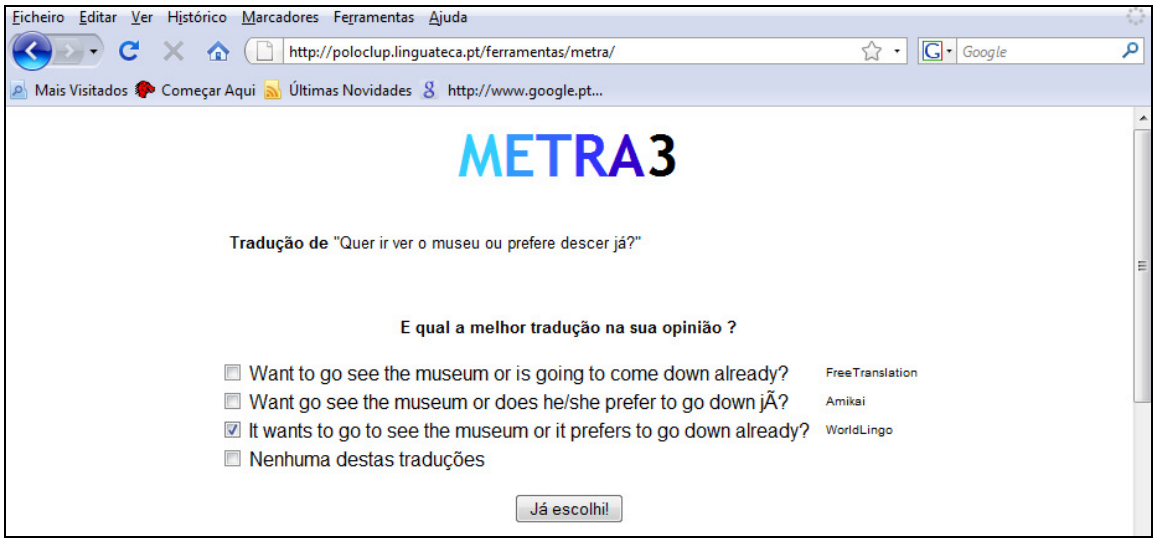

Figure 31: METRA results for Portuguese-English translation of a sentence with a verb (paraphrase of the sentence with a support verb construction)

The methodology is repeatable; it uses familiar software. The spreadsheet files can be re-used. It is a consistent test and user-friendly. It was thought to minimize the time that the tester would spend in performing the test. The data captured in the spreadsheet file is easy to process, and is complete. It does not need the screen shots for the analysis of the results. The screen shots are the documentation that verifies the quality and veracity of the results data.

For English to Portuguese translation, testers considered that 26% of the sentences with support verb constructions presented the best results; 57% of the sentences with their paraphrases presented the best results, and 17% presented equally good results. For Portuguese to English translation, testers considered that 30% of the sentences with support verb constructions presented the best results; 51% of the sentences with their paraphrases presented the best results, and 19% presented equally good results. The evaluation results clearly confirmed that the paraphrases have helped the systems produce better results. It appears that linguistically well studied paraphrases increase the quality of machine translation results. Human evaluation is difficult to attain. In order to obtain significant results, it would require a minimum of 5 testers per support verb group for each language. It was not possible to gather that many testers. Not all testers have the same judgement capability. Only testers with good linguistic knowledge and translation skills are able to perform good tests. Advanced students of translation studies performed

the English to Portuguese side of this experiment and bilingual or near bilingual English-Portuguese speakers with good linguistic knowledge performed the Portuguese to English side of the experiment.

9.1.4. **Experiment on Machine Translation of Automatically Generated Paraphrases**

This experiment consists of two steps: the automated pre-processing of support verb constructions and conversion into (strong) verb expressions and on the machine translation of the automatically generated paraphrases. We selected from COMPARA, a parallel corpus of English-Portuguese fiction, all sentences where the infinitive form of the Portuguese verbs *fazer* (*to do*), *dar* (*to give*), *pôr* (*to put*), *tomar* (*to take*) and *ter* (*to have*) occurred with a noun or with a left modifier and a noun. First, we manually classified these combinations as to whether they corresponded to support verb constructions or not. We confirmed that these verbs occur very frequently in a support verb construction. 89% of the occurrences of *dar*, 88% of *tomar*, 77% of *pôr*, 47% of *fazer* and 20% of *ter* were in a support verb construction. This means that globally in 64.2% of their occurrence, these verbs are used as support verbs.

Subsequently we selected randomly a sub-corpus with 500 sentences (100 for each selected verb), containing instances of only support verb constructions. We classified them manually and compared these results with the results obtained automatically. We tried to have constraining recognition rules so that paraphrasing would be more precise. Currently, our paraphrasing tools can recognize 62.6% of the support verb constructions with high scores in precision. Furthermore, they not only recognize the support verb constructions, as they also paraphrase them with high degree of success. Table 12 shows the results of the support verb construction recognition (precision and recall) and the results (precision) of our automatic paraphraser.

	SVC Recognition Precision	SVC Recognition Recall	SVC Paraphrasing Precision
Pôr	73/73 - 100%	73/100 – 73%	72/73 - 98.6%
Tomar	75/75 - 100%	75/100 – 75%	68/73 - 93.1%
Ter	65/65 - 100%	65/100 – 65%	59/65 - 90.7%
Dar	57/60 - 95%	57/100 – 57%	46/51 - 90.1%
Fazer	43/45 – 95.5%	43/100 – 43%	40/45 - 88.8%
Average	62.6/63.6 - **98.4%**	62.6/100 - **62.6%**	57/61 - **93.4%**

Table 12: Evaluation of simultaneous recognition and paraphrasing of support verb constructions

After the paraphrasing of support verb constructions was done by ReWriter and the results were evaluated by us, we tried to translate some of the automatically paraphrased sentences and compare the results against the results of the original sentences with the support verb constructions. The pre-processing of support verb constructions into strong verbs of the first experiment was reused, so the original sentences are from the same source, COMPARA. We selected only 50 Portuguese sentences, randomly. The procedures included the implementation of the following routines: (i) ReWriter suggested paraphrases for the support verb constructions in parallel with original sentences; (ii) both pre-processed sentences (paraphrases) and the original text are submitted to machine translation process on METRA and the output results for both original and pre-processed sentences are compared. From the total of 50 pairs of sentences, 29 (58%) of the best translations were of automatically generated paraphrases, only 9 (18%) were of support verb constructions and 12 (24%) were equally bad or equally good. When support verb constructions were identified and replaced with semantically equivalent verbs or other verbal expressions as a pre-processing step to translating, an average 40% improvement was observed in the evaluated quality of the results of the machine translation of the expressions replacing the support verb constructions. So, we can conclude that the automated paraphrasing by ReEscreve helps both the machine and the results. This experiment confirms the tendency shown by the previous experiment where paraphrasing was performed manually.

Even if the previously described experiments indicate that paraphrases help improve translation scores, they do not tell us about the qualitative aspects of the paraphrases. In order to do that, we have created a paraphrase suitability index, which measures the type and relevance of each paraphrase.

9.2. **Paraphrase Suitability Index**

A paraphrase suitability index was created to help select the best support verb construction paraphrase in a particular corpus (domain-specific) and produce the appropriate text for the user. Different types of text show different distributions of alternatives. An alternative is the paraphrasing capability that meets the requirements for a certain domain. It meets controlled language standards. The best alternative is the most appropriate result that takes into consideration both the source and the target language.

The paraphrase suitability index ranks paraphrases from the closest in meaning to the most distant paraphrase ranging/varying from 1 to 5. Index 1 corresponds to an identical or exact paraphrase. It represents the best paraphrasing capability, where there is a full semantic match with:

(i) a semantically and morpho-syntactically related verb, such as *dar um abraço a* = *abraçar* (*give a hug to* = *hug*);

(ii) a semantically related but morphologically unrelated verb, such as *dar aulas* = *ensinar / leccionar* (*to give classes* = *to teach*) or *abrir caminho* = *prosseguir* (*to make way* = *proceed*);

(iii) a related predicate adjective construction, such as *ter cansaço* = *estar cansado* (∓ **to have tiredness* = *to be tired*) or *ter preocupação com* = *estar preocupado com* (∓ **to have concern about* = *to be concerned about*), that only use *be* in English;

(iv) a related predicate noun construction, such as *fazer uma viagem* = *efectuar/realizar uma viagem* (*to go on a trip*);

(v) a semantically related verb with an adverbial that is morphologically related to an adjective in the original support verb construction, such as *dar uma olhada (rápida)* = *olhar rapidamente* (*give a quick glance* = *glance quickly*) or *fazer uma aterragem segura* = *aterrar em segurança* (*to make a safe landing* = *to land safely*.

In index 1 there is normally 100% of meaning preservation and, in most cases, word count reduction. Except for the cases where the verb is not morphologically related to the nominal in the support verb construction, the root of the predicate remains the same, viz.

V-N pair *visita-visitar*; V-A pair *cansaço-cansado*). Meaning preservation and/or morpho-syntactic closeness (same morphological root, same syntactic behavior) is considered a good criterion to define index 1. Index 2 is used for approximately equal paraphrases. 'Approximately equal' means that there is a very close semantic match, such as in *tomar um duche* ≈ *lavar-se* (*to take a shower* ≈ *to shower*). *Lavar-se* can mean either *take a bath* or *take a shower*. Index 3 is used for inferred meaning paraphrases. These paraphrases express indirectly the same meaning. They contain, involve or indicate by inference, association, or necessary consequence. For example, in the paraphrasing capability *preparar uma bebida quente* ⊃ *beber qualquer coisa quente* (*to make a hot drink* ⊃ *to drink something hot*), there is preservation of at least two words and the correspondance has a relatively high accuracy. Index 4 is used for approximate paraphrases such as *estar cansado* = *cansar-se*. Index 5 is used for quasi paraphrases. Quasi-paraphrases are not semantically identical, but they convey similar information, viz. *fazer uma sessão de boas-vindas* > *saudar* (∓ *to make a welcome session* => *to greet*).

It is important to establish a measurement representation so that paraphrases can be evaluated scientifically, notwithstanding the human judgment behind such representation, even if such an index is not consensual.

Paraphrases that have high suitability index values help produce a translation that uses information from the corpus without changing its meaning, and help increase machine translation accuracy by avoiding preventable errors in machine translation, such as literal translation.

The simplest types to implement are the paraphrasing capabilities between support verb constructions and verbs. These verbal constructions are the most appropriate for automatic paraphrasing and machine translation because of many reasons, among them:

(i) they make sentences easier for the machine to process clearer;
(ii) they help reduce sentence length and create shorter paragraphs;
(iii) they reduce the number of prepositions, which make translation considerably more complicated. For instance, *fazer uma apresentação do livro* = *apresentar o livro* (*make a presentation of the book* = *present the book*), where the second phrase is shorter and trouble-free;
(iv) they reduce the number of nominalizations;

(v) they replace general verbs with precise verbs;

(vi) in some cases they help text cohesion.

However, an elementary support verb construction that can be paraphrased by a non-elementary support verb construction that uses a stronger support verb is also a very good candidate for paraphrasing and machine translation. Even though they do not help reduce the number of words in the sentence, they may help make the sentences clearer and can be very useful for stlylistic reasons.

9.3. **Other Evaluation Measures**

In our study, we also marked the instances where:

(i) the paraphrase improves clarity (the > symbol means that the paraphrase has more fitting usage)

(ii) the paraphrase maintains clarity (the = sign means that both the original support verb construction and its paraphrase are equally fitting)

(iii) the paraphrase reduces clarity (the < symbol means that the original support verb construction has a more suitable usage than its paraphrase).

Additionally, we established a usage category as well, marking original support verb constructions as regular (R), figurative (F), colloquial (C), slang (S), idiomatic (I), technical (T) or metaphorical (M). The tables below show a classification of paraphrases taking into account Portuguese to English (Table 13) and English to Portuguese (Table 14) manual translations of sentences where the support verb constructions occur, extracted from parallel corpora COMPARA. Support verb constructions were identified in bold, both in the "Source and Target Sentences" columns. "Support Verb Construction Structure" column illustrates the source sentence support verb construction internal arguments. "Possible Translation" column shows a few "good" translations for the source sentence support verb construction. "Paraphrase" column shows simple short paraphrases only, mostly simple verbs for the source sentence support verb construction. "Paraphrase Rating" column shows if the paraphrase is better, worse or the same as the original source support verb construction. "Suitability Index" column grades the appropriateness

of the paraphrase. Finally, "Usage Category" column specifies the kind of paraphrase, in terms of its usage in the text language.

Source Sentence			Target Sentence		
Quero ter primeiro uma conversa preliminar, que lhe **tomará** pouco **tempo**.			I want a preliminary talk first, which will **take** only a little of your **time**.		
Support verb construction Structure	**Possible Translation**	**Paraphrase**	**Paraphrase Rating**	**Suitability Index**	**Usage Category**
Vsup(tomar) (Adv) Npred(tempo)	Be short	Ser breve/rápido	>	2	(R)egular

Source Sentence			Target Sentence		
Devias **dar uma vista de olhos** pela literatura do Médio Oriente.			You should **take a look at** some Middle Eastern literature.		
Support verb construction Structure	**Possible Translation**	**Paraphrase**	**Paraphrase Rating**	**Suitability Index**	**Usage Category**
Vsup(dar) uma vista de olhos	Take a look at Look at	Ver	>	1	(I)diomatic

Source Sentence			Target Sentence		
A comadre informou de semelhante coisa ao Leonardo-Pataca, e este apresentou-se para **tomar conta de** seu filho.			The comadre informed Leonardo-Pataca of the situation, and he presented himself to **take charge of** his son.		
Support verb construction Structure	**Possible Translation**	**Paraphrase**	**Paraphrase Rating**	**Suitability Index**	**Usage Category**
Vsup(tomar) Npred(conta) Prep(de) NP	Take care of	Cuidar de NP	>	2	(R)egular

Table 13: Classification of Portuguese support verb construction paraphrases

Source Sentence		Target Sentence			
It was up to me to **make a decision**.		Cumpria-me **decidir**.			
Support verb construction Structure	**Possible Translation**	**Paraphrase**	**Paraphrase Rating**	**Suitability Index**	**Usage Category**
Vsup(make) Det(a) Npred(decision)	Tomar uma decisão = Decidir	Decide	=	1	(R)egular

Source Sentence		Target Sentence			
Greatly surprised, I endeavored to **make my way** toward it, as it appeared to be but a few feet from my position.		Espantado, esforcei-me por **me dirigir** para essa luz que parecia estar apenas a alguns pés de distância.			
Support verb construction Structure	**Possible Translation**	**Paraphrase**	**Paraphrase Rating**	**Suitability Index**	**Usage Category**
Vsup(make) ONE'S Npred(way)	Dirigir-se	Proceed	>	1	(I)diomatic

Source Sentence		Target Sentence			
But he's so transparent, you can't **take offence**.		Mas é tão óbvio, que ninguém **leva a mal**.			
Support verb construction Structure	**Possible Translation**	**Paraphrase**	**Paraphrase Rating**	**Suitability Index**	**Usage Category**
Vsup(take) Npred(offense)	Ofender-se = ficar ofendido = levar a mal	Be offended	=	2	(I)diomatic

Source Sentence		Target Sentence			
Well, **take my word for** it, they're dead boring.		Pois olhe, **acredite** no que eu lhe digo, são chatos de morrer.			
Support verb construction Structure	**Possible Translation**	**Paraphrase**	**Paraphrase Rating**	**Suitability Index**	**Usage Category**
Vsup(take) N(word) for it	Acreditar	Believe	>	1	(I)diomatic

Table 14: Classification of English support verb construction paraphrases

Chapter 10

Conclusion

*

Chapter Ten summarizes the main points treated in the dissertation and emphasizes the relevance of paraphrases to natural language processing, particularly to machine translation, presenting a list of the most important benefits. Finally, it presents a synopsis of future work goals, based on the work completed here.

*

The use of semantically weak verbs and the lack of predicate-argument knowledge give rise to ambiguity that compromises precision and spoils machine translation. In this dissertation, we have tried to answer the question of whether paraphrase information can improve machine translation output and how the analysis and formalization of their paraphrases can contribute to machine translation in general. We have demonstrated how linguistic analysis and computational formalization of paraphrases for support verb constructions represent a good basis for machine translation development and machine translation evaluation. The discovery process has provided results in two areas. First, it has led to the creation of a primitive multiword expression electronic dictionary that addresses monolingual Portuguese and bilingual Portuguese-English-Portuguese paraphrases of equivalent meaning between support verb constructions and their verbal counterparts, such as *fazer/realizar/efectuar uma investigação sobre* = *investigar sobre* > *to make/perform an investigation on* = *to investigate*. Second, it has helped to further specify the definition of multiword expressions, and extend the scope of current dictionary functionality. The interface between user and software that is presented is not finished yet, but we believe that the sub-task is well-understood and this interface can be optimized and, hopefully, be usable and integrated into the larger task of machine translation, just as the single word electronic dictionary has already been integrated into machine translation systems.

Our work based on support verb constructions illustrates what can be done with ReWriter for any kind of multiword expression. The method we used is repeatable and extendable, and we believe it will provide good results when applied to a bigger general language corpus. Furthermore, the tools we created are extensible to cover larger and more complex linguistic phenomena, including sentence level paraphrases that can be used in improving authoring aids or translation and can be also integrated in translation memory software. Linguistic knowledge is useful to machine translation development, because it permits deeper understanding of source text, and it provides a successful methodology to analyze paraphrasing, given that paraphrasal intelligence is crucial in both machine translation development and machine translation evaluation. On the other hand, ParaMT can evolve in a way similar to ReWriter and be used as on-line linguistic aid for translators so they can determine the best translation (for human evaluation purposes), and for automated machine translation evaluation.

While this research is intended to find a place in ideal machine translation, both resources and tools developed so far are useful from a monolingual and bilingual point of view. Monolingual benefits consist in simplification of pre-translated source text. Paraphrasing renders the text less complex, less wordy and sometimes less ambiguous. Converting support/weak verbs into lexical strong verbs helps to simplify and reduce the number of words in a text, providing better quality results. Converting elementary support verb constructions into less neutral stylistic variants brings some additional semantic weight to the expression, resulting also in an easier translation. These paraphrasing techniques also cause a positive impact on translation cost, in circumstances where word count or "white space" is sensitive (in contexts where the cost of paper, or computer display space are relevant). In addition, the rewriting software is environmentally friendly in the sense that it reduces the need to print the text to be edited. The edits can be made directly online and the color factor helps visualize and distinguish the different types of edits (green, when the expression was unchanged, blue when it was provided automatically and red when the user added as a suggestion).

Bilingual benefits consist in reducing ambiguity and verbosity and providing better automated machine translation. Standardized technical and specialized languages, such as the biomedical technical language, which use more restricted lexicons or terminologies and syntactic conventions, can add to controlled rules, a few more sophisticated and

linguistic knowledge-based resources and techniques to improve machine translation of support verb constructions. The examples we presented in the paper on this type of language served to prove unmet demands on the sophistication of machine translation resources. But in this domain, support verb constructions might be used in clearer and controlled language texts written by technical writers and/or professionals who have a good knowledge of the terminologies and writing conventions for these texts. However, the major contribution of this work seems to be the improvement of machine translation of everyday more mundane, popular and less conventional language, namely of the language used on the Internet, blogs, e-mails, etc. where people are less preoccupied with linguistic and stylistic conventions. That obviously includes a more relaxed style of reporting technical issues. Machine translation has proved to become increasingly popular for this oral-type and more ephemeral communication. Linguistically enriched systems which handle support verb constructions, frozen and semi-frozen expressions and idioms well will enable the general public to communicate more freely and more understandably across the Internet with people speaking different languages in situations where they would not require a professional translator or a machine translation optimized package. We believe that, a machine translation program that offers a correct translation of support verb constructions, either via direct phrasal translation or paraphrases is moving a step forward.

There is a vast amount of work to be done to complete the bodies of knowledge associated with natural language processing of the major languages spoken worldwide, including English and Portuguese. Until that work is done, the best that machine translation can offer is to sell the ideas of gisting, and extend electronic dictionaries, and so try to bridge the gap caused by the linguistic limitations of current machine translation systems. The underlying science of linguistics is still relatively young and the number of people trained in the science is small in comparison to the task at hand. However, it is required to build linguistic intelligence into machine translation systems. The work that is unavoidable centers on creating system logic that can process natural language. This work implies focusing on grammar rules and formalisms. Present results indicate that we need to focus on the linguistic contributions we can make to machine translation systems in overcoming challenges imposed by areas such as syntax and semantics, and learn to find synergistic ways to use the benefits offered by different approaches.

As we enlarge existing resources and create new ones and improve our tools for new functionalities, simplify and fine-tune the preliminary results that we have presented in this dissertation, we are aware of the value of bringing together several functions. We see the usefulness of NooJ as a basis on which to build language resources and enhancement of its machine translation capabilities, but we believe that only collaborative efforts that bring specific, relevant skill sets together will benefit from the creation of larger and more sophisticated linguistic resources so that we can achieve our goals in paraphrasing and machine translation. We see the need to better integrate human interfaces and computing systems to accomplish better quality results in natural language processing, especially in complex applications such as machine translation.

10.1. **Relevance of Paraphrases**

This section needs to be finalized by pointing out the relevance of paraphrases to natural language processing in general. They are most crucial to machine translation, but they are important in several applications for different aspects because they contribute to: (i) simplification and homogenization of sentence production used in controlled-language; (ii) pre-editing that shortens text - this is especially valuable if the translator is reading non-native sources, such as support verb constructions are difficult to understand for the non-native speaker and they are uncommon or non-existent in certain languages which do not have the same kind of syntactic-semantic behavior; (iii) facilitating machine translation and reducing word count without changing the meaning of the text; (iv) objectivity and elimination of redundancy or undesirable stylistic choices, presenting other possible and often preferable substitutes; (v) offering alternatives that enhance meaning to machine translation source text; (vi) establishing metrics for machine translation evaluation. In addition to the usefulness of paraphrases in paraphrasing models in the domains of natural language processing, paraphrases are also useful in other domains, such as the teaching of both native and foreign language.

10.2. **Future Goals**

As we look to employ what we have learned from machine translation, we uncover more detail regarding the specificity of paraphrases. Some paraphrases are always first-rate,

precise and do not have variable meanings. Other paraphrases require adverbial or adjectival qualification or other inserts to fix their appropriateness and/or usability. As a result of our current ongoing research in analyzing, paraphrasing and translating support verb constructions, we anticipate several outcomes:

(1) Pooled bilingual lexical resources. This includes the assembling of a bilingual corpus of support verb construction lexicalizations, such as Portuguese *fazer uma apresentação de* (*to make a presentation of*) or English *take a walk* when compiling sets of support verb construction translations and equivalent paraphrases for a particular language pair. This resource can be used by ParaMT for refinement of the automated extraction and generation of short paraphrases.
(2) Rules and requirements for finding short paraphrases for the support verb constructions. We are aware that creating more precise definitions of support verb constructions/paraphrases and other subtypes of multiword expressions is crucial for the advancement in machine translation. The field of machine translation needs the matter-of-fact practical principle described above or a similar one; linguistic enhancement to create more sophisticated machine translation standards so that systems produce more human-like results.
(3) Identification and construction of Portuguese-English and English-Portuguese syntactic-semantic correspondence rules for the support verb constructions and understanding of study transfer mechanisms between the two languages.
(4) Acquisition of a stable and functional methodology, usable by anyone working on machine translation or bilingual lexicography. This research is an approach which can operate within the ideal machine translation tool and can be applicable to different levels of linguistic phenomena.
(5) The ultimate benefits of this research are the creation of useful resources that drive evolution of machine translation systems, a work in progress.

The empirical evidence presented shows that linguistic knowledge concerning the paraphrasal characteristics of support verb constructions offers a most promising avenue for research besides being a practical aid to translators in the pre-editing of machine translation input in a controlled language environment. It also brings with it an

opportunity to enhance electronic dictionaries by adding new attributes to the lexicon, and, eventually it can enable precision in the area where we consider we can best improve machine translation quality. Even though it was not our goal to discuss the benefits of support verb construction paraphrases to spoken language machine translation in this dissertation, we would like to point out that it is an area where the results of this study could effectively be implemented.

Bibliography

Bibliographic references were automatically processed with SUPeRB [Cabral et al., 2008].

[Albertoni et al. 2006]
R. Albertoni, E. Camossi, M. De Martino, F. Giannini & M. Monti. "Semantic Granularity for the Semantic Web". *Workshops 2006, LNCS* (2006), pp. 1863-1872. Springer. http://www.springerlink.com/content/e5737k872t008142/

[Allemang & Hendler 2008]
Dean Allemang, James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann. 2008-05-09. ISBN 9780123735560.

[Allerton 1989]
D.J. Allerton. "Three (or four) levels of word co-occurrence restrictions". *Lingua* **63** (1989), pp. 17-40.

[Antoniou & van Harmelen 2004]
Grigoris Antoniou, Frank van Harmelen, *A Semantic Web Primer*, The MIT Press, Cambridge, Massachusetts, April 2004, ISSN: 0-262-01210-3.

[Antoniou et al. 2005]
Grigoris Antoniou, Enrico Franconi & Frank van Harmelen. "Introduction to Semantic Web Ontology Languages". In *Reasoning Web 2005* 2005, pp. 1-21.

[Antoniou & Bikakis 2007]
Grigoris Antoniou & Antonis Bikakis. "DR-Prolog: A System for Defeasible Reasoning with Rules and Ontologies on the Semantic Web". In *Proceedings of the IEEE Trans. Knowl. Data Eng* 19 (2), 2007, pp. 233-245.

[Azizinezhad 2006]
Massoud Azizinezhad. "Is Translation Teachable?". *Translation Journal* **10.2** (2006). http://accurapid.com/journal/36edu.htm

[Bacelar do Nascimento et al. 1993]
Maria Fernanda Bacelar do Nascimento, Amália Mendes & Diana Santos. "O corpus e a classificação sintáctica dos verbos". In *Actas do 1º Encontro de Processamento da Língua Portuguesa (escrita e falada) (EPLP'93)* (Lisboa, 25-26 February 1993), pp. 125-129.

[Banerjee & Lavie 2005]
Satanjeev Banerjee & Alon Lavie. "METEOR: An Automatic Metric For MT Evaluation With Improved Correlation With Human Judgments". In *Proceedings of the ACL; Workshop On Intrinsic And Extrinsic Evaluation Measures For Machine Translation And/or Summarization* (Ann Arbor, Michigan, EUA, June 2005), Association for Computational Linguistics, pp. 65-72.
http://www.aclweb.org/anthology-new/W/W05/W05-0909.pdf

[Bannard & Callison-Burch 2005]
Colin Bannard & Chris Callison-Burch. "Paraphrasing with Bilingual Parallel Corpora". In *Proceedings of the 43rd Annual Meeting of the ACL* (Ann Arbor, Michigan, EUA, June 2005), Association for Computational Linguistics, pp. 597-604.
http://acl.ldc.upenn.edu/P/P05/P05-1074.pdf

[Baptista et al. 2004]

J. Baptista, A. Correia & G. Fernandes. "Frozen Sentences of Portuguese: Formal Descriptions for NLP". In *Workshop on Multiword Expressions: Integrating Processing, International Conference of the European Chapter of the Association for Computational Linguistics* (Barcelona, Spain, 26 July 2004).

[Baptista 2004]

J. Baptista. "Compositional vs Frozen sequences". In *Lexicon-Grammar Workshop* (Beijing, China, 14-19 October 2004).

[Baptista 2005]

Jorge Baptista. *Sintaxe dos Nomes Predicativos com verbo-suporte SER DE*. Lisboa: Fundação para a Ciência e a Tecnologia/Fundação Calouste Gulbenkian. 2005.

[Barnstone 1993]

Willis Barnstone. *The Poetics of Translation: History, Theory, Practice*. New Haven and London: Yale University Press. 1993. ISBN: 0-300-05189-1.

[Barreiro et al. 1996]

Anabela Barreiro, Luzia Helena Wittmann & Maria de Jesus Pereira. "Lexical differences between European and Brazilian Portuguese". *INESC Journal of Research and Development* **5.2** (1996).
http://www.linguateca.pt/Repositorio/Barreiroetal95.rtf

[Barreiro & Ranchhod 2005]

Anabela Barreiro & Elisabete Ranchhod. "Machine Translation Challenges for Portuguese". *Linguisticæ Investigationes* **28.1** (2005), pp. 3-18. (Machine Translation, Controlled Languages and Specialised Languages). Amsterdam/Philadelphia: John Benjamins Publishing Company. ISSN: 0378-4169.

[Barreiro 2008b]

Anabela Barreiro. "Port4NooJ: Portuguese Linguistic Module and Bilingual Resources for Machine Translation". In *Proceedings of the 2007 International NooJ Conference* (Barcelona, Spain, 7-9 June 2007), Cambridge Scholars Publishing.

[Barreiro 2008a]

Anabela Barreiro. "ParaMT: a Paraphraser for Machine Translation". In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira & Paulo Quaresma (eds.), *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)* Vol. 5190, (Aveiro, Portugal, 8-10 September 2008), Springer Verlag, pp. 202-211. Slides

[Barreiro 2008c]

Anabela Barreiro "Novas Ferramentas e Recursos Linguísticos para a Tradução Automática: Por ocasião d'O Fim do Início de uma Nova Era no Processamento da Língua Portuguesa". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 13-23.
http://www.linguateca.pt/Linguateca10anos/ResumosAlargados/BarreiroL10.pdf

[Barzilay & McKeown 2001]

Regina Barzilay & Kathleen McKeown. "Extracting paraphrases from a parallel corpus". In *Proceedings of the ACL/EACL* (Toulouse, 2001), pp. 50-57.

[Barzilay 2001]

Regina Barzilay. Multidocument summarization by information fusion. PhD. Department of Computer Science; Columbia University. 2001.

[Barzilay & Lee 2003]

Regina Barzilay & Lillian Lee. "Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment". In *Proceedings of Human Language Technology conference (HLT- NAACL 2003)* 2003, pp. 16-23. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.4603

[Barzilay 2003]

R. Barzilay. Information Fusion for Multidocument Summarization. PhD Thesis. Columbia University. 2003.

[Bar-Hillel 1959]

Y. Bar-Hillel. "Report on the state of machine translation in the United States and Great Britain". Jerusalem Hebrew University. 15 February 1959. Technical report.

[Beck 1997]

David Beck. "Rheme, Theme, and communicative structure in Lushootseed and Bella Coola". *Recent Trends in Meaning-Text Theory* (1997), pp. 93-135. Amsterdam: Benjamins. http://www.ualberta.ca/~dbeck/CommS.pdf

[Berners-Lee et al. 2001]

Tim Berners-Lee, James Hendler & Ora Lassila. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In *Scientific American Magazine*, May 17, 2001. http://www.sciam.com/article.cfm?id=the-semantic-web&print=true

[Bhagat & Ravichandran 2008]

Rahul Bhagat & Deepak Ravichandran. "Large Scale Acquisition of Paraphrases for Learning Surface Patterns". In *Proceedings of Association for Computational Linguistics (ACL)* (Columbus, OH, 2008). http://www.isi.edu/~rahul/Papers/Paraphrases_ACL2008.pdf

[Biber et al. 1999a]

Douglas Biber, Susan Conrad & Randi Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press. 1999.

[Biguenet & Schulte 1989]

John Biguenet & Rainer Schulte. *The Craft of Translation*. Chicago Guides to Writing, Editing, and Publishing. 1989. ISBN: 9780226048697. http://www.press.uchicago.edu/presssite/metadata.epl?mode=synopsis&bookkey=62846

[Boonthum 2004]

Chutima Boonthum. "iSTART: paraphrase recognition". In *Proceedings of the ACL 2004 workshop on Student research* (Barcelona, Spain, 2004). http://www.aclweb.org/anthology-new/P/P04/P04-2006.pdf

[Boyer & Lapalme 1985]

M. Boyer & G. Lapalme. "Generating paraphrases from meaning-text semantic networks". *Computational Intelligence* **1.1** (1985), pp. 103-117.

[Brandel 2008]

Brandel, Mary. "Stormy Weather." *Computerworld* 3 Nov. 2008: 22-28.

[Braz et al. 2005]

Rodrigo de Salvo Braz, Roxana Girju, Vasin Punyakanok, Dan Roth & Mark Sammons. "An Inference Model for Semantic Entailment in Natural Language". In *Proceedings of the PASCAL RTE Challenge* 29-32, pp. 29-32. http://l2r.cs.uiuc.edu/~danr/Papers/BGPRS05.pdf

[Butt 2003]

Miriam Butt. "The Light Verb Jungle". *Working Papers in Linguistics* **9** (2003), pp. 1-49. Papers from the GSAS/Dudley House
http://ling.uni-konstanz.de/pages/home/butt/harvard-work.pdf

[Buvet 2003]
Pierre-André Buvet. "La possessivation dans les constructions à support". *Lingvisticae investigationes. Actes du Colloque Grammaires et Lexiques Comparés.* **26.1** (2003), pp. 47-70. Benjamins. ISSN: 0378-4169.

[Cabral et al. 2008]
Luís Miguel Cabral, Diana Santos & Luís Costa. "SUPeRB: using an automated publication helper in 9012 at SINTEF ICT". (SINTEF, Oslo, 21 November 2008).
http://www.linguateca.pt/documentos/SUPeRB21Nov08.pdf

[Callison-Burch et al. 2006a]
Chris Callison-Burch, Philipp Koehn & Miles Osborne. "Improved Statistical Machine Translation Using Paraphrases". In *NAACL-2006* 2006.

[Callison-Burch et al. 2006b]
C. Callison-Burch, M. Osborne & P. Koehn. "Re-evaluating the Role of BLEU in Machine Translation Research". In *11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006* 2006, pp. 249-256.

[Callison-Burch 2007]
Chris Callison-Burch. Paraphrasing and Translation. PhD Thesis. University of Edinburgh. 2007.
http://www.cs.jhu.edu/~ccb/publications/callison-burch-thesis.pdf

[Callison-Burch 2008]
Chris Callison-Burch. "Syntactic Constraints on Paraphrases Extracted from Parallel Corpora". In *Proceedings of EMNLP 2008* 2008.
http://www.cs.jhu.edu/~ccb/publications/syntactic-constraints-on-paraphrases.pdf

[Carvalho 2007]
Paula Cristina Quaresma da Fonseca Carvalho. *Análise e Representação de Construções Adjectivais para Processamento Automático de Texto. Adjectivos Intransitivos Humanos*. PhD dissertation. Universidade de Lisboa. 2007.
http://www.linguateca.pt/Repositorio/TeseDoutPaulaCarvalho.pdf

[Casteleiro 1981]
Casteleiro, João Malaca. *Sintaxe transformacional do adjectivo – regência das construções completivas*, Lisboa: INIC. 1981.

[Catford 1965]
John Catford. *A Linguistic Theory of Translation: an Essay on Applied Linguistics*. London: Oxford University Press. 1965.

[Chacoto 2005]
Lucília Chacoto. O Verbo Fazer em Construções Nominais Predicativas. PhD dissertation. Universidade do Algarve, Portugal. 2005.

[Cherry & Quirk 2008]
Colin Cherry, Chris Quirk, Discriminative, Syntactic Language Modeling through Latent SVMs, in Proceeding of AMTA, Association for Machine Translation in the Americas, 23 Oct. 2008

[Chomsky 1957]
Noam Chomsky. *Syntactic Structures*. Mouton. 1957.

[Chomsky 1965]

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press. 1965. ISBN: 0-262-53007-4.

[Chomsky 1975]
Noam Chomsky. *The Logical Structure of Linguistic Theory*. 1975.

[Chomsky 1980]
Noam Chomsky. *Rules and Representations*. 1980.

[Chomsky 1999]
Noam Chomsky. *Derivation by Phase*. Ms., MIT. 1999.

[Cohen & Hunter 2006]
K Bretonnel Cohen & Lawrence Hunter. "A critical review of PASBio". *BMC Bioinformatics* (2006).
http://www.biomedcentral.com/content/pdf/1471-2105-7-S3-S5.pdf

[Cohen et al. 2008]
K. Bretonnel Cohen, Martha Palmer, Lawrence Hunter. Nominalization and Alternations in Biomedical Language. PLoS ONE 3(9): (2008)

[Collins et al. 2005]
Michael Collins, Philipp Koehn & Ivona Kučerová. "Clause restructuring for statistical machine translation". In *Proceedings of ACL* 2005, pp. 531-540.
http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/clause-acl05.pdf

[Crego & Habash]
Josep M. Crego & Nizar Habash. "Using Shallow Syntax Information to Improve Word Alignment and Reordering for SMT". In *Proceedings of the Third Workshop on Statistical Machine Translation, ACL-08: HLT*. June 19, 2008. The Ohio State University. Columbus, Ohio, USA. pp. 53-61.
http://www.mt-archive.info/ACL-SMT-2008-Crego.pdf

[Cross 1992]
Marilyn Cross. "Choice in lexis: computer generation of lexis as most delicate grammar". *Language Sciences* **14.4** (1992), pp. 579-607.

[Cruse 1986]
D. Alan Cruse. *Lexical Semantics*. Cambridge University Press. 1986.

[Culioli 1990]
A. Culioli. *Pour une Linguistique de l'énonciation*. Paris: Ophrys. Opérations et représentations. 1990.

[Davidson 1978]
Donald Davidson. "What Metaphors Mean". *Critical Inquiry* **5** (1978), pp. 31-48.

[Davies et al. 2006]
John Davies, Rudi Studer & Paul Warren. *Semantic Web Technologies: Trends and Research in Ontology-based Systems*. Wiley. 2006. ISBN: 0470025964.

[Denoul & Lepage 2005]
E. Denoul & Y. Lepage. "BLEU in characters: towards automatic MT evaluation in languages without word delimiters". In *Companion Volume to the Proceedings of the Second International Joint Conference on Natural Language Processing* 2005, pp. 81-86.

[Dillinger 2008]
M. Dillinger. *Hands-on Research Methods*. San Jose State University: Unpublished manuscript. 2008.

[Dolan et al. 2004]

William B. Dolan, Chris Quirk & Chris Brockett. "Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources". In *COLING 2004* (Geneva, Switzerland, 2004).

[Dras 1995]

Mark Dras. "Automatic Identification of Support Verbs: A Step Towards a Definition of Semantic Weight". In *Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence* (Canberra, Australia, 1995), World Scientific Press, pp. 451-458. http://xxx.lanl.gov/abs/cmp-lg/9510007

[Duboué & Chu-Carroll 2006]

Pablo Ariel Duboué & Jennifer Chu-Carroll. "Answering the question you wish they had asked: The impact of paraphrasing for Question Answering". In *Proceedings of HLT-NAACL 2006* 2006. http://acl.ldc.upenn.edu/N/N06/N06-2009.pdf

[Dugast et al. 2007]

Loïc Dugast, Jean Senellart, & Philipp Koehn: Statistical post-editing on SYSTRAN's rule-based translation system. In Proceedings of the Second Workshop on Statistical Machine Translation, ACL 2007, June 23, 2007, Prague, Czech Republic; pp. 220-223 http://www.mt-archive.info/ACL-SMT-2007-Dugast.pdf

[Dyer et al. 2008]

Christopher Dyer, Smaranda Muresan & Philip Resnik. "Generalizing Word Lattice Translation". In *in Proceedings of ACL-08: HLT* (Columbus, Ohio, USA, June 2008), Association for Computational Linguistics, pp. 1012-1020. http://www.aclweb.org/anthology-new/P/P08/P08-1115.pdf

[Edmonds 1999]

Philip Edmonds. Semantic Representations of Near-Synonyms for Automatic Lexical Choice. PhD dissertation. University of Toronto, Canada (1999). http://ftp.cs.toronto.edu/pub/gh/Edmonds-PhDthesis.pdf

[Edmonds & Hirst 2002]

P. Edmonds & G. Hirst. "Near-Synonymy and Lexical Choice". *Computational Linguistics* **28.2** (2002), pp. 105-144. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.5469

[Eisele et al. 2008]

Andreas Eisele, Christian Federmann, Hervé Saint-Amand, Michael Jellinghaus, Teresa Herrmann, Yu Chen. "Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System" In *Proceedings of the Third Workshop on Statistical Machine Translation, ACL-08: HLT*. June 19, 2008. The Ohio State University. Columbus, Ohio, USA. pp. 179-182. http://www.mt-archive.info/ACL-SMT-2008-Eisele.pdf

[Evert & Kermes 2003]

Stefan Evert & Hannah Kermes. "Experiments on Candidate Data for Collocation Extraction". In *Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics* 2003, pp. 83-86.

[Fellbaum 1998]

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Cambridge: The MIT Press. 1998.

[Fernandes & Baptista 2007]

Graça Fernandes & Jorge Baptista. "Frozen sentences on large corpus: an experience". In *26th International Colloquium on Compared Lexicon and Grammar* (Bonifacio (Corse du Sud), 2-6 October 2007).

[Fernandes 2007]

G. Fernandes. Léxico-Gramática das Frases Fixas do Português Europeu. Contruções Intransitivas. Master's dissertation. Universidade do Algarve/FCHS. Faro, Portugal. 2007.

[Fillmore et al. 2002]

Charles J. Fillmore, Collin F. Baker & Hiroaki Sato. "The FrameNet database and software tools". In *Proceedings of the Third International Conference on Language Resources and Evaluation*, (Las Palmas, Canary Islands, Spain, 2002) 1157-1160.

[Fillmore et al. 2003]

Fillmore, C. J., Johnson, C. R. & Petruck, M. R. L. "Background to Framenet". *International Journal of Lexicography* **16.3** (2003), pp. 235-250.

[Firth 1957]

J.R. Firth. "A Synopsis of Linguistic Theory, 1930-1955. Studies in Linguistic Analysis". *Special Volume, Philological Society* (1957), pp. 1-32.

[Firth 1968]

J.R. Firth. "Linguistic Analysis and Translation". In F.R. Palmer (ed.). 1968, pp. 74-84.

[Frankenberg-Garcia & Santos 2003]

Ana Frankenberg-Garcia & Diana Santos "Introducing COMPARA, the Portuguese-English parallel translation corpus". In Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds.), *Corpora in Translation Education*. Manchester:St. Jerome Publishing, 2003, pp. 71-87.
http://www.linguateca.pt/documentos/Frankenberg-GarciaSantos2000.pdf

[Fuchs 1982]

Catherine Fuchs. *La Paraphrase*. Presses Universitaires de France. 1982. ISBN: 9782130371038.

[Fuchs 1985]

Catherine Fuchs. *Aspects de l'ambiguïté et de la paraphrase dans les langues naturelles*. Berne. Sciences pour la Communication. 1985.

[Fuchs 1987]

Catherine Fuchs. *L'Ambiguïté et la Paraphrase: opérations linguistiques, processus cognitifs et traitements automatisés*. Caen, Centre de Publications de l'Université. 1987.

[Fuchs 1994]

Catherine Fuchs. *Paraphrase et énonciation*. Paris, France: Ophrys. Coll. L'Homme dans la Langue. 1994.

[Fujita et al. 2004]

Atsushi Fujita, Kentaro Furihata, Kentaro Inui, Yuji Matsumoto & Koichi Takeuchi. "Paraphrasing of Japanese Light-verb Constructions Based on Lexical Conceptual Structure". In *in Proceedings of the ACL Workshop on Multiword Expressions* (Barcelona, Spain, 2004).

[Fung 1995]

Pascale Fung. "Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus". In *Proceedings of the Third Annual Workshop on Very Large Corpora* 1995, pp. 173-183. http://www.ee.ust.hk/~pascale/wvlc95.ps

[Giry-Schneider 1978]
J. Giry-Schneider. *Les nominalisations en français: l'opérateur "faire" dans le lexique*. Geneve, Switezerland: Droz. 1978.

[Giry-Schneider 1987]
Jacqueline Giry-Schneider. *Les prédicats nominaux en français. Les phrases à verbe support*. Droz. 1987.

[Glass & Hazen 1998]
Glass, James R. & Timothy J. Hazen. "Telephone based conversational speech recognition in the Jupiter domain". In *Proceedings of the International Conference on Spoken Language Processing* (Sydney, Australia, 1998).

[Green et al. 2004]
Rebecca Green, Bonnie J. Dorr & Philip Resnik. "Inducing Frame Semantic Verb Classes from WordNet and LDOCE". In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics* 2004.

[Grimshaw & Mester 1988]
Jane Grimshaw & Armin Mester. "Light Verbs and Theta-marking". *Linguistic Inquiry* **19** (1988), pp. 205-232.

[Gross 1975]
Maurice Gross. *Méthodes en Syntaxe - Régime des constructions complétives*. Paris: Hermann. 1975.

[Gross 1981]
Maurice Gross. "Les bases empiriques de la notion de prédicat sémantique". *Formes Syntaxiques et Prédicat Sémantiques, Langages* **63** (1981), pp. 7-52. Paris, França: Larousse.

[Gross 1982a]
Maurice Gross. "Simple sentences. Discussion of Fred W. Householder". *Text Processing* (1982), pp. 297-315. Almqvist Wiksell.

[Gross 1986]
Maurice Gross: Lexicon Grammar. The Representation of Compound Words. COLING 1986: 1-6

[Gross 1996]
Maurice Gross. "Les formes être Prép X du français". *Lingvisticæ Investigationes* **20.2** (1996).

[Gruber 1993]
Helmut Gruber. "Political language and textual vagueness". *Pragmatics* **3.1** (1993), pp. 1-28.

[Guttenplan 2005]
Samuel Guttenplan. *Objects of Metaphor*. Oxford University Press. 2005. ISBN: 0199280894. http://ndpr.nd.edu/review.cfm?id=7504

[Haas 1999]
Haas, S. W. "Knowledge Representation, Concepts, and Terminology: Toward a Metadata Registry for the Bureau of Labor Statistics". 1999. Final report for the Bureau of Labor Statistics contract, "Investigation into the Requirements and Structure of a Knowledge Organization for BLS Published Information" http://ils.unc.edu/~stephani/bls/fin-rept-99.pdf

[Haas 2000]

Haas, S. W. "A Terminology Crosswalk for LABSTAT: Mapping General Language Words and Phrases to BLS Terms". 2000. Final report for the Bureau of Labor Statistics contract http://ils.unc.edu/~stephani//bls/fin-rept-00.pdf

[Haas & Hert 2000]
Haas, S. W. & Hert, C. A. "Terminology development and organization in multi-community environments: The case of statistical information". In *Proceedings of the 11th American Society for Information Science & Technology SIG/CR Classification Workshop* 2000, pp. 51-72.

[Hale & Keyser 1993]
Ken Hale & Samuel J. Keyser. "On the argument structure and the lexical expression of syntactic relations", in: Hale, K. & S. Keyser (eds.) The view from Building 20. Cambridge, MA: The MIT Press. 1993, pp. 53-109.

[Halliday 1985a]
M.A.K. Halliday. *An introduction to functional grammar*. London; Baltimore, MD, USA: Edward Arnold. 1985.

[Halliday 1985b]
M.A.K. Halliday. *Spoken and written language*. Victoria: Deakin University. 1985.

[Halliday 1985c]
M.A.K. Halliday "Systemic background". In James D. William S. & Greaves Benson (eds.), *Systemic perspectives on discourse*. Norwood, N.J.: Ablex, 1985.

[Harris 1951]
Z. Harris. *Methods in Structural linguistics*. Chicago: University of Chicago Press. 1951.

[Harris 1957]
Zellig Harris. "Co-occurrence and transformation in linguistic structure". *Language* **33** (1957), pp. 293-340.

[Harris 1964]
Zellig Harris. "Transformations in Linguistic Structure". *Proceedings of the American Philosophical Society* (1964), pp. 418-422. Repr. in 1970a.472-481

[Harris 1968]
Zellig Harris. *Mathematical Structures of Language*. New York: Wiley. 1968.

[Hasan 1987]
Ruqaiya Hasan "The grammarian's dream: lexis as most delicate grammar". In Michael Halliday & Robin Fawcett (eds.), *New developments in systemic linguistics: theory and description*. London: Pinter, 1987.

[Hermans 1985]
Theo Hermans. "Images of Translation: Metaphor and Imagery in the Renaissance Discourse on Translation". *The Manipulation of Literature: Studies in Literary Translation* (1985), pp. 103-35. London/Sydney: Croom Helm.

[Heyer et al. 2001]
Gerhard Heyer, Martin Läuter, Uwe Quasthoff, Wittig, Th. & Christian Wolff. "Learning Relations using Collocations". In *Proceedings of the IJCAI Workshop on Ontology Learning*. Seattle, WA, August 2001, pp. 19-24.

[Hirao et al. 2004]
T. Hirao, T. Fukusima, M. Okumura, C. Nobata & H. Nanba. "Corpus and Evaluation Measures for Multiple Document Summarization with Multiple Sources". In *Proceedings of the COLING-2004* 2004, pp. 535-541.

[Hoey 2005]
Michael Hoey. *Lexical Priming: A New Theory of Words and Language*. London: Routledge. 2005.

[Hutchins & Somers 1992]
W. J. Hutchins & H. L. Somers. *An Introduction to Machine Translation*. London: Academic Press. 1992.

[Ibrahim et al. 2003]
Ali Ibrahim, Boris Katz & Jimmy Lin. "Extracting structural paraphrases from aligned monolingual corpora". In *Proceedings of the Second International Workshop on Paraphrasing (ACL 2003)* 2003.

[Imamura et al. 2004]
Kenji Imamura; Hideo Okuma; Taro Watanabe; Eiichiro Sumita. "Example-based Machine Translation Based on Syntactic Transfer with Statistical Models". In Proceedings of the 20th International Conference On Computational Linguistics. Geneva, Switzerland, No. 99, 2004.
http://people.csail.mit.edu/koehn/smt/imamura2004.pdf

[Iordanskaja & Polguère 1988]
L. Iordanskaja & A. Polguère. "Semantic Processing for Text Generation". In *Proceedings of the 1st International Computer Science Conference* (Hong Kong, 19-21 December 1988), pp. 310-318.

[Iordanskaja et al. 1991]
Lidija Iordanskaja, Richard Kittredge & Alain Polguère "Lexical Selection and Paraphrase in a Meaning-Text Generation Model". In Cécile L. Paris, William R. Swartout & William C. Mann (eds.), *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. Kluwer Academic Publishers, 1991, pp. 293-312.

[Iordanskaja et al. 1996]
L. Iordanskaja, M. Kim & A. Polguère. "Some Procedural Problems in the Implementation of Lexical Functions for Text Generation". *Lexical Functions in Lexicography and Natural Language Processing* (1996), pp. 279-297.

[Kasparek 1983]
Christopher Kasparek. "The Translator's Endless Toil". *The Polish Review* **XXVIII.2** (1983), pp. 83-87.

[Kearns 2002]
Kate Kearns. "Light verbs in English". 2002.
http://www.ling.canterbury.ac.nz/documents/LIGHT%20VERBS%20IN%20ENGLISH%20KSK.pdf

[Kerzazi-Lasri 2003]
Rafika Kerzazi-Lasri. "La métaphore dans le commentaire politique [Texte imprimé]: articles extraits de "L'Express" et du "Point"". *Langage* (2003), pp. 173-176. préf. de Frédéric François

[Kilgarriff & Tugwell 2001]
Adam Kilgarriff & David Tugwell. "WASP-Bench: an MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation". In Proceedings of the MT Summit VIII, Santiago de Compostela, Spain. 2001. pp. 187-190.

[Kingsbury et al. 2002]

P. Kingsbury, M. Palmer & M. Marcus. "Adding semantic annotation to the Penn treebank". In *Proceedings of the Human Language Technology Conference (HLT'02)* 2002.

[Koehn & Knight 2002]
Philipp Koehn & Kevin Knight. "Learning a translation lexicon from monolingual corpora". In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition* July 2002, Philadelphia: Association for Computational Linguistics, pp. 9-16. ACL Special Interest Group on the Lexicon (SIGLEX)
http://www.aclweb.org/anthology-new/W/W02/W02-0902.pdf

[Lakoff & Johnson 1980]
George Lakoff & Mark Johnson. *Metaphors We Live By*. Chicago & Londres: University of Chicago Press. 1980.

[Lakoff & Turner 1989]
George Lakoff & Mark Turner. *More Than Cool Reason: A Field Guide to Poetic Metaphor*. Chicago: University of Chicago Press. 1989. ISBN: 0226468127.

[Lakoff 1992]
George Lakoff "Metaphors and war: The metaphor system used to justify was in the gulf". In Martin Pütz (ed.), *Thirty Years of Linguistic Evolution: Studies in Honor of Rene Dirven on the Occasion of his Sixtieth Birthday.* Amsterdam: John Benjamins, 1992, pp. 463-481.

[Lakoff 1996]
George Lakoff "Reflections on metaphor and grammar". In Masayoshi Shibatani & Sandra A. Thompson (eds.), *Essays in Semantics and Pragmatics*. Amsterdam: John Benjamins Publishing Company, 1996, pp. 133-144.

[Landers 2001]
Clifford E. Landers. *Literary Translation: A Practical Guide*. Multilingual Matters Ltd. Topics in Translation. 2001. ISBN: 1-85359-519-5.

[Lareau 2002]
F. Lareau. La synthèse automatique de paraphrases comme outil de vérification des dictionnaires et grammaires de type Sens-Texte. Mémoire de maîtrise. Département de linguistique et de traduction, Université de Montréal. 2002.

[Larson 1988]
Richard Larson. On the double object construction. *Linguistic Inquiry* 19, 1988, pp. 335-391.

[Lee & Przybocki 2005]
A. Lee & M. Przybocki. "NIST 2005 machine translation evaluation official results". 2005.

[Leighton 1990]
Lauren G. Leighton. "Translation as a Derived Art". In *Proceedings of the American Philosophical Society* 134 (4), December 1990, American Philosophical Society, pp. 445-454. http://www.jstor.org/pss/986898

[Lin & Pantel 2001]
Dekang Lin & Patrick Pantel. "DIRT-discovery of inference rules from text". In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 2001.

[Lyons 1995]

John Lyons. *Linguistic semantics: An introduction*. Cambridge, England: Cambridge University Press. 1995.

[Macleod et al. 2000]

Catherine Macleod, Ralph Grishman & Adam Meyers. "An Electronic Lexicon of Nominalizations: NOMLEX". In *Conference Abstracts of the 12th Joint International Conference of Association for Literary and Linguistic Computing and Association for Computers and the Humanities (ALLC/ACH 2000)* (Glasgow, 2000).

[Madnani et al. 2007]

Nitin Madnani, Necip Fazil Ayan, Philip Resnik & Bonnie Dorr. "Using Paraphrases for Parameter Tuning in Statistical Machine Translation". In *ACL Workshop on Statistical Machine Translation* (Prague, 2007).

[Maia & Barreiro 2007]

Belinda Maia & Anabela Barreiro "Uma experiência de recolha de exemplos classificados de tradução automática de inglês para português". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal: IST Press, 2007, pp. 205-216.

[Maia et al. 2007]

Belinda Maia, Rui Sousa Silva, Anabela Barreiro & Cecília Fróis. "N-grams in search of theories". In *Proceedings of the 6th International Conference of Practical Applications in Language and Computers (PALC 2007)* (Lodz University, Lodz, Poland, 19-22 April 2007).

[Maia 2007]

Belinda Maia. "A Tradução Automática - amiga ou inimiga do tradutor?". In *X Seminário de Tradução Científica e Técnica em língua Portuguesa: Tradução e Multilinguismo* 2007, Direção de Terminologia e Indústrias da Língua - DTIL; União Latina. http://dtil.unilat.org/Xseminariofct_ul/belinda_maia.htm

[Maia 2008]

Belinda Maia. "A Tradução Automática - amiga ou inimiga do tradutor?". In *X Seminário de Tradução Científica e Técnica em língua Portuguesa: Tradução e Multilinguismo* 2007, Direção de Terminologia e Indústrias da Língua - DTIL; União Latina. http://dtil.unilat.org/Xseminariofct_ul/belinda_maia.htm

[Maia & Matos 2008]

Belinda Maia & Sérgio Matos. "Corpógrafo V4 - Tools for Researchers and Teachers using Comparable Corpora". In Pierre Zweigenbaum, Éric Gaussier & Pascale Fung (eds.), *LREC 2008 Workshop on Comparable Corpora (LREC 2008)* (Marrakech, 31 May 2008), European Language Resources Association (ELRA), pp. 79-82. http://www.linguateca.pt/documentos/MaiaMatosW12LREC08.pdf

[Marcus et al. 1993]

Mitchell P. Marcus, Beatrice Santorini & Mary Ann Marcinkiewicz. "Building a Large Annotated Corpus of English: The Penn Treebank". In *Computational Linguistics*. Special Issue on Using Large Corpora, **19:2** (June 1993), pp. 313-330.

[Martin 1976]

R. Martin. *Inférence, antonymie, et paraphrase. Éléments pour une théorie sémantique*. Paris: Klincksieck. 1976.

[Martin 1983]

Robert Martin. *Pour une logique du sens*. Paris: P.U.F. 1983.

[McCawley 1976]

McCawley, James D. *Syntax and semantics, 7: Notes from the linguistic underground*. New York: Academic Press. 1976.

[McKeown et al. 2002]

Kathleen McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith Klavans, Ani Nenkova, Carl Sable, Barry Schiffman & Sergey Sigelman. "Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster". In *Proceedings of the Human Language Technology Conference* (San Diego, CA, USA, March 2002).

[Melamed 2004]

Dan Melamed. Algorithms for Syntax-Aware Statistical Machine Translation. In *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation* (TMI'04), Baltimore, MD. 2004.

[Mel'čuk 1988]

Igor Mel'čuk. *Dependency Syntax: Theory and Practice*. State University of New York Press. 1988.

[Mel'čuk 1996]

Igor A. Mel'čuk "Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon". In L. Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/Philadelphia: Benjamins, 1996, pp. 37-102.

[Mel'čuk 2003]

Igor Mel'čuk. "Collocations dans le dictionnaire". In T. Szende (ed.), *Les écarts culturels dans les dictionnaires bilingues*. Paris, France: H. Champion, 2003, pp. 19-64.

[Menezes & Quirk 2008]

Menezes, Arul, Quirk, Chris. Syntactic Models for Structural Word Insertion and Deletion during Translation, in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Honolulu, Hawaii, Oct. 2008, pp. 735–744.

[Meyers et al. 2004a]

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young & Ralph Grishman. "Annotating noun argument structure for NomBank". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)* (Lisboa, Portugal, May 26-28, 2004).

[Meyers et al. 2004b]

A. Meyers, R. Reeves & C. Macleod. "NP-External Arguments: A Study of Argument Sharing in English". In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing* (Barcelona, Spain, 26 July 2004), pp. 96-103.

[Mieder 1993]

Wolfgang Mieder. *International Proverb Scholarship: An Annotated Bibliography, with supplements*. New York: Garland Publishing. 1993.

[Mieder 2001]

Wolfgang Mieder. *International Proverb Scholarship: An Annotated Bibliography*. Bern, New York: Peter Lang. Supplement III (1990-2000). 2001.

[Mieder 2004]

Wolfgang Mieder. *Proverbs: A Handbook*. Greenwood Press. Greenwood Folklore Handbooks. 2004.

[Milićević 2007a]
Jasmina Milićević. *La paraphrase. Modélisation de la paraphrase langagière*. Bern: Peter Lang. 2007. ISBN: 978-3-03911-197-8.

[Milićević 2007b]
Jasmina Milićević. *Semantic Equivalence Rules in Meaning-Text Paraphrasing*. Amsterdam and Philadelphia: Benjamins. 2007. ISBN: 978-90-272-3094-2. In Honour of Igor Mel'čuk

[Miller et al. 1990]
Miller, George A. Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. "Introduction to WordNet: an on-line lexical database". *International Journal of Lexicography* **3.4** (1990). Revised August 1993
ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

[Miller 1995]
George A. Miller. "WordNet: A Lexical Database for English". In *Communications of the ACM* 1995.
http://l2r.cs.uiuc.edu/~danr/Teaching/CS598-05/Papers/miller95.pdf

[Miller et al. 1995]
George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine Miller. "WordNet: an On-line Lexical Database". *International Journal of Lexicography*. http://www.cfilt.iitb.ac.in/archives/english_wordnet_5papers.pdf

[Mota 2008]
Cristina Mota. *How to keep up with language dynamics? A case study on Named Entity Recognition*. PhD dissertation. IST, Lisboa, Portugal. December 2008.

[Nadeau & Sekine 2007]
Nadeau, David and Satoshi Sekine. "A survey of named entity recognition and classification". *Named Entities: Recognition, classification and use: Special issue of Lingvisticæ Investigationes* 30:1 (2007), Sekine, Satoshi and Elisabete Ranchhod (eds.), pp. 3–26.

[Nagao 1984]
Makoto Nagao "A framework of a mechanical translation between Japanese and English by analogy principle". In A. Elithorn & R. Banerji (eds.), *Artificial and Human Intelligence*. Elsevier Science Publishers B. V. 1984.

[Nasr 1996]
Alexis Nasr. Un modèle de reformulation automatique fondé sur la théorie Sens-Texte: applications aux langages contrôlés. Thèse de doctorat en informatique. Université Paris 7. Paris. 1996.

[Nazar et al. 2008]
R. Nazar, L. Wanner & J. Vivald. "Two Step Flow in Bilingual Lexicon Extraction from Unrelated Corpora". In *Proceedings of the EAMT 2008 Conference (European Association for Machine Translation)* (Hamburg, Germany, 22-23 September 2008). http://www.mt-archive.info/EAMT-2008-Nazar.pdf

[Newmark 1988]
P. Newmark. *A Textbook of Translation*. Hertfordshire: Prentice Hall. 1988.

[Nida 1969]
Eugene A. Nida "Science of Translation". In A. S. Diz (ed.), *Language Structure and Translation: essays by Eugene A. Nida*. Stanford, CA, USA: Stanford University Press, 1969.

[Nord 1997]
Christiane Nord. "A Functional Typology of Translation". *Benjamins Translation Library* **26** (1997), pp. 43-66.
http://www.benjamins.com/cgi-bin/t_bookview.cgi?bookid=BTL%2026

[Orliac & Dillinger 2003]
Brigitte Orliac & Mike Dillinger. "Collocation extraction for machine translation". In *MT Summit IX,* New Orleans, USA, 23-27 September 2003; pp. 292-298.

[Orliac 2004]
Brigitte Orliac. *Automatisation du repérage et de l'encodage des collocations en langue de spécialité*. Doctoral dissertation presented at the University of Montreal, Montreal, Canada. 2004

[O'Grady et al. 1996]
W. O'Grady, M. Dobrovolsky & F. Katamba. *Contemporary Linguistics: An Introduction*. Longman. 1996.

[Palmer et al. 2005]
Palmer M, Kingsbury P, Gildea D. "The Proposition Bank: An Annotated Corpus of Semantic Roles". *Computational Linguistics* **31** (1), 2005, pp: 71–106.

[Pang et al. 2003]
Bo Pang, Kevin Knight & Daniel Marcu. "Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences". In Marti A. Hearst and Mari Ostendorf (ed.). In *Proceedings of the Human Language Technology Conference (HLT-NAACL'03).*Edmonton, Canada, 27 May - 1 June 2003, pp. 181-188.
http://www.isi.edu/~marcu/papers/mta-hlt-naacl03.pdf

[Papineni et al. 2002]
K. Papineni, S. Roukos, T. Ward & W. J. Zhu "BLEU: a method for automatic evaluation of machine translation". In Proceedings of the *40th Annual meeting of the Association for Computational Linguistics (ACL-2002)* 2002, pp. 311-318.

[Paşca 2003]
Marius Paşca. "Open-Domain Question Answering from Large Text Collections". *Computational Linguistics* **29.4** (2003), pp. 665-667.

[Paşca & Dienes 2005]
Marius Paşca & Péter Dienes. "Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web". In *Natural Language Processing - IJCNLP 2005* 2005, Berlin, Germany: Springer, pp. 119-130.

[Paşca 2005]
Marius Paşca. "Mining paraphrases from self-anchored web sentence fragments". In *Proceedings of the 9th European conference on Principles and Practice of Knowledge Discovery in Databases* (Porto, Portugal, 2005).

[Poibeau 2004]
Thierry Poibeau. "Automatic extraction of paraphrastic phrases from medium size corpora". In *Computational Linguistics Conference (COLING 2004)* (Genève, Switzerland, 2004).

[Polguère 2000]
A Polguère. "A "Natural" Lexicalization Model for Language Generation". In *Proceedings of the Fourth Symposium on Natural Language Processing 2000 (SNLP'2000)* (Chiangmai, Thailand, 10-12 May 2000), pp. 37-50.

[Quirk & Corston-Oliver 2006]

Chris Quirk, Simon Corston-Oliver, "The impact of parse quality on syntactically-informed statistical machine translation", *in Proceedings of EMNLP 2006*, ACL/SIGPARSE, Jul. 2006

[Ranchhod 1983]

Elisabete Ranchhod. "On the support verbs ser and estar in Portuguese". *Linguisticæ Investigationes* **VII** (1983), pp. 317-353. Amsterdam/Philadelphia: John Benjamins Publishing Company. ISSN: 0378-4169.

[Ranchhod 1990]

Elisabete Ranchhod (ed.). *Sintaxe dos predicados nominais com Estar*. Lisboa, Portugal: INIC. 1990.

[Ranchhod & Carvalho 2006]

Elisabete Marques Ranchhod & Paula Carvalho. "Expressões Multipalavra - Questões lexicais e sintácticas". *Material de ensino na Primeira Escola de Verão da Linguateca* (UP, Porto, Portugal, 10-14 July 2006). http://www.linguateca.pt/escolaverao2006/Sintaxe/EDV2006SintaxeRanchhodCarvalho.pdf

[Rapp 1995]

Reinhard Rapp. "Identifying word translation in non-parallel texts". In *Proceedings of the 33rd Meeting of the Association for Computational Linguistics* 1995, Cambridge, MA, USA, pp. 320-322. http://www.ims.uni-stuttgart.de/~prescher/clustering-papers/rapp:1995.ps

[Reiss & Vermeer 1991]

Katarina Reiss & Hans J. Vermeer. *Fundamentos para una teoría funcional de la traducción*. Madrid: Akal. 1991.

[Roth & Sammons 2007]

D. Roth & M. Sammons. "Semantic and Logical Inference Model for Textual Entailment". In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* 2007, pp. 107-112.

[Russo-Lassner et al. 2005]

Grazia Russo-Lassner, Jimmy Lin & Philip Resnik. "A Paraphrase-based Approach to Machine Translation Evaluation". LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57 College Park, MarylandUniversity of Maryland. 2005. http://www.umiacs.umd.edu/~jimmylin/publications/Russo-Lassner_etal_TR2005.pdf

[Sag et al. 2002]

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. "Multiword Expressions: A Pain in the Neck for NLP". In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)* 2002. http://lingo.stanford.edu/pubs/WP-2001-03.pdf

[Salkoff 1990]

Morris Salkoff. "Automatic translation of support verb constructions". In *13th International Conference on Computational Linguistics (COLING 1990)* (University of Helsinki, Finland, 20-25 August 1990), pp. 243-246.

[Salkoff 1999]

Morris Salkoff. *A French-English Grammar: A Contrastive Grammar On Translational Principles*. Amsterdam, Philadelphia: John Benjamins Publishing Company. 1999.

[Santorini 1990]

Santorini, B. "Part-of-speech tagging guidelines for the Penn Treebank Project". Technical report MS-CIS-90-47, 1990. Department of Computer and Information Science, University of Pennsylvania.

[Santos 1988]
Diana Maria de Sousa Marques Pinto dos Santos. A fase de transferência de um sistema de tradução automática do inglês para o português. Tese de Mestrado. Instituto Superior Técnico, Universidade Técnica de Lisboa. October 1988.

[Santos 1990]
Diana Santos. "Lexical gaps and idioms in Machine Translation". In Hans Karlgren (ed.), *Proceedings of the 14th International Conference on Computational Linguistics (COLING'90)* 2, (Helsínquia, 20-25 August 1990), pp. 330-335.
http://www.linguateca.pt/Diana/download/Santos1990COLING.pdf

[Santos 1992]
Diana Santos. "Broad-coverage machine translation". *INESC Journal of Research and Development* **3.1** (1992), pp. 43-59.

[Santos 1999]
Diana Santos. "Towards language-specific applications". *Machine Translation* **14.2** (1999), pp. 83-112. Dordrecht: Kluwer Academic Publishers. ISSN: 0922-6567.

[Santos et al. 2004]
Diana Santos, Belinda Maia & Luís Sarmento. "Gathering empirical data to evaluate MT from English to Portuguese". In Lambros Kranias, Nicoletta Calzolari, Gregor Thurmair, Yorick Wilks, Eduard Hovy, Guðrún Magnúsdóttir, Anna Samiotou & Khalid Choukri (eds.), *Proceedings of LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora* (Lisboa, Portugal, 25 May 2004), pp. 14-17.
http://www.linguateca.pt/documentos/SantosMaiaSarmentoAmazing2004.pdf

[Santos & Inácio 2006]
Diana Santos & Susana Inácio. "Annotating COMPARA, a grammar-aware parallel corpus". In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik & Daniel Tapias (eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)* (Genoa, Italy, 22-28 May 2006), pp. 1216-1221.
http://www.linguateca.pt/Diana/download/SantosInacioLREC2006.pdf

[Sarmento et al. 2004]
Luís Sarmento, Belinda Maia & Diana Santos. "The Corpógrafo - a Web-based environment for corpora research". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)* (Lisboa, Portugal, 26-28 May 2004), pp. 449-452.
http://www.linguateca.pt/Diana/download/SarmentoMaiaSantosLREC2004.pdf

[Sarmento 2006]
Luís Sarmento. "BACO - A large database of text and co-occurrences". In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik & Daniel Tapias (eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)* (Genoa, Italy, 22-28 May 2006), pp. 1787-1790. http://www.linguateca.pt/documentos/SarmentoLREC2006_BACO.pdf

[Sarmento et al. 2006]

Luís Sarmento, Belinda Maia, Diana Santos, Ana Pinto & Luís Cabral. "Corpógrafo V3: From Terminological Aid to Semi-automatic Knowledge Engine". In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik & Daniel Tapias (eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)* (Genoa, Italy, 22-28 May 2006), pp. 1502-1505. http://www.linguateca.pt/Diana/download/SarmentoetalLREC2006.pdf

[Sarmento 2007]

Luís Sarmento "Ferramentas para experimentação, recolha e avaliação de exemplos de tradução automática". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal: IST Press, 2007, pp. 193-203.

[Sarmento et al. 2007]

Luís Sarmento, Anabela Barreiro, Belinda Maia & Diana Santos "Avaliação de Tradução Automática: alguns conceitos e reflexões". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal: IST Press, 2007, pp. 181-190.

[Schulz et al. 2004]

Stefan Schulz, Kornél Markó, Eduardo Sbrissia, Percy Nohama & Udo Hahn. "Cognate Mapping - A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon". In *Proceedings of COLING 2004* (Geneva, Switzerland, 23-27 August 2004), pp. 813-819. http://www.aclweb.org/anthology/C04/C04-1117.bib

[Scott 1989]

B. E. Scott. "The Logos System". In *MT Summit II* (Munich, Germany, 1989).

[Scott 2003]

Bud Scott. "The Logos Model: An Historical Perspective". *Machine Translation* **18** (2003), pp. 1-72. http://www.springerlink.com/content/r01317149734884p/

[Sekine 2005]

Satoshi Sekine. "Automatic Paraphrase Discovery based on Context and Keywords between NE Pairs". In *International Workshop on Paraphrase* (Jeju Island, Korea, 2005). http://nlp.cs.nyu.edu/publication/papers/iwp05_final.pdf

[Sgall et al. 1986]

Petr Sgall, Eva Hajicova & Jarmila Panevova. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: Reidel - Prague: Academia. 1986.

[Smadja 1993]

Frank Smadja. "XTRACT: an Overview". In *Computer and the Humanities*. Vol.26, 1993.

[Shinyama et al. 2002]

Y. Shinyama, S. Sekine, K. Sudo & R. Grishman. "Automatic Paraphrase Acquisition from News Articles". In *Proceedings of Human Language Technology Conference* (San Diego, CA, USA, 2002).

[Shinyama & Sekine 2003]

Yusuke Shinyama & Satoshi Sekine. "Paraphrase Acquisition for Information Extraction". In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications* (Sapporo, Japan, 2003). http://nlp.cs.nyu.edu/publication/papers/shinyama-acl03.pdf

[Shinyama & Sekine 2005]

Yusuke Shinyama & Satoshi Sekine. "Using Repeated Patterns across Comparable Articles for Paraphrase Acquisition". New York University. Proteus Technical Report 2005. http://nlp.cs.nyu.edu/publication/papers/shinyama-none05.pdf

[Silberztein 2004]

Max Silberztein. "NooJ: A Cooperative, Object-Oriented Architecture for NLP". *INTEX pour la Linguistique et le traitement automatique des langues*. Presses Universitaires de Franche-Comté. Cahiers de la MSH Ledoux, 2004. Besançon, France.

[Sinclair 1996]

J. M. Sinclair. "The search for units of meaning". *Textus* **1.9** (1996), pp. 75-106.

[Tateisi et al. 2004]

Yuka Tateisi, Tomoko Ohta & Jun-ichi Tsujii. "Annotation of Predicate-argument Structure of Molecular Biology Text". In *Proceedings of the IJCNLP-04 workshop on Beyond Shallow Analyses* 2004.

[Temmerman 2001]

Rita Temmerman "Metaphors the live sciences live by". In *Translation and Meaning*. 2001, pp. 43-52. http://cvc.ehb.be/pub/temmerman_art_tm03.pdf

[Temmerman 2002]

Rita Temmerman. "Metaphorical Models and the Translation of Scientific Texts". *Linguistica Antverpiensia* **1** (2002), pp. 211-226.
http://cvc.ehb.be/pub/temmerman_art_metaph.pdf

[Teubert 2001]

Wolfgang Teubert. "Corpus Linguistics and Lexicography". *Text Corpora and Multilingual Lexicography: Special issue of International Journal of Corpus Linguistics* **iv** (2001), pp. 125-153.
http://www.benjamins.com/cgi-bin/t_articles.cgi?bookid=IJCL%206%3ASI&artid=202054669

[Ullmann 1962]

St. Ullmann. *Semantics. An Introduction to the Science of Meaning*. Oxford. 1962.

[van der Plas & Tiedemann 2006]

L. van der Plas & J. Tiedemann. "Finding synonyms using automatic word alignment and measures of distributional similarity". In *Proceedings of COLING/ ACL 2006* 2006.

[Venuti 2004]

Lawrence Venuti. *The Translation Studies Reader*. Routledge. 2004. 2nd edition

[Venuti 2008]

Lawrence Venuti. *The Translator's Invisibility: A History of Translation*. UK: Routledge (Taylor and Francis). 2008. ISBN: 9780415394536. Originally Published On: December 1994

[Wang et al. 2007]

C. Wang, M. Collins & P. Koehn. "Chinese syntactic reordering for statistical machine translation". In *Proceedings of EMNLP-CoNLL* 2007, pp. 737-745.

[Weaver 1955]

W. Weaver "Translation". In *Machine Translation of Languages*. Cambridge, MA, USA: MIT Press, 1955.

[Weaver 2002]

William Weaver. "The Art of Translation". *The Paris Review* **3, Issue 161** (2002). http://www.theparisreview.com/viewinterview.php/prmMID/421

[Wu & Wang 2005]
Hua Wu; Haifeng Wang. "Improving Statistical Word Alignment with a Rule-Based Machine Translation System". *In Proceedings of COLING-04*. August 23-27, 2004. pp. 29-35.

[Xia & McCord 2004]
F. Xia & M. McCord. "Improving a statistical MT system with automatically learned rewrite patterns". In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)* 2004.

[Yamamoto 2002]
Kazuhide Yamamoto. "Machine Translation by Interaction between Paraphraser and Transfer". In *Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002* 1, (Taipei, Taiwan, 24 August - 1 September 2002), Morristown, NJ, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology-new/C/C02/C02-1163.pdf

[Yamamoto 2004]
Kazuhide Yamamoto. "Interaction between paraphraser and transfer for spoken language translation". *Journal of Natural Language Processing* (2004). http://nlp.nagaokaut.ac.jp/arc/04/04JNLP.pdf

[Zhang 1998]
Qiao Zhang. "Fuzziness - vagueness - generality - ambiguity". *Journal of Pragmatics* **29** (1998), pp. 13-31.

[Zhou et al. 2006]
Liang Zhou, Chin-Yew Lin & Eduard Hovy. "Reevaluating machine translation results with paraphrase support". In *Proceedings of EMNLP* 2006. http://www.isi.edu/natural-language/people/hovy/papers/06EMNLP-MTeval-paraphrases.pdf

[Zhou et al. 2006b]
Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu & Eduard Hovy. "PARAEVAL: Using paraphrases to evaluate summaries automatically". In *Proceedings of HLT-NAACL* 2006. http://acl.ldc.upenn.edu/N/N06/N06-1057.pdf

[Zolkovskij & Mel'čuk 1965]
Aleksandr Zolkovskij & Igor Mel'čuk. "O vozmozhnom metode i instrumentax semanticheskogo sinteza [On a possible method an instruments for semantic synthesis (of texts)]". *Nauchno-texnicheskaja informacija [Scientific and Technological Information]* **6** (1965), pp. 23-28.

[ ]
"MSN Encarta Dictionary". http://encarta.msn.com/encnet/features/dictionary/dictionaryhome.aspx

[ ]
"Dictionary.com". http://dictionary.reference.com/

[ ]
"Merriam-Webster OnLine". http://www.merriam-webster.com/

[ ]
"Roget's Online Thesaurus". http://thesaurus.reference.com/

[ ]

"The Free Dictionary". http://www.thefreedictionary.com/