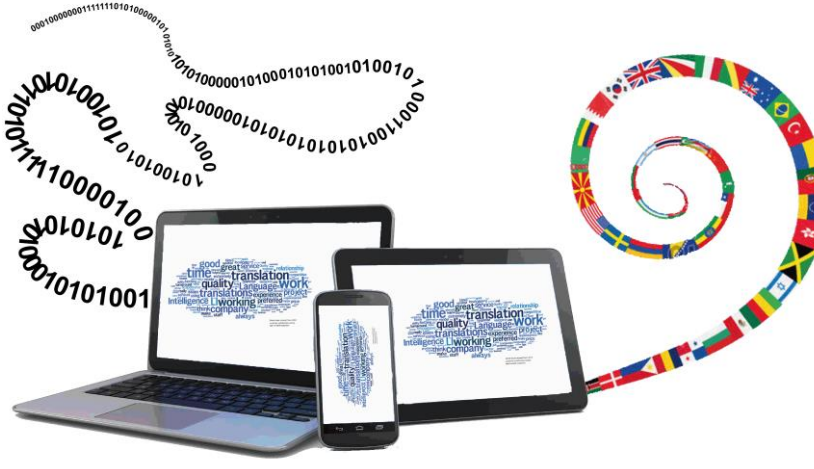
UNIVERSITÀ DEGLI STUDI DI SALERNO

MULTI-WORD UNIT PROCESSING IN MACHINE TRANSLATION

Developing and using language resources
for multi-word unit processing in Machine Translation

Johanna Monti



Supervisor
Prof. Annibale Elia

Coordinator
Prof. Alessandro Laudanna

XI Ciclo – Nuova Serie
2009-2012

Table of Contents

List of figures and tables

Figures

Tables

Acknowledgements

First of all, I would like to thank my Supervisor, Professor Annibale Elia, who has always been very helpful with his suggestions, comments and contributions. He has always supported me with his experience in Computational Linguistics, answered all my question and solved all my doubts.

I would also like to thank Professor Laudanna, coordinator of the Doctoral School of Communication Sciences, for his patience and his support, especially during the last few months of my doctoral research.

I am also very grateful to all my colleagues at the Maurice Gross Laboratory in the Department of Social, Political and Communication Sciences at the University of Salerno, and in particular Dr. Mario Monteleone, Dr. Federica Marano and Lorenza Melillo, with whom I have worked on a number of different research projects which were very successful from a theoretical and technical point of view.

I would also like to thank the co-authors of various papers presented in national and national conferences, Prof. Alberto Postiglione, Professor of Computer Science at the University of Salerno and Dr. Anabela Barreiro, researcher at INESC ID's Spoken Language Systems Laboratory (L^2F - Laboratório de sistemas de Língua Falada)- Portugal in addition to Prof. Elia, Mario and Federica.

I am particularly grateful to my Portuguese friend and colleague, Anabela, for her continuous support and help, the fruitful and interesting exchange of ideas on common research interests, her suggestions and the challenging discussions regarding possible strategies for implementing

new resources and tools for Machine Translation.

I am especially indebted to Professor Max Silberztein of the University of Franche-Comté for his precious suggestions which have enabled me to make significant progress in my research thanks to his help and his patience in answering all my emails concerning the use of NooJ, the NLP tool, which I have based all my research on and the strategies to adopt when processing various types of multi-word units.

I would also like to thank Prof. Ruslan Mitkov, Professor of Computational Linguistics and Language Engineering and Head of the Research Group in Computational Linguistics and Director of the Research Institute of Information and Language Processing at the University of Wolverhampton for the opportunity to be visiting lecturer at his Institute and work with him and his successful research group which is well-known for its innovative research in different areas of the field and its NLP tools and resources.

I would also like to mention Dr. Marco Turchi, researcher at the Fondazione Bruno Kessler in the Human Language Technology group, and thank him for his suggestions and interesting discussions related to my thesis.

Special thanks also go to Bud Scott, the father of the MT Logos system, with whom I had the pleasure of working many years ago and who introduced me, as a young computational linguist, to the exciting world of Machine Translation.

Finally, a special mention for my husband, who supported and encouraged me in the achievement of this important goal and my children, Giuseppe and Inge, who have been very understanding towards their busy mother, especially in the last few months of thesis writing.

Salerno, 31/12/2012 *Johanna*

Sommario

La Traduzione Automatica si è evoluta insieme alle diverse tipologie di applicazioni di Traduzione Assistita e sono stati raggiunti notevoli progressi nel miglioramento della qualità delle traduzioni prodotte da questi sistemi.

Tuttavia, nonostante i recenti sviluppi positivi nell'ambito delle tecnologie per la traduzione, non tutti i problemi sono stati risolti ed in particolare l'identificazione, interpretazione e traduzione delle cosiddette polirematiche, ovvero di quegli elementi lessicali costituiti da più di una parola come ad esempio *anima gemella*, *carta di credito*, *acqua e sapone*, che hanno una particolare coesione strutturale e semantica interna, rappresenta ancora una sfida aperta, sia da un punto di vista teorico che pratico.

La scadente qualità dell'analisi e traduzione di queste unità lessicali nell'ambito delle tecnologie per la traduzione ed in particolare della traduzione automatica indica che c'è la ancora la necessità di investire in ulteriore ricerca allo scopo di migliorare le prestazioni delle diverse applicazioni per la traduzione.

Le polirematiche rappresentano un fenomeno linguistico complesso, che spazia da unità lessicali con una relativa variabilità di co-occorrenza delle parole a espressioni fisse o semi-fisse. Tali unità sono molto frequenti sia nel linguaggio di tutti i giorni che nelle lingue per scopi speciali. La loro interpretazione e traduzione presenta talvolta ostacoli inaspettati anche per i traduttori umani, soprattutto a causa di intrinseche ambiguità, di asimmetrie strutturali e lessicali tra lingue ed infine di differenze culturali.

Un approccio efficace al problema deve tener conto dei

seguenti aspetti: (i) le polirematiche hanno diversi gradi di composizionalità e, in diversi casi, significati opachi; (ii) la traduzione delle polirematiche è talvolta imprevedibile e una traduzione parola-per-parola può produrre gravi errori; infine, (iii) le loro proprietà morfosintattiche consentono, in alcuni casi, un certo numero di variazioni formali con la possibilità di dipendenze di elementi anche se distanti tra loro all'interno di una frase.

Le attuali tendenze teoriche su questo argomento riguardano tecniche e formalismi diversi, rilevanti per il trattamento delle polirematiche in traduzione automatica, così come anche per altre applicazioni per la traduzione, come ad esempio: il riconoscimento automatico delle polirematiche in contesti monolingui e bilingui, metodologie di allineamento e parafrasi, sviluppo e usabilità di risorse linguistiche monolingui e bilingui e grammatiche sviluppate manualmente; uso delle polirematiche nella traduzione automatica di tipo statistico per scopi di adattamento al dominio, così come ricerche di tipo empirico che riguardano l'accuratezza del modello e l'adeguatezza descrittiva tra varie lingue.

A livello pratico, la questione delle polirematiche è stata affrontata nell'ambito dei diversi approcci alla traduzione automatica: si tratta infatti di una questione di cruciale importanza sia per i sistemi basati su conoscenze, sia per quelli di tipo statistico (*word-based*, *phrase-based* o *factored-based*) nonché per i nuovi sistemi ibridi.

Benché la traduzione delle polirematiche sia un problema noto fin dagli albori della traduzione automatica, rimane ancora irrisolto e dunque la ricerca su questo argomento è suscettibile ancora di possibili significativi miglioramenti.

Recentemente si registra una crescente attenzione verso il trattamento delle polirematiche nell'ambito della traduzione

automatica e delle tecnologie per la traduzione, essendo stato riconosciuto che non è possibile sviluppare applicazioni su vasta scala senza affrontare in maniera adeguata questo problema.

La presente dissertazione, basata sui principi teorici e metodologici della teoria del Lessico-Grammatica, si propone di analizzare quest'area critica della traduzione automatica e presenta un lavoro di ricerca fondato su un'analisi linguistica contrastiva inglese-italiano relativa ai diversi tipi di polirematiche, confrontando i diversi approcci utilizzati per risolvere le difficoltà poste dal trattamento di questo particolare fenomeno lessicale in traduzione automatica.

Il risultato di questa ricerca è rappresentato dallo sviluppo di una strategia di trattamento computazionale delle diverse forme di polirematiche che utilizza fondamentalmente due diversi tipi di risorse: un dizionario bilingue Inglese-Italiano delle polirematiche e un insieme di grammatiche locali per l'identificazione e la traduzione delle stesse.

Tutte le informazioni linguistiche sono state sviluppate con l'ambiente per il Trattamento Automatico del Linguaggio (TAL) NooJ NLP e sono particolarmente utili per superare le attuali limitazioni delle tecnologie traduzione automatica allo stato dell'arte.

Keywords:
Traduzione automatica, Trattamento del linguaggio naturale, polirematiche, dizionari elettronici, NooJ, Trasduttori a stati finiti, Automi a stati finiti, Reti di transizioni ricorsive, grammatiche libere da contesto.

Abstract

Machine Translation (MT) has evolved along with different types of computer-assisted translation tools and significant progress has been made in improving the quality of translations.

However, in spite of recent positive developments in translation technologies, not all problems have been solved and the identification, interpretation and translation of multi-word units (MWUs), i.e a group of two or more words or terms in a language lexicon that generally conveys a single *meaning*, such as the Italian expressions *anima gemella*, *carta di credito*, *acqua e sapone*, in particular still represent open challenges, both from a theoretical and a practical point of view. The low standard of analysis and translation of MWUs in translation technologies suggest that there is a need to invest in further research in order to improve the performance of various translation applications.

MWUs are a complex linguistic phenomenon, ranging from lexical units with a relatively high degree of internal variability to expressions that are frozen or semi-frozen. Such units are very frequent both in everyday language and in languages for special purposes. Their interpretation and translation sometimes present unexpected obstacles even to human translators, mainly because of intrinsic ambiguities, structural and lexical asymmetries between languages and, finally, cultural differences.

An effective processing approach has to take into account issues such as the following: (i) MWUs have different degrees of compositionality and, in many cases, opaque meanings; (ii) translations of MWUs are very often

unpredictable and a word-for-word translation may result in severe mistranslations; finally, (iii) their morpho-syntactic properties allow, in some cases, a certain number of formal variations with the possibility of dependencies of elements even when distant from each other in the sentence.

The current theoretical work on this topic deals with different formalisms and techniques relevant for MWU processing in MT as well as other translation applications such as automatic recognition of MWUs in a monolingual or bilingual setting, alignment and paraphrasing methodologies, development, features and usefulness of handcrafted monolingual and bilingual linguistic resources and grammars and the use of MWUs in Statistical Machine Translation (SMT) domain adaptation, as well as empirical work concerning their modelling accuracy and descriptive adequacy across various language pairs.

On a practical level, the issue of MWUs has been addressed in various MT approaches, whether knowledge-based, statistical (word-based, phrase-based or factored-based) or hybrid.

Although MWU translation is a well-known problem since the beginnings of MT, research on this topic is not yet mature. In general, MWU identification and translation problems are far from being solved and there is still considerable room for improvement. Recently, increasing attention has been paid to MWU processing in MT and Translation Technologies since it has been acknowledged that large scale applications cannot be created without proper handling of MWUs of all kinds.

The present dissertation, grounded in the theoretical and methodological principles of Lexicon-Grammar Theory, investigates this critical area of Machine translation.

The research was based on a contrastive linguistic

analysis of different types of multi-word units and compares the different current approaches to solving the difficulties posed by multi-word unit processing in MT.

The results of this knowledge-driven approach to MWU processing are the development of different processing strategies for the different forms of MWUs using basically two different types of linguistic resources, i.e. a dictionary of English-Italian MWUs and a set of local grammars for the identification and translation of MWUs.

All linguistic information was developed using the NooJ NLP environment which is particularly useful for overcoming the current limitations of state-of-the-art MT technology.

Keywords:
Machine translation, Natural Language Processing, multi-word units, electronic dictionaries, NooJ, Finite-State Transducers (FST), Finite State Automata (FSA), Recursive Transition Networks (RTN) and Context Free Grammars (CFG).

Index of abbreviations and acronyms

British National Corpus	BNC
Computer Assisted (Aided) Translation	CAT
Context Free Grammar	CFG
Enhanced Recursive Transition Networks	ERTN
Finite State Automata	FSA
Finite State Transducers	FST
Google Translate	GT
Google Translator Toolkit	GTT
Information and Communication Technology	ICT
Information Extraction	IE
Information Retrieval	IR
Instant Messaging	IM
Lexicon-Grammar	LG
Linguistic Resource	LR
Machine Aided Human Translation	MAHT
Machine Translation	MT
Multi-word Units	MWU
Natural Language Processing	NLP
OpenLogos	OL
Recursive Transition Networks	RTN
Regular Expression	RegEx
Part Of Speech	POS
Semantico-syntactic Abstract Language	SAL
Semantic Web	SW
Source Language	SL
Source Text	ST
Statistical Post Editing	SPE
Target Language	TL
Target Text	TT

Translate, Edit, Publish	TEP
Translation Memory	TM
Regular Expression	RegEx
Rule-based Machine Translation	RBMT
Statistical Machine Translation	SMT
Example-based Machine Translation	EBMT
Phrased-based Statistical Machine Translation	PB-SMT
Text Annotation Structure	TAS
The Web as Corpus	WebCorp

Symbols

*	Ungrammatical construction
?	Unnatural construction
→	'translates into'

Language codes

En.	English
It.	Italian

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.

MULTI-WORD UNIT PROCESSING IN MACHINE TRANSLATION

Chapter 1 – Introduction

This chapter presents the main topic of this dissertation, i.e. multi-word unit (MWU) processing in Machine Translation (MT).

The starting hypothesis is that proper processing of MWUs applied to an MT process, and in particular to Statistical Machine Translation (SMT), improves output quality.

The method focuses on the analysis and processing of MWUs of different types and the subsequent formalisation of these particular lexical constructions and their translations within the framework of the Lexicon Grammar Theory (Gross, 1975 and 1981).

The goal is to demonstrate that the use of linguistic knowledge of MWU morpho-syntactic and semantic behaviour is of crucial importance to MT processing.

Paragraph 1.1 of this chapter describes the importance of MWUs for MT applications. Paragraph 1.2. provides a description of the scope, original contributions and goals of the work. Finally, the last paragraph summarises the structure of this document and presents previously published works.

1.1. Motivations

Machine Translation (MT), namely the translation process that is performed by a software without any human intervention, is now a reality that is offered to the wide

public by web services which include *E-translation services*, i.e. on-line MT services offered sometimes even for free by rival companies such as Google with *Google Translate* (GT),[1] an on-line MT service for translating text and web pages, or Microsoft which offers the same type of service with *Bing Translator*,[2] or with the *Microsoft Translator webpage widget*,[3] a small MT device, which can be placed inside a web page to allow for simultaneous translation into multiple languages without the user needing to use a separate translation web site.

The actual turning point in the spread of this type of system occurred with the offer of free on-line MT services (Monti, 2004) by some vendors who had realised that the Internet could be a powerful means of advertising their products and services: while the first translation systems were used by a limited number of users, typically large organisations or companies with large translation needs, now, thanks to the success of free on-line MT services, it has gained unexpected popularity with the general public.

The Internet acted as an effective springboard for this kind of service in the Information Society, creating a more extensive and widespread market demand: there are currently about 60 on-line services (Hutchins, 2010), offering automatic translation services both of text and web pages. This figure does not take into account the offer of MT services integrated into other types of services such as, for example, multilingual Instant Messaging (IM) or Cross-lingual Information Retrieval (CLIR) services.

Although these "disposable" translations are still qualitatively very variable and sometimes quite poor,

[1] http://translate.google.it/

[2] http://www.bing.com/translator

[3] http://www.microsofttranslator.com/widget/

millions of people use this type of translation service offered on the Internet on a daily basis. These users, who accept the current limitations of technology and have low expectations as far as the quality of the results is concerned, resort to these services to obtain a translation, which, although of poor quality, enables them to get a rough idea of a text written in an unknown foreign language. The equivalence relation between this type of translation and the source text is very faint, but clearly meets the requirement of overcoming language barriers to a certain degree.

In this way, MT performs the function of an *Assimilation Tool* (Hutchins, 2005:3), i.e. on-line MT services for electronic documents in plain text or web sites, currently offered by many different suppliers such as Microsoft, Google and *Systran* meet the needs of users who want a quick understanding of any text while browsing and surfing the Internet.

It is interesting to note that this feature, which has always been considered a side effect of the main purpose of MT, i.e. producing raw translations as a basis for scientific-technical publishable translations, has been emphasised in recent years by the spread of these free on-line services which gave a significant contribution to the acceptance of MT systems by the general public. Every day, millions of requests are made by users who want to know the content of the texts that circulate in various languages on the Internet in real time.

This can be considered the main function of MT on the Internet since it is the best-known and widespread one, but in recent years MT is being also used as a tool for:

- exchanging information in real time (Interchange tool),

- facilitating the access to information (Information access tool)
- producing publishable translations (Dissemination tool)[4].

As a tool for the rapid exchange of information in chat room discussions or Instant Messaging (IM) systems, as well as in e-mails, MT allows users to communicate in real time with foreign people. These services are known as Cross-Language Instant Messaging (CLIM) applications offered, for example, in the three-dimensional multi-user on-line virtual world by Linden Lab, *Second Life*, but also by Microsoft (*Windows Live Messenger*) and Google (*Google Talk*). MT is provided by means of the so-called (ro)bot translation software or codes that act as contacts in the chat conversations and offer useful features or entertainment to users connected to the network. The so-called MTBots act as partners in IM conversations with the role of translators of the messages exchanged by the participants in real-time.

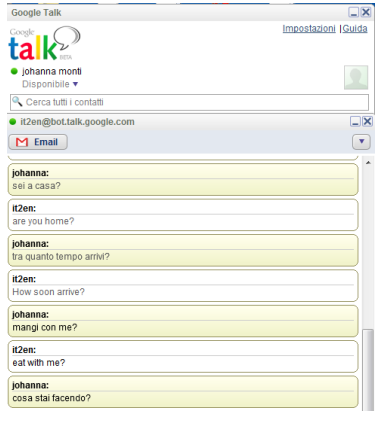


Figure 1 - Google Talk

[4] The different uses have been described by (Hutchins, 2005)

A further recent application of MT is the translation of keywords in a search query. This type of service, known as Cross-Language Information Retrieval (CLIR), has been made available, for instance, by Google to facilitate research and access to information on the Internet[5]: you can search for something simply by entering the search element in your own language and the system immediately translates this element into the desired foreign language. This application is achieved by means of the integration of MT systems with Information Retrieval systems (search and retrieval of information), textual databases, querying systems of structured databases and, finally, search engines.

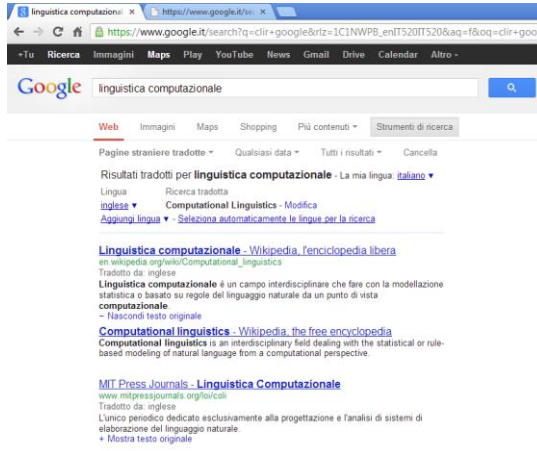


Figure 2 - Translated Search in Google

Until a few years ago, the function for which MT was designed from its beginnings, i.e. the production of raw translations to be used as a basis of high-quality translations,

[5]http://translate.google.com/translate_s?hl=it&layout=1&eotf=1&source=translation_tab

had not yet taken off.

This is the main purpose for which it was developed, integrated into a complete translation process which foresees human intervention either in an early stage of the translation process prior to MT (pre-editing) or in a subsequent phase (post-editing), where the human revision of the so-called raw translation produced by the system takes place.

MT is in this sense only one part of a more complex process (or project) of translation, divided into several phases: analysis of the translation and retrieval of reference material, updating the system, preparation of the text to be machine translated, the translation itself, editing of the translations by professional translators and subject matter experts and quality controls

Commercial systems (including for example Logos, *Systran*, etc.) were primarily designed to serve this function especially in the field of technical-scientific translations and non-literary text types, whose main purpose is to communicate unambiguous and precise information to readers as is the case for software instruction manuals or operating manuals for an aeroplane, characterised by a simple syntax, highly repetitive contents and, if anything, by considerable complexity and density from a terminological point of view.

The possibilities of MT systems in recent years have been enhanced thanks to its integration in translators' workplaces that include other support tools such as electronic dictionaries, translation memories (TM) and tools for the quality control of translation among others.

Recently, the *dissemination function* of MT has been proposed on the web by some MT on-line services, which are based on the collaborative development and maintenance of multilingual content, as is the case for Google, which in 2009

launched a new service called *Google Translate*[6] based on the Google MT system, *Google Translate*, integrated in an on-line translators' workplace where the user can:

1. use additional tools, such as TMs and dictionaries;
2. invite other people (via email) to edit or view the translations for work or revision purposes;
3. edit documents on-line in collaborative mode with other translators, and then publish them in on-line blogs;
4. publish translations directly on Wikipedia.

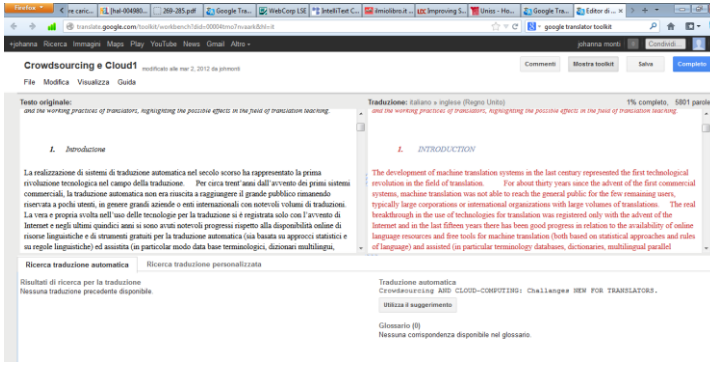


Figure 2 - Google Translator Toolkit

Chapter 2 analyses in detail these recent technological developments, which combine automatic translation with crowdsourcing or collaborative practices on a large scale by which users voluntarily provide feedback on the quality of the translations performed by MT systems.

[6] http://translate.google.com/toolkit?hl=it

Since MT technology is being used so extensively both by the wide general public, which uses MT mainly for information purposes on the Web, i.e. to grasp the general meaning of a text or a web site in a foreign language, and by the public specialised in translation, mainly Language Service Providers and translators, for dissemination purposes, i.e. to produce publishable translations, the quality of MT output should be as accurate as possible.

In these last decades, research in MT has evolved considerably and has led to remarkable improvements in translation quality, but there are still many weak linguistic areas that should be addressed. One of these areas is represented by MWU disambiguation and translation. MWUs designate a wide range of lexical constructions, composed of two or more words with an opaque meaning, i.e. the meaning of a unit is not always the result of the sum of the meanings of the single words that are part of the unit. They are a very frequent and productive linguistic phenomenon both in everyday languages and in languages for special purposes as highlighted by many scholars and are the result of human creativity which is not ruled by algorithmic processes, but by very complex processes which are not fully representable in a machine code since they are driven by flexibility and intuition.

More than a half century has passed since Bar Hillel made the following statement concerning the translation of idioms which are part of a wide class of MWUs: "The only way for a machine to treat idiom successfully is – not to have idioms!"[7], where he settles the problem with the intraducibility of idioms by MT, but MT still presents many

[7] In: "The treatment of 'idioms' by a Translating Machine" presented at the Conference on Mechanical Translation at the Massachusetts Institute of Technology, in June 1952

shortcomings when translating this particular type of lexical unit.

MWUs are not always easy to identify since co-occurrence among the lexemes forming the units may vary a great deal. The most straightforward typology is represented by idioms or idiomatic expressions and proverbs since they represent a unit established by usage as having a meaning not deducible from that of the individual words and with a very limited variability of co-occurrence among the words in the units. Idiomatic expressions like *Hold your tongue* (→ It. *Frena la tua lingua*) and proverbs like *The early bird gets the worm* (→ It. *Chi dorme non piglia pesci*) pose many problems to non-native speakers especially since they cannot be translated literally.

These problems have been widely discussed in translation theory since they are used to analyse the question of the link between meaning and translation, the tiny border between translatability and untranslatability, equivalence and cultural implications. Indeed, the translation of idiomatic expression and proverbs presents many difficulties especially when there is no "natural" equivalent in the Target Language (TL) and it cannot be a simply and mechanical replacement of strings translated word-for-word from the Source Language (SL).

Vinay and Darbelnet (1958) were among the first scholars in translation theory who suggested considering idioms as a unique Translation Unit (TU) and using strategies of oblique translation such as modulation[8], equivalence[9] and

[8] Modulation is a semantic shift, i.e. a variation through change of viewpoint, perspective and very often category of thought, introducing a clarification with respect to the original formulation. (Example: En. *from cover to cover* → It. *dalla prima all'ultima pagina).*

[9] Equivalence is the translation procedure for idioms 'par excellence' since it is used to substitute a TL statement for a SL statement which accounts for the

adaptation[10] to overcome linguistic and cultural obstacles to translating idioms, clichés, proverbs and nominal and adjectival phrases.

Later, other scholars analysed the problem of translating idioms and idiomatic expressions and in general agreed on the fact that literal translation is the worst translation strategy in this case, suggesting different translation strategies (Nida & Taber, 1969; Bassnett-McGuire, 1980; Newmark, 1981 among others).

These concepts expressed in translation theory should be taken into account in MWU processing by MT, which still produces unacceptable translations. For instance, if we try to translate the abovementioned English proverb *The early bird gets the worm* in Italian with *Google Translate*, the result is quite disappointing: **L'early bird ottiene il worm*. Idioms and idiomatic expression should be treated as a single unit and should not be translated literally.

Compound words represent another type of very frequent MWU both in everyday language and in language for special purposes. Here too, the general meaning of the compound word cannot be inferred from the meanings of the different elements of the compound.

These units can have different syntactic functions and can therefore be classified as noun compounds, verb compounds, adverbial compounds, and so on.

Noun compounds are sequences in which a head noun is modified by other elements such as nouns (En. *credit card* → It. *carta di credito*), adjectives (En. *perfect pitch* → It. *orecchio assoluto*) or adjectival locutions (En. *amount of*

---

same situation, even though there is no formal or semantic correspondence. (Example: En. *It's raining cats and dogs* → It. *Piove a catinelle*).

[10] Adaptation is used to replace a SL situation with an analogous TL situation. (Ex.: En. *Yours sincerely* → It. *Cordiali saluti*)

time → It. *quantità di tempo*).

Verb compounds, on the other hand, are lexical units in which the verb is modified by some other elements such as particles (En. *give up* → It. *rinunciare*), prepositions (En. *adapt to* → It. *adattarsi a*), nouns (En. *advance a project* → It. *presentare un progetto*) among others.

Phrasal verbs and light verb constructions or support verb constructions belong to verb compounds. In phrasal verbs, the original meaning of the verb is modified by a particle or a preposition as in the English verb *give*, for which we can have the following phrasal constructions: En. *give away* → It. *dar via, donare*; En. *give back* → It. *restituire, rendere, ridare*; En. *give in* → It. *consegnare, arrendersi*; En. *give off* → It. *emettere, sprigionare*; En. *give out* → It. *distribuire*; En. *give over* → It. *dedicare, consegnare*; En. *give up* → It. *cedere, arrendersi, smettere*; En. *give way* → *cedere.*

In light verbs or support verb constructions, the actual meaning is not expressed by the verb, which has little semantic content, but by some additional expression which is usually a noun, as in the English construct *to give a presentation*, which can be paraphrased by means of an intralinguistic translation in English, with *to present.*

Support verb constructions are very frequent in English and there are several verbs which are semantically weak such as *get, have, make, do* among others. The syntactic properties of sentences with support verbs and predicative nouns have been described, from a linguistic point of view, for a number of languages, in particular for French (Giry-Schneider, 1978, 1987, 2005; Gross, 1984; L. Danlos, 1992), Italian (Elia et al., 1985; De Angelis, 1989; Vietri, 1996), Portuguese (Ranchhod, 1989, 1990), and Korean (Hong, 1991; Shin, 1994; Han, 2000) and in English (Macleod et al., 1997 1998; Danlos, 1992; Krenn & Erbach, 1994; Mel'čuk , 1996) but

they are also very frequent in many other languages.

Word compounds with the grammatical function of adjectives (En. *good-looking*), prepositions (En. *in order to*), adverbs (En. *arm in arm*) and conjunctions (En. *in spite of*) are also quite common.

A particular type of word compounds are term compounds, i.e. various types of compounds, but mainly noun compounds, which belong to a special language. In all languages there is a close relationship between terminology and multi-words and, in particular, word compounds. In fact, word compounds account in some cases for 90% of the terms belonging to a special language.

Contrary to generic simple words, terminological word compounds are mono-referential, i.e. they are unambiguous and refer only to one specific concept in one special language, even if they may occur in more than one domain. For instance, if we consider the word *pay scale* in the financial domain, it can only refer to "the different levels of pay for a particular job, relating to different degrees of skill or experience",[11] denoting therefore a specific and unique concept as opposed to the simple words *pay* and *scale* which are part of the term compound noun, with therefore a one-to-one correspondence, for instance, with its Italian translation *scala dei salari.*[12] Their meaning, similar to all compound words, cannot be directly inferred by a non-expert from the different elements of the compounds because it depends on the specific area and the concept it refers to.

Processing and translating these different types of compound words is not an easy task since their morpho-syntactic and semantic behaviour is quite complex and varied

11 http://dictionary.cambridge.org/dictionary/business-english/pay-scale

12http://iate.europa.eu/iatediff/FindTermsByLilId.do?lilId=1739342&langId=en

according to the different types and their translations are practically unpredictable.

Collocations, defined in Sag et al. (2002) as "any statistically significant co-occurrence" of words, are also non-casual, restricted, arbitrary and recognisable combinations of words (collocates) and represent a wide sub-class of MWUs, for instance Mel'čuk (1998: 24) claims that "collocations make up the lion's share of the phraseme inventory".

Collocations are indispensable in many applications, but particularly in MT, where they can be considered "the key to producing more acceptable output" (Orliac & Dillinger, 2003: 292).

Though collocations are usually semantically compositional, they are notoriously difficult to understand and used by non-native speakers and have therefore a crucial role in the acquisition of a foreign language. Learners need to memorise these word patterns in order to attain fluency in the foreign language.

Collocations are also particularly relevant in translation practice since they cannot always be translated literally, as the following English-Italian examples clearly show: En. *anticipate the salary* → It. *anticipare lo stipendio*; En. *anticipate a pleasure* → It. *pregustare un piacere*; En. *anticipate Ving* → IT. *prevedere di Vinf*.

Several scholars in translation theory have stressed that collocations are one of the translator's major problems such as Newmark (1988), who claims that a key issue in translation is to find a suitable collocation or Hatim and Mason (1990) who state that SL interference can easily lead to an unnatural collocation in the TL.

The unpredictability of word co-occurrence on the basis of syntactic or semantic rules is one of the main

characteristics of collocations, for instance *I did my homework* is correct in English whereas *I made my homework* is not. This means that the verb *to do* cannot be replaced with *to make* in this context, even if they share the same meaning. The translation of collocations requires a correct interpretation of their meaning which is determined by the co-text. Here too, MT presents many shortcomings. If we consider the translation from English into Italian of the English collocation *anticipate a pleasure*, *Google Translate* translates in Italian with **anticipare un piacere*, i.e. a literal translation which is totally wrong.

All these different types of MWU pose serious challenges to MT, especially now that MT is becoming an increasingly more widespread tool used by the general public on the Web for different purposes. Therefore, the main aim of this dissertation is to propose a knowledge-based methodology to correctly identify and translate MWUs.

1.2. Dissertation contribution

The original contributions of this thesis in the field of computational linguistics, and in particular MWU processing in MT with respect to the related work mainly described in Chapter Five is represented by a set of theoretical concepts concerning MWU processing as well as a knowledge-based exemplification of MWU processing in a bilingual context (from English to Italian), using lexical resources and a set of local grammars to handle different types of MWU. This knowledge-based approach allows MWU identification, which currently represents a "pain in the neck" (Sag et al., 2002) for many state-of-the-art MT systems, to be improved.

1.3. Dissertation Structure

1.3.1. Published work

Part of the work presented in this thesis has previously been published in conferences, workshops, journals and as a chapter in books. In some of them, the work was not carried out individually, but in collaboration with colleagues. Therefore, before describing the structure of the dissertation content chapters, I would like to first acknowledge previously published articles and their respective co-authors.

Barreiro et al. (2010) and Monti et al. (2011) addressed the difficulties MWUs present to MT by comparing translations performed by systems adopting different approaches to MT and proposed a solution for improving the quality of the translation of MWUs which adopted a methodology that combined Lexicon Grammar resources with *OpenLogos* (OL) lexical resources and semantico-syntactic rules. It highlighted and discussed the need to create new evaluation metrics and a new MT evaluation tool to correctly evaluate the performance of MT engines with regard to MWU processing and thus to contribute to the improvement of translation quality.

Elia et al. (2011), presented at the WIMS2011 Conference, addressed the problem of MWU processing in Information Retrieval (IR) applications. The shortcomings are mainly due to the fact that these units are often considered extemporaneous combinations of words retrievable by means of statistical routines. On the contrary, several linguistic studies, dating back to the ’60s, show that MWUs, and mainly compound nouns, are almost always fixed meaning units with specific formal, morphological,

grammatical and semantic characteristics. These units can be processed as dictionary entries and become in this way concrete lingware tools which are useful for efficient semantic information retrieval (IR). Elia et al. (2011) present and describe in detail a methodology for MWU processing using tailor-made Linguistic Resources (LR). The LRs developed in this way can be used in NLP applications such as IR, Information Extraction (IE), Information Storage, Machine Translation (MT), ontology development, lexicon-dependent Semantic Web, query-free procedures for knowledge structuring and question answering.

The identification, interpretation and translation of MWUs are crucial aspects in the work of translators, who, in spite of the vast amount of content and knowledge available in electronic format and on the web in recent years, still do not have friendly and targeted tools at their disposal for the various aspects of a translation process, i.e., the analysis phase, automatic creation and management of the linguistic resources needed and automatic updating with the relevant information generated by the computer translation tools used in the process (MT, TMs, and so on). Monti et al. (2011) explore a new approach to helping translators look for different types of information (glossaries, corpora, Wikipedia, and so on) related to the specific translation work they have to perform which can then be used to update the lexical base needed for the translation workflow (both human or machine-aided). This new approach aims to improve the documentary competence of translators in order to process unstructured (textual) information, and make the information on the web or in texts accessible and is based on the automatic identification and disambiguation of MWUs in the texts to be translated by means of CATALOGA, a text mining tool, based on extensive MWU resources for various

domains, which can be combined with an IR application and/or an MT/TM system and used for different purposes.

Finally, Monti (2010) and (2012) analyses the relevant changes that are taking place in MT under emerging phenomena of the Web such as crowdsourcing, i.e., the exploitation of a community/group of people to perform tasks normally performed by employees and cloud computing technologies, which enable ubiquitous access to digital content and on-line multilingual translation tools.

1.3.2. Overview of chapters

This dissertation is divided into seven different chapters. Chapter 1 introduces the research problem and describes the main issues that are addressed in the study. It presents the motivation behind the dissertation and its contribution, summarises the previous published papers in conferences and journals and, finally, gives a brief overview of the dissertation structure. Chapter 2 presents a brief historical review of MT, highlighting the different theoretical and computational approaches in the course of time. The remainder of this chapter is focused on the different uses of MT. Chapter 3 describes all the different MT models from direct translations systems to the current hybrid approaches. Chapter 4 discusses the importance of MWUs for natural language processing and explains what MWUs are in general. Specific paragraphs are devoted to the definition of MWUs, which is still under discussion, and their properties together with the different classifications proposed so far. An in-depth analysis is devoted to the Lexicon-Grammar approach to MWUs. It presents the theoretical background and describes how MWUs are dealt with in this formal natural language analysis framework. Chapter 5 discusses the different approaches to MWU processing in MT from

knowledge-driven approaches to probabilistic models. Chapter 6 describes the empirical and practical work in our research project in detail and, specifically, the linguistic resources used including a dictionary of MWUs and local grammars developed to identify, disambiguate and translate MWUs from English into Italian. The last chapter in this dissertation describes the conclusions of the research project together with a reference to future research work.

Chapter 2 – Machine Translation: state of the art

MT represents in technological and contemporary terms one of the most ancient dreams of man: the possibility to design and construct a machine able to think and act as a human being. The underlying assumption is that the most complex mental mechanism that governs the human activity of translation from one language to another one can be brought back to a set of procedures which can be executed by a computer program.

MT is a translation performed from one natural language to another one by a computer application without any intervention by a human being during the process.

The history of MT, i.e. the attempt to automate the whole translation process is characterised, on the one hand, by the enthusiasm of the researchers involved in the design and development of the systems (who in the beginning especially hoped to obtain results that were comparable to those of translations by professional translators) and, on the other, by the mistrust of the wide audience and the fears of translators.

In this chapter, we provide a brief overview of the history of MT and discuss some current trends in MT technology.

2.1. Brief history of Machine Translation

The first idea of a dictionary based on numeric codes to be used for translation can be traced back to the European Enlightenment and was developed by Descartes and Leibniz. This concept was inspired by a movement that theorised a

“universal language”, i.e. a language based on universally comprehensible logical principles and iconic symbols. This idea was developed with the arrival of the computers and with technological advances in the last century: language universals, i.e.: rules common to all natural languages, seemed to be an ideal basis for the implementation of a software that was able to translate from one language to another without human intervention.

The first step in the development of MT systems occurred in the last century. The first notable attempt took place in the 1930s when Pëtr Smirnov-Troyanskii, a French engineer of Armenian origin, patented a translation machine called the "Mechanical Brain".

The historians date the actual origins of MT back to 1947 and in particular to the conversations and correspondence between Andrew D. Booth, an English crystallographer, and Warren Weaver, director of the Natural Science division of the Rockfeller foundation. In 1949, Weaver wrote a memorandum in which the future of MT was discussed for the first time. This document paved the way for research into MT: in the early ’50s, the first research groups were formed in the USA and in Europe.

These groups received significant funding from local governments and, with the development of Information Technology applications, the first results were obtained and interest in MT grew very rapidly. In 1952, the first conference on MT was held where the first public demonstration of MT was performed. The system used for the demonstration was developed by IBM and Georgetown University in the USA. Its linguistic base contained only 250 words and 6 syntactic rules and it translated a selected set of 49 Russian sentences. This demonstration led to large-scale funding of research in MT in the USA.

Since then, important developments in the design and functioning of translation programs have been achieved. Between the 50s and the 60s there was great unrest among the research groups in the United States (Georgetown University, MIT, Harward, Texas and Berkley Universities), the Soviet Union (Linguistic Institute of Moscow and Leningrad) and the UK (Cambridge Research Unit), who were working on the development of prototypes to demonstrate the feasibility of MT.

In the wake of the transformational-generative grammar conceived by the American linguist Noam Chomsky, the research in the field of MT was mainly directed towards the development of formal grammars, and the syntactical aspects of language. However, this approach reached a dead end in the mid 60s when it became clear that syntax alone did not represent a satisfactory basis for MT: the results produced by syntax-rule based MT systems were very discouraging. The idea of a universal language based on syntactical principles came into conflict with the problem of the polysemy, ambiguity and complexity of the natural language.

Overcoming the semantic barriers became a real problem for all the researchers involved in this field. Once it was clear that a syntactic approach was completely useless in the comprehension and interpretation of a text, necessary steps in translation, the initial enthusiasm subsided and very negative judgments concerning the future of MT gained the upper hand.

In 1966, the Automatic Language Processing Advisory Committee Report (ALPAC) on the future prospectives of MT did not foresee any particular usefulness in MT and did not consider further investment and research funding in this field necessary.

The scientific community in other countries did not agree

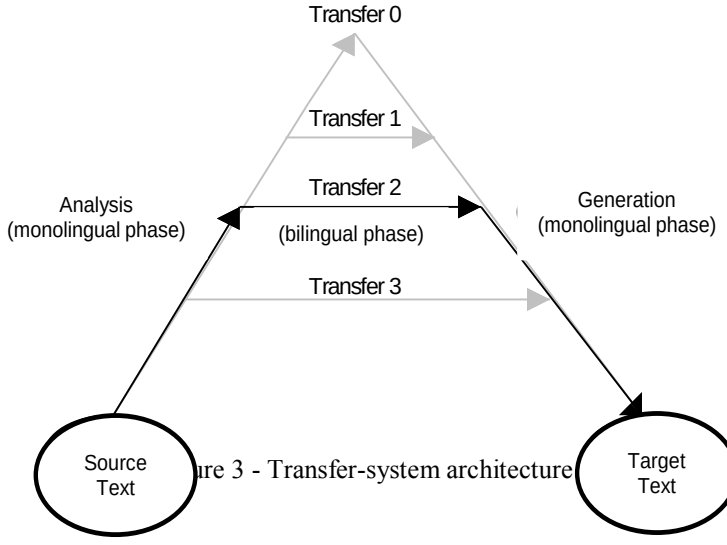
on this negative judgment and other research groups were formed mainly in Canada and Europe, where the issue of multilingualism was important.

In Canada, the Meteo system was developed to translate the meteorological bulletins of the broadcasting service. At the same time, the EEC heavily invested in MT, first of all with the adoption of SYSTRAN for the translation of legal, scientific, technical and administrative documentation. The EEC then decided to fund the ambitious EUROTRA project whose main aim was the development of a pre-industrial advanced multilingual MT system for all European languages based on an interlingua structure.

However, the project did not produce an operative system and it ended in the late ’80s. Even if EUROTRA did not achieve its main aim, i.e. the development of a multilingual translation system from and into all European languages, nevertheless it stimulated transnational research in the field of computational linguistics.

Here too, research that focused on the development of systems based on semantic models which were believed to be more effective than syntactic ones, reached a dead end and the results were quite disappointing.

Hence, the idea of a metalanguage based on linguistic universals was abandoned in favour of a less ambitious but more pragmatic approach which produced better results in terms of quality, i.e. the transfer approach, still used by many commercial systems on the market, such as *Systran*, for example. The transfer systems are based on a structure divided into three stages.



ure 3 - Transfer-system architecture

The first stage is represented by analysis of the source language which has as a result the transfer from a natural language to an abstract representation of the language itself, from a lexical, syntactical and sometimes semantic point of view: this intermediate abstract representation is the basis of a subsequent stage of transfer in which the abstract representation of the source language is transformed into the corresponding abstract representation of the target language. The last stage of generation converts the abstract representation of the target language into the corresponding natural language.

The most significant development in the ’80s was the availability of the first MT commercial systems including the LOGOS system in the USA, SYSTRAN in France and METAL in Germany.

In the late ’80s in France an MT service, called Minitel, was offered by the French postal service to a wide audience on its network. This MT service based on *Systran* for various

language pairs had many disadvantages: it was quite expensive and slow and it was not integrated into a PC environment.

The ’90s were characterised by a variety of different approaches in MT research: systems based on word-for-word translation and transfer-based systems coexisted with systems which were experimenting new theories. From the mid ’90s onwards, the number and the typologies of applications for translation increased rapidly: MT systems, assisted translation systems, translator’s work environments, translation memories and on-line MT systems became available on the market.

The myth of a machine able to translate like a human being was abandoned in favour of a more pragmatic approach to the translation problem which mainly addressed the real needs of users and made usable tools for the translation process available on the market.

The real turn in MT took place with the spread of the Internet and the development of on-line MT services by various software developers. In 1996, *The Language Engineering Directory – A resource guide to Language Engineering Organisations, products and services* presented the results of a research performed to outline the state of the art in the language industry. In the section devoted to the services of language engineering in the category "Machine Translation via Modem/Minitel" six companies are mentioned that, at that time, already offered this type of service: Compuserve Inc., Globalink Inc., Language Engineering Corporation, Nec Corporation – C&C IT Research Laboratories, Smart Communications Inc. and *Systran* SA.

It was *Systran* that gave a decisive boost to the spread of on-line MT system services by entering a joint venture with

AltaVista, the famous research engine and by offering the first free MT service in real time to the wide audience on a domain called *Babelfish*, a concept taken from *The Hitchhiker's Guide to Galaxy* by the science fiction author Douglas Adams, where galactic hitchhikers were able to understand every language simply by activating a small yellow fish in their ears.

This first experiment, which started in 1997, became a great success in a very short time, as highlighted by Jin Yang and Elke D. Lange, in two articles about the on-line MT services offered by *Babelfish* (1998, 2003): the number of translation requests by users increased from 500,000 in May 1998 to 1.3 million per day in 2000.

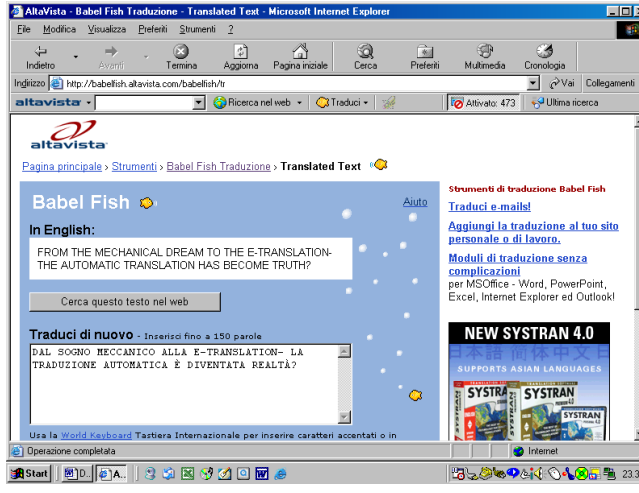


Figure 4 - Babelfish

Starting from 1997, other free MT services spread very rapidly to promote the products which these services were based on and create a market by attracting the internauts to use and test MT. Some of these services also offered human

revision of the translations produced by MT at an extra price. MT began to be widely used by the general public, stimulating the language industry towards new and more user-friendly solutions.

Nowadays, almost all MT vendors offer free MT services and provide MT gateways to large customers. Millions of Web pages are translated on the fly every day into more than sixty language pairs.

The Internet has changed the way in which this technology is considered by a wide audience and the way in which it is used, opening up unexpected perspectives for MT and contributing to its qualitative improvement.

Over the last decades, MT has become a fast moving research area characterised by:

1. the availability of open-source systems, like *Openlogos*,[13] *Moses*,[14] *Apertium*,[15] *Asia Online*[16] MT systems which not only contributed to stimulating academic debate and experimentation in this field, but also to attracting potential users and translation industry investors.

13 An open-source derivative MT system of the former Logos RBMT engine, offered by DFKI at the following web address: http://logos-os.dfki.de/

14 An open-source SMT, developed within the framework of the European project EuroMatrix, a large scale EC funded effort to develop MT engines for all possible European language pairs: http://www.statmt.org/moses/.

15 An open-source RBMT system developed in the OpenTrad project by the Universidade de Vigo, the Universitat Politècnica de Catalunya and the Universitat Pompeu Fabra): www.apertium.org/

16 Host for a series of consumer-oriented portals that provides public access to global information, news, science, education, literature and more in local Asian languages. It uses state-of-the-art translation systems and crowd-sourcing.

2. the increasing offer of MT on the web as a standalone tool or integrated in other applications such as IM systems, CLIR and Translators' virtual workplaces, among others.
3. the increasing offer of translation services based on MT technology, both for information purposes, with the commitment of important actors in the scenario such as Google, Microsoft and social networks like *Facebook*, *Twitter*, *LinkedIn*, and for dissemination purposes thanks to the integration of MT with other translation tools such as TMs and linguistic resources of various types (corpora, electronic dictionaries, glossaries). In this scenario, the use of new translation methodologies, and in particular crowdsourcing, and advanced IT technologies, i.e cloud computing, is a sign of a new technological turn.

The next section of this chapter will address in detail this latter trend which is particularly important for the improvement of MT engines and SMT in particular.

2.2. Current trends: crowdsourcing and cloud computing

Over the last fifteen years, we have witnessed a complete turn in the availability of linguistic resources and free machine and assisted translation tools on the Internet. Emerging web phenomena such as crowdsourcing, i.e., the exploitation of a community/group of people to perform tasks normally performed by employees (Howe, 2006) and cloud computing, which allows users ubiquitous access to services and on-line tools for translation and multilingual

digital content, are further changing the scenario. In particular, the combination of crowdsourcing and cloud models of automatic/assisted translation is taking place on a large scale inside collaborative translation platforms.

In the translation field, crowdsourcing refers to the use of professional and non-professional translators to perform typical translation and localisation tasks either on a paid or a voluntary basis. Common Sense Advisory, an American market research company, has coined the acronym CT3, or "community, crowdsourcing, and collaborative translation," which collects the different denominations used to highlight the main feature of this emerging phenomenon, i.e., the collaborative aspect within a community of professionals or occasional translators who belong to a "crowd" of volunteers willing to contribute to translation tasks.[17]

Generally, this practice of exploitation of collective intelligence in the field of translation is performed as follows:

- the documents to be translated are shared on the web. This sharing can occur either within dedicated environments and is therefore addressed to a group of professional translators or on sites open to the public where the work takes place on a voluntary basis and, in this case, is aimed at non-professional and occasional translators;
- the work performed by professional, occasional and non-professional translators is then submitted to a review process which can again be assigned to professionals and non-professionals depending on the type of text and the purpose of translation;

17 http://www.dqglossary.com/simple/thoughtData/3734.html

- professional translators are usually paid in a conventional way, but volunteer translators, working for free, are paid through non-conventional forms of social gratification such as the attribution of a score in the list of the people who contribute to the translation up to public recognition of leadership when they reach the top of the list, or simply the opportunity to learn something new.

The idea of using crowdsourcing in translation is based on the need to execute translation projects in a short time. It allows large volumes of translations to be produced in a short time, at low cost with acceptable quality. Therefore, it seems to be an adequate alternative in terms of costs and quality both to MT which produces large amounts of translations which are of low quality and professional translators who produce quality translations but at high costs. On the contrary, it very often requires the combination of both these elements, i.e., professional or occasional translators edit MT outputs.

Since 2006, this form of exploitation of collective intelligence in the field of translation has paved the way to collaborative practices of translation on a large scale, which, on one hand, are based on the active involvement of translators, including non-professionals, in localising open-source products and on-line platforms and, on the other hand, voluntary feedback by users regarding the quality of MTs.

Examples of this alliance are now widespread, but the true pioneers of this practice were social networks such as *Facebook*, *LinkedIn* and *Twitter* which were localised in many different languages thanks to the work of their followers. In particular, in 2008, Facebook launched its application *Translations*, in order to localise the interface in

new languages and translate the continuous updates to the platform. In this way, Facebook has been localised in over 70 different languages (with about 100,000 words for each version) at a surprising speed (for instance the entire French version was translated by 4,000 users in 14 hours). The localisation and translation strategies used by Facebook are based, on one hand, on the free work of its supporters and, on the other, more recently, on *Microsoft Bing Translate* for the translation of posts.

In *Translations - Go vote on translations* users can choose the best translation from the possible solutions suggested by the system or, if they do not like them, translate from scratch. The social dimension of the activity is fed by the *Facebook Translations Team* group which is used by the members of the management team to communicate with translators on various technical aspects and in which translators can discuss their problems, ask for tips and give advice on possible translation solutions.

In the abovementioned examples, crowdsourcing is used not only in order to reduce costs, but also to translate in commercially unattractive languages and finally as a means to increase and loyalise users by giving them the possibility to shape the image of Facebook according to their tastes and expectations. Thanks to the active involvement of users in the localisation of the French, German and Spanish versions, Facebook, for example, recorded an increase that went from 52 to 124 million hits (Britton & McGonegal, 200; Eskelsen , Marcus, & Fereee, 2008).

Therefore, localisation is the main engine of crowdsourcing since this new way of translating offers considerable advantages for large companies with regard to the localisation of website contents and their products, but also to the development of language resources for translation

projects and the training of translation software applications. For instance, IBM launched the project *no.Fluent* to build a multilingual parallel corpus[18] using its voluntary employees around the world. One year after the start of the project about 3,000 volunteers had contributed with approximately 36 million words (mainly chat messages and translations done collaboratively), editing the translations done by the IBM MT system.

However, localisation is no longer the only aim of crowdsourcing since it is also used in subtitling, e.g. in *dotSUB*[19] or *TED*,[20] and even for literary translation, e.g. for the translation of the Harry Potter saga into German.[21]

Crowdsourcing is thus adopted as a novel approach to performing all the different phases of a complete localisation/translation process, as highlighted by Désilets (2011), who identifies several forms of crowdsourcing that affect translation from organisational TeamWare and specialised sites for translation to the availability of platforms for:

- creating and sharing terminology resources and translation memories, i.e., Wiki platforms such as the *Worldwide Lexicon Project*,[22] an open source collaboration platform based on a huge database of translations usable for any website or web

[18] A parallel corpus is a collection of source texts and corresponding target texts in which sentences are aligned and constitute parallel sentences.
[19] http://dotsub.com/
[20] http://www.ted.com/translate/languages/it
[21] www.had-community.de/HaD
[22] http://www.worldwidelexicon.org/home

application. Other examples are *UrbanDictionary*,[23] *TermWiki*,[24] *WeBiText*,[25], TAUSData Association;[26]

- distributing parts of large translation projects to professional or occasional translators such as *My Genco*[27] (for professional translators) or *Mechanical Turk*[28] (for non-professionals), virtual platforms where buyers can communicate and conduct transactions with translation suppliers;

- providing translations or editing MTs such as the collaborative translation environments *Google Translator Toolkit*, or *Geoworkz*[29] by Lionbridge.

This latter type of platform, based on the interaction of crowdsourcing and cloud models of assisted translation systems, requires a closer examination for its impact on the improvement of MT and MAHT applications.

As an example, *Google Translate* is a free collaborative translation environment where users can submit their documents to MT and MAHT processes, revise, edit and store translations in translation memories and invite other people (via email) to share the translation or editing work.

Translation memories created by users contain invaluable information for the Google MT engine which is based on the use of parallel corpora, i.e., original texts aligned with the corresponding translations, stored by users on the platform made available by Google.

---

[23] http://www.urbandictionary.com/
[24] http://it.termwiki.com/
[25] http://www.webitext.com
[26] http://www.tausdata.org
[27] http://mygengo.com/
[28] https://www.mturk.com/mturk/welcome
[29] http://www.geoworkz.com/support/training.aspx#Translator

There are therefore clear benefits both for users, who can access a free repository where they can process their translation work, using what has been previously translated by themselves or by other users, and for Google, which draws on the translations stored in translation memories to improve the performance of its system. However, this is a collaborative environment for occasional translators since there are limits to the amount of data and formats that can be used, there are no typical translation memory features such as fuzzy matching and no quality control procedures and, finally, there are data confidentiality issues.

Nevertheless, it was one of the first translation platforms of its kind and it has inspired collaborative professional translation environments that allow ubiquitous access to digital content and on-line multilingual translation tools within a team such as *Geoworkz* by Lionbridge, a fee-paying environment for translation service providers and professional translation companies, based on SaaS solutions[30] which provides access and real-time updates to translation memories, glossaries and features for data sharing within a team, and also between customers and suppliers. More and more software translation tool vendors are incorporating their products into collaborative translation platforms, including *MemoQ Server*[31] by Kilgray Translation Technologies, *The Translation Network* by LingoTek,[32] *Crowdin*,[33] *Wordbee Translator*,[34] *Wordfast Anywhere*,[35]

---

[30] ***Software as a service*** (*SaaS*) is a software distribution model in which a software company develops, and manages a web application available to customers on the Internet, allowing ubiquitous access to products as a fee-paying service.

[31] http://kilgray.com/products/memoq-server

[32] http://www.lingotek.com/

[33] http://crowdin.net/

[34] http://www.wordbee.com/

XTM Cloud.[36]

The process of translation has been significantly changed by the use of this new generation of translation technologies and in particular by collaborative environments in which the interaction man/machine is particularly significant. Cloud applications offer useful tools to translators such as automatic/assisted translation tools, glossaries, translation memories, editing features together with software applications for cooperation between the different actors in a translation process (translators, editors, terminologists, customers and so on) such as IM applications.

The first change concerns the use of automatic translation and translation memories in the translation process. These are no longer an option for translators but an integral part of the workflow. The combination of translation memories with automatic translation and the terminological resources, prepared in the preliminary phases of the translation process, is a main feature in all the various models of collaborative environments available on the web, from *Google Translate* to commercial environments such as *Geoworkz* or *MemoQ server*.

This means that the translation process now has to be carried out using translation technologies, something that was unthinkable up to a few years ago. As a matter of fact, translation memories were only used in advanced technological sectors of the translation market such as in localisation processes (Monti, 2007) whereas MT was only used for technical translations by large companies or bodies.

Nowadays, these technologies, integrated into collaborative environments, are used for every type of translation by a large audience of specialists (translation

[35] http://www.wordfast.net/?whichpage=anywhere

[36] http://www.xtm-intl.com/xtmcloud

service providers, professional translators, editors) who have to adapt their interaction and work practices to the new work modalities.

With regard to this point, Kelly et al. (2011) highlight the change from the Translate, Edit and Publish (TEP) linear model of the translation process to a new model based on the abovementioned translation technologies and cloud computing applications in which the work is performed at the same time by different members of a translation team, even on the same document, as happens for instance in *Google Translate*, where modifications are made available to all the people who share a document.

This new way of working is called "parallel translating", which not only refers to the traditional distribution of a large amount of translation work in a translation group, but also to translating and editing the same documents simultaneously and in real time. It considerably shortens the translation process and offers further advantages such as the availability in real time of the editor's changes or quality controls to translators.

The traditional concept of the translation group, based on vertical management of translation jobs, and in particular of big translation jobs where a project manager organises the translation process according to the TEP model and where the translators' task is limited to the part of work they have been assigned to, is replaced by the concept of a translation community.

In the community, translators interact continuously and in real time with peers and contribute through exchanging ideas, suggesting best practices, searching for relevant information and solving translation problems. The concept of "community", which highlights the social dimension of interaction in order to achieve a common goal, was initially

used to refer to the communities of occasional translators who voluntarily joined a translation project. Nowadays, it also refers to communities of professional translators who take advantage of being members of these communities in different ways: by finding information, developing language resources (glossaries, terminological resources, translation memories) on a collaborative basis and interacting with the other members of the community.

The community is based on the use of new translation technologies so that translators become post-editors of translation produced by machine or machine assisted translation systems. Post-editing becomes, indeed, the main activity of translators whose creativity, usually used to solve translation problems, is now expressed in quite a different way from the past since it has to take into account ready-made solutions identified by the translation systems. Many scholars have recently analysed this issue from a theoretical point of view (Austermühl, 2001, 2006; Corpas Pastor & Varela Salinas 2003; Esselink, 2000; Pym, 2003; Torres del Rey, 2005), but also with reference to translators' training (O'Brien, 2002).

A new element in this context is represented by the fact that, thanks to the crowdsourcing used by companies, the translations edited by translators are used to improve the outputs of translation technologies, something that was previously not possible. As an example, in *Google Translate*, the edited translations are valuable resources and, more specifically, valuable parallel corpora used to train the statistical engine of *Google Translate* so that its outputs become more and more reliable. It is a virtuous circle put in place by translation software developers as highlighted in the following figure:

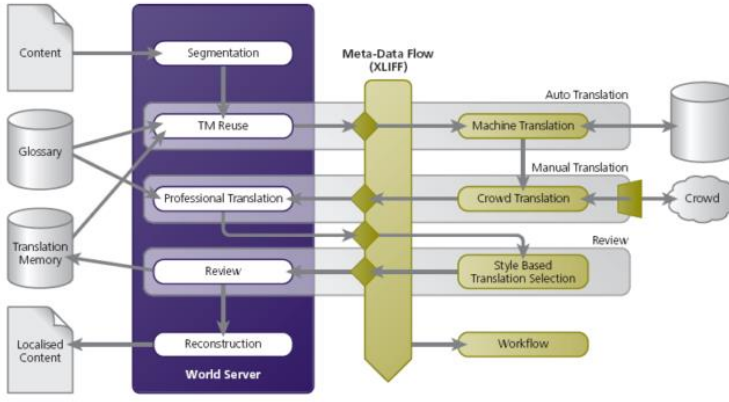


Figure 5 - Use of crowdsourcing in a translation process (Carson-Berndsen et al. 2010)

Translators use MT and MAHT to speed up the various stages of a translation task. In this way, they provide valuable linguistic resources which can be used to tune the products and can be reused in new translation projects.

Translation vendors are increasingly choosing to use and, in some cases, develop proprietary collaborative translation environments in order to ease the management work of translation orders, data control and the quality of the final product. Significant improvements in the quality of translations based on post-editing MT and MAHT outputs have been achieved thanks to the re-use of translator's knowledge.

2.3. Conclusions

In the digital age, MT has found its raison d'etre and has abandoned the pretensions of an impossible dream, namely

that of a "thinking machine" able to produce results comparable to human thought, in order to become a means, with all its limitations, of effective multilingual communication.

The Internet and MT have formed a strong alliance, becoming indispensable for each other: the Internet has contributed to the knowledge of MT and made it, beyond any reasonable expectation, usable by the general public and useful for overcoming language barriers in the global village. MT has also made the Internet a very efficient communication and information retrieval tool.

Chapter 3 – From direct translation to hybrid MT systems

In the previous chapter, we provided a brief history of MT together with an analysis of current trends.

The goal of the present chapter is to provide a general overview of the different approaches that have been used and are currently adopted in MT architectures. What follows is a schematic view of the operation of MT systems which is considered sufficient for the purposes of this dissertation. For a more detailed description, the reader is referred to the various introductions to MT technology which have been produced so far (Hutchins & Somers, 1992; Koehn,2009; Wilks, 2009; Nirenburg, Somers, & Wilks, 2003).

In Section 3.1, we start with a brief overview of the various linguistic approaches which have been adopted starting with the first type, i.e. the direct approach to the most widely used one in commercial systems, i.e. the transfer approach. In Section 3.2. we describe the knowledge based approach which was developed in the framework of Artificial Intelligence methodologies. Section 3.3. describes the empirical approaches, i.e. the example-based approach and the statistical one. The final section of this chapter presents the current trend in the field of MT research, i.e. the hybrid approach which tries to merge linguistic and empirical methodologies and capitalise on the advantages of both.

3.1. Linguistic approaches

From its beginnings until the early 1990s, almost all MT technology relied on a linguistic approach based on large collections of linguistic resources, both dictionaries and grammars rules, to analyse the source language and then map the syntactic and semantic structure into the target language.

This type of approach, known also as Rule-based Machine Translation (RBMT) has three different strategies: the direct translation method which maps input to output with very simple rules, the interlingua method which uses an abstract meaning representation and finally, the transfer approach which relies on an intermediate abstract representation of the natural language and uses morphological, syntactic and sometimes semantic information. The architecture of these different approaches is graphically represented in the so-called *Vaquois triangle*, as illustrated in the figure below:

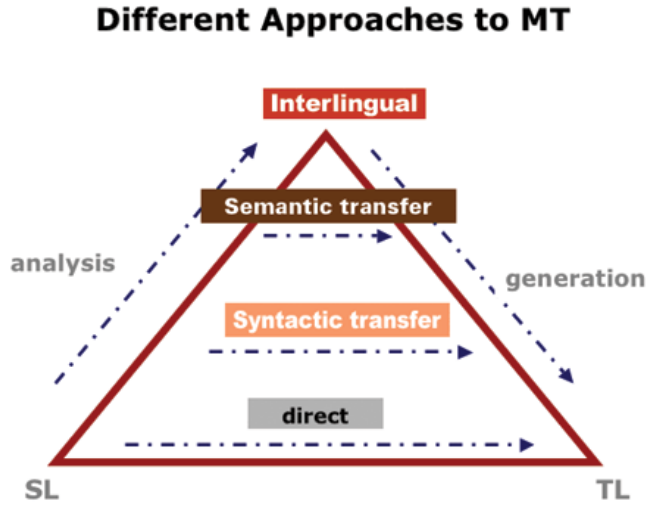


Figure 6 - The Vaquois triangle

The linguistic approaches basically follow the same procedure, albeit with some differences concerning the depth of linguistic analysis: the SL is analysed on the basis of the linguistic resources stored in the MT, there is then some sort of transfer and finally, generation in the TL takes place.

The analysis of the SL produces a linguistic annotation of the ST with information which is mainly morpho-syntactic, i.e. the morphological inflection pattern, and the syntactic function of the word, stored in a monolingual lexical database.

Some sophisticated lexical databases, like the *Openlogos* one, may also contain semantic information concerning the conceptual formalisation of things, ideas, relationships, dispositions, conditions, processes, etc. During this phase, the words of the ST are matched against the dictionary and identified, in some cases also by contextual rules when words are ambiguous, i.e. they may have different parts of speech, as in, for example, the Italian word *porta* which can be both a noun and the 3rd person singular of the present simple tense of the verb *portare*. A simple disambiguation rule resolves this type of ambiguity, by establishing, for instance, that if the word *porta* is preceded by a determiner, then it is a noun.

Subsequently, a parser that applies more complex syntactic rules identifies and segments all the main structural constituents of the ST, i.e. noun, verb, prepositional phrases. Once all constituents are identified, the transfer phase matches the SL elements with the TL ones. It can be very simple as in direct systems, where it consists of a simple match at lexical level between ST and TT, or more complex as in the transfer approach where it is based on transformational syntactic rules to produce the equivalent information in the SL.

The last phase is represented by the generation in the TL which takes into account the results of transfer phase, i.e. a sequence of annotated words or larger units with a description of their morpho-syntactic and semantic features and produces the equivalent output in the TL.

3.1.1. Direct translation systems (First generation systems)

Historically, direct translation systems are the first type of MT designed in the ’50s and ’60s and known as first generation MT systems. They are called direct translation systems because translation was performed directly from one natural language to another, trying to avoid intermediate and lengthy passages where possible, as occurred in other types of MT systems. These systems produced a word-for-word translation on the basis of a bilingual dictionary and only later on, they used some sort of syntactic analysis of the ST. They are basically systems developed for a specific language pair in only one direction (unidirectional systems), i.e. the source texts are only analysed to generate texts in one specific TL.

Direct systems are characterised by the lack of:

- a complex intermediate stage in the translation process. Translations are performed by the simple transposition of a sequence of source words in an equivalent sequence in the target language.
- the analysis of the syntactic and semantic relations both in the SL and the TL. As shown in the following image, there are no analysis and generation modules both at the syntactic and semantic level.

Figure 7 - Direct MT flow chart

The translation process consisted of three stages: morphological analysis, translation using a bilingual dictionary and re-arrangement of word order of the text obtained in this way so that it respected the word order in the TL.

During morphological analysis, analysis of the text was performed by recognising the inflected forms of the words in the text, brought back to their canonical form (for instance the infinitive form for the verbs). At the end of this phase, a string of words in canonical form was obtained and used for the subsequent phase of finding the equivalents in the TL. Once the word string in canonical form in the ST was obtained, the MT system went on directly with the translation using a bilingual dictionary, without going through a further analysis phase of the ST from a semantic and syntactic point of view. Starting from the canonical form in the SL, the equivalent form was looked for in a bilingual dictionary and finally the local organisation of the word order was performed on the basis of very simple rules (for example in English: move all adjectives before nouns) to reflect the word order in the TL.

Obviously, this type of system produced low quality results, first of all because the computational capacities of

former computers (we're talking about the '50s) were quite limited, but also because there was no linguistic approach to translation problems: it was a word-for-word translation with no disambiguation process on a semantic and syntactic level. The limitations of this approach were evident, even though it has to be said that, in the course of time, this type of system has evolved and has left traces in some more recent uni-directional bilingual systems such as the Météo system: these systems basically exploit the similarities between SL and TL to translate, making use of the syntactic and grammatical modules to analyse the parts where the two languages differ.

3.1.2. Indirect systems (Second Generation Systems)

The poor results obtained by direct systems forced researchers to move in other directions. The idea that the text of the SL could be turned into an abstract intermediate representation that could then in turn be transformed back into the text of the TL appeared in academic research. This idea was clearly expressed by C. Cherry in his essay *On human communication* (1966: 117): these systems "transform from a source language A to a target language B, using rules expressed in a third language C".

Unlike direct systems described in previous paragraph, indirect systems make use of an intermediate step between the SL analysis and TL generation. Depending on the degree of abstraction of this intermediate stage, we have two types of indirect systems: interlingua systems and transfer systems.

3.1.3. Interlingua systems

Interlingua systems are based on the belief that texts can be converted from and into an abstract representation common

to more than one language. The intermediate representation (interlingua) contains all the information required to generate a text in any target language without having to go back to the text. The interlingua, which is an abstract representation common to several languages, is neutral with respect to both the SL and the TL. The initial idea was to develop an interlingua that was truly universal and thus universally valid for all possible combinations of languages but this goal soon proved to be utopian.

Interlingua systems consist of two phases:

1. the first phase is represented by the transition from the SL to the interlingua: the process of analysing the ST on lexical, semantic and syntactic levels results in an abstract representation that, despite being neutral with respect to the ST, contains all the information needed for subsequent generation of the text in the TL,
2. the second phase is represented by the transition from the interlingua to the TL: the process of generating the TL at lexical, semantic and syntactic levels starting from the interlingua, results in the production of a text in the TL.

Several artificial languages have been used as interlingua and in some cases Esperanto too.

Interlingua systems are characterised by:

- modules for analysis and generation: the analysis of the SL produces an independent representation of both the SL and TL;
- the possibility to add new languages in an economical way: the interlingua systems are by

definition multilingual systems, being an abstract representation that is independent of the specific natural languages and common to all languages.

As shown in the figure below, the development of new modules of generation and analysis exponentially increases the possibility of translation to and from different languages.

If we add for instance a Spanish analysis module, we will immediately get three more language pairs: English-Spanish, Spanish-French, Spanish-Italian.

Figure 8 - Interlingua system architecture

However, this feature of interlingua systems also represents a limitation since it is difficult to define an interlingua that is truly universal or at least common to more languages, even for languages that belong to the same family such as the Romance ones.

3.1.4. Transfer systems

In the ’80s, the majority of MT commercial systems available on the market, such as *Logos*, *Systran* and *Metal*, were based on the transfer approach which was to a certain extent advantageous due to the high degree of modularity and reusability of its components.

The linguistic data used in this type of system were essentially monolingual and bilingual dictionaries and grammars. Based on the linguistic information provided by their linguistic modules, the transfer systems performed an analysis of the sentence in the SL both on a morphological and syntactic level. The structure of the source sentences, identified in this way, was then transformed into a meta-language and finally from this meta-language into the TL.

Unlike interlingua systems, transfer systems are based on a three-phase structure

- Analysis of the source language text
- Transfer
- Generation of the target language text.

The first phase is therefore represented by an analysis of the SL text which results in a transition from the natural language to an abstract representation of the language itself, both from a lexical and grammatical point of view (and in some cases from a semantic point of view as well). This intermediate abstract representation is the basis for the next transfer phase where the conversion of the abstract representation of the SL into the corresponding abstract representation of the TL takes place. The last phase of generation transforms the intermediate abstract representation of the TL in the corresponding text into the

target natural language.

Analysis and generation modules are generally independent of the specific language pairs and can be reused for other language pairs whereas transfer is specific to a language pair.

These systems are therefore more complex but obtained better translation results compared with the systems previously described, for a number of reasons:

- unlike direct systems, transfer systems were characterised by a great modularity and reusability of linguistic data since the various components were not related to specific language pairs with the exception of the transfer module. The analysis module of a language, as well as the generation one, could be reused for other language pairs.
- unlike interlingua systems, the transfer module allowed for greater flexibility in the definition of the intermediate representation since the level of abstraction needed for the definition of a universally valid structure for all languages (independent of any language) was not necessary and the definition of an intermediate representation valid for a specific language pair (depending on the language pair) was sufficient.

So far we have discussed systems based on linguistic approaches to the problem of MT, but in recent decades, different research approaches have emerged, essentially related to the development of knowledge bases, statistical models and large textual bilingual and multilingual corpora.

3.2. Knowledge-based systems (Artificial

Intelligence)

The main problem with systems based on a purely linguistic approach is the inability to solve the problems of the so-called "semantic barrier": linguistic analysis alone makes it possible to disambiguate texts from a morpho-syntactic point of view, but it does not provide a real understanding of the text. The main objective of research in the field of artificial intelligence is the solution to problems related to the understanding of texts on the basis of knowledge and that is why there has been a growing interest in the application of major developments in this field to translation since the ’70s.

The assumption is that the integration of artificial intelligence technologies and in particular knowledge bases for specific domains will help to "understand" the real meaning of a text and therefore produce more accurate translations.

Since the understanding of the meaning of a text is the main goal of AI research, semantics plays a leading role in the syntax. In this perspective, semantic models become the central element of MT systems. The development of semantic representations of the meaning of texts and the use of knowledge bases represent the pillars of the semantic analysis needed for the interpretation of a text and become a priority with respect to the development and use of syntactic analysis models of natural language.

According to AI researchers, syntactic patterns alone are not able to disambiguate natural language since they are not able to grasp all the complexity of meanings in different contexts. The content (meaning) of a text and its function are important elements to consider in a correct and effective analysis phase but lie outside the mere syntactic analysis since they are not linguistically expressed in the text.

During the ’70s, AI researchers begun to work on MT projects with Yorick Wilks carrying out research at Stanford University and Roger Shank at Yale University. In the ’80s, there was a growing interest in AI applied to translation in Europe (with the Eurotra project), Japan and the United States, and in particular at the Carnegie-Mellon University in Pittsburgh, where Jaime Carbonell and Sergei Nirenburg developed a prototype based on a methodology described as "meaning-oriented MT in an interlingua paradigm", i.e. a methodology oriented towards the meaning in the context by an interlingua system. Most of the research was devoted to the development of knowledge bases, hence the name, "Knowledge-based Machine Translation". The Carnegie-Mellon University prototype, developed for the English translation of Japanese PC manuals, was based on:

- an ontology of concepts;
- analysis and generation dictionaries (English and Japanese);
- analysis and generation grammars (English and Japanese);
- rules of correspondence between the interlingua and the syntax of English/Japanese.

The system was based on a very limited vocabulary of 900 words and contained 1500 concepts, mainly related to interaction between users and the computer.

The concepts are represented in the forms of conceptual structures (which provide an intrinsic characterisation of the concepts) linked together in a hierarchical network. The importance of this prototype is that it explored the feasibility of an MT system based on a conceptual interlingua, specific

to a domain (the computer) but independent of any particular language.

Currently, several systems use features and semantic rules, but this does not necessarily imply that they use AI technologies. This is because even if the semantic features in some way contribute to the definition of the attributes of a real object, in this type of approach, their definition is not the ultimate goal, but it is much more important to define the hierarchical semantic network which governs use in specific domains. We can affirm that at present AI technologies have not been used extensively in MT systems, but are limited to a few prototypes.

The main reason is that MT systems designed to translate any type of special language text require the development of large databases that can store huge amounts of data.

3.3. Empirical approaches

The approaches described so far are mainly based on linguistic knowledge whereas the most recent research trends in MT are based on empirical approaches to MT that exploit the growing availability of data in electronic format such as corpora of different types including both parallel and comparable corpora.

Empirical approaches are grounded on the belief that the empirical analysis of real texts and their translations makes a significant contribution to solving the problem. There are two types of empirical approach: an approach based on examples and a statistical one.

3.3.1. Example-based systems (EBMT)

Unlike the systems described above, the basic idea developed in this type of system is that source texts and their translations offer a huge database that can be exploited to achieve the translation of new texts. This idea was first expressed by Nagao (1984) in the International NATO Symposium on Artificial and Human Intelligence, who suggested following an "analogy principle" rather than a rule driven approach to MT.

The analogy principle is based on the assumption that the human translation process works by means of the decomposition of a source text into fragmental phrases and by recalling the equivalent translation, rather than on a deep linguistic analysis of the text to be translated. Like the cognitive process of human translators, the EBMT system architecture is therefore based on examples extracted from a database of original texts aligned with their translations (parallel corpora) and on a mechanism that provides translators with the most probable translation of a text string.

EBMT exploits the same type of knowledge as TMs, i.e. a bilingual database, but it differs from TM in that it is not interactive, i.e. it does not allow translators to choose from possible translation suggestions and provides translators with ready translated texts.

In order to use this database, however, a structural analysis stage is required in order to identify translation units in the ST and their equivalent translations in the TT and subsequently align them so that they can be offered to translators as possible solutions for texts that have to be translated from scratch.

The translation process carried out in this type of system is organised as follows:

- Matching stage: analysis of the input text to identify the translation units already present in the bilingual database and show similarities with the ST;
- Alignment stage: the translation units identified in the ST are automatically aligned with translation examples extracted from the bilingual database;
- Recombination stage: the system offers translators the translations identified in its database as possible translations of the translation unit of the ST recombined in order to comply with the TL structure.

3.3.2. Statistical Machine Translation (SMT)

The statistical approach to MT is the dominant methodology in MT nowadays, being used in several very popular MT systems such as *Google Translate*, *Bing* by Microsoft and *Moses*, currently funded in MosesCore, an EU-funded Coordination Action. SMT is focused on a data-driven approach and machine learning techniques which again rely on the use of bilingual corpora of reliable translations, but also on monolingual corpora, with different purposes: it uses bilingual corpora (training corpus) to compute the most probable translation for a given input sentence and monolingual corpora (language model corpus) to assign the proper word order in the target language.

In order for SMT to give reliable results, the system must be trained using large corpora, now more and more freely available on the Internet in many different languages.[37] The systems automatically learn which are the most appropriate

[37] For a list of freely available corpora see http://www.ecpc.uji.es/EN/links.php?language=en

translations by computing their probability of occurrence in the training corpus, which contains a suitable number of aligned source and corresponding target sentences. Therefore, if there is more than one possible translation for an input string, an SMT system will choose the one which is ranked as the most probable translations for it. In pure SMT no linguistic knowledge is used to disambiguate the source text and subsequently to translate it in the target language. Every input sentence is analysed as a sequence of words (*n-grams*) which matches the most probable sequence of words in the target language.

CANDIDE was the first SMT system, developed by Brown (Brown et al., 1988; 1990, 1993) on the basis of the Bayes' law (Bayes & Price, 1763) which is used to calculate the probability that an analysed phenomenon is true or will be true according to a certain set of circumstances.

Given this general description of the SMT approach, there are different types of SMT approaches: word-based, phrase-based and factored-based .

In the word-based SMT model, used for the first time by the IBM CANDIDE project in the late '80s, words are translated as atomic units, i.e. translation is based on word-for-word mapping (*alignment*) between the source and the target text, using the so-called lexical *translation probability distribution* concept, i.e. for each word in the source text, the system computes the number of instances of the equivalent translations in the corpus and estimates a probability distribution by means of the *maximum likelihood estimation*, a method to seek the probability distribution that makes the observed data most likely. This was the first approach adopted in SMT and has now been superseded by other methodologies. Nevertheless, it has developed some basic principles that are still valid for SMT.

In the phrase-based SMT model, the atomic units are represented by phrases. The input text is fragmented into phrases and then translated in the target language using the most likely translation on the basis of a probabilistic weight assigned to co-occurrent translations in the corpus for the same input phrase.

The phrase-based method can be dated back to Och's alignment template model which inspired other scholars including Yamada (Yamada & Knight, 2001) who used phrase translation in a syntax-based model and Marcu (Marcu & Wong, 2002) who introduced a joint-probability model for phrase translation. This method tries to overcome some evident limits of the word-based model, mainly due to the fact that a word cannot be considered the best translation unit. If we look at the English verb *mix up,* for instance, we can easily see that it assumes different meanings in different contexts and therefore needs different translations according to the words and the nature of the words it occurs with:

(1) *try not to* ***mix up*** *all the different problems together*
(2) ***mix up*** *the ingredients in the cookie mix*
(3) *Tom* ***mixes*** *John up with Bill*
(4) *I'm all* ***mixed up***

In (1), it means to change the order or arrangement of a group of things, especially by mistake or in a way that you do not want. In (2), it means to prepare something by combining two or more different substances. In (3), it means to think wrongly that somebody/something is somebody/something else. In (4), it means to be in a state of confusion.

All these different meanings of mix up represented in (1)-(4) correspond, obviously, to different translations in Italian:

*(5) cerca di non **mischiare** i diversi problemi*
*(6) **mescola** gli ingredienti nell' impasto dei biscotti*
*(7) Tom **confonde** John con Bill*
*(8) Sono molto **confuso**.*

On the basis of the different context of use and the co-text, the verb *mix up* can have the following translations in Italian:

mix up(VT) N in = **mescolare** N in
mix up(VT) N with = **confondere** N con
mix up(VT) N(ingredient) = **mescolare** N
mix up(VT) N(medicine)= **preparare** N
mix up(VT) with = **confondere** con
mix up(VT) N(human,info) with = **confondere** N con
mix(VT) up(part) = **confondere**

In SMT, phrases are not considered as a linguistic concept, but as a pure sequence of co-occurrent and contiguous words, as shown in the following example extracted from a phrase table of *Moses*:

" , per la gestione del presente ||| " for the management of this ||| 0.245841 0.000386953 0.245841 0.0788203 2.718 ||| ||| 1 1
" , per la gestione del ||| " for the management of ||| 0.245841 0.000632227 0.245841 0.0841843 2.718 ||| ||| 1 1
" , per la gestione ||| " for the management ||| 0.245841 0.00310736 0.245841 0.143357 2.718 ||| ||| 1 1
" , per la quale sono richiesti ||| " , requiring ||| 0.718868 3.33037e-08 0.718868 0.00289219 2.718 ||| ||| 4 4
" , per la ||| " for the ||| 0.0491683 0.00479878 0.245841 0.210926 2.718 ||| ||| 5 1
" , per ||| " for ||| 0.0491683 0.039521 0.245841 0.339868 2.718 ||| ||| 5 1

The concept of phrase corresponds to a text chunk, i.e. a mere sequence of consecutive words which the equivalent phrase in the target language is assigned to on the basis of a probabilistic computation of the occurrences.

For instance, in the phrase table above, the Italian source phrase "per la gestione del presente" is followed by its translation in English "for the management of this" and finally by the system phrase translation probability distributions φ(f|e) and φ(e|f) and by additional phrase translation scoring functions such as lexical weighting, word penalty, phrase penalty, etc.

Phrases are mainly identified by word alignment in a parallel corpus (Tillmann, 2003; Zhang, Vogel, & Waibel, 2003; Zhao & Vogel, 2005; Zhang & Vogel, 2005; Setiawan, Li, & Zhang, 2005) or directly by sentence-aligned corpora using a probabilistic model (Shin, Han, & Choi, 1996), pattern mining methods (Yamamoto, Kudo, Tsuboi, & Matsumoto, 2003) or matrix factorisation (Goutte, Yamada, & Gaussier, 2004).

In this model, no linguistic information is used, but current trends in SMT are re-considering its use in order to improve the results as will be discussed later on in this chapter.

Nowadays, almost all MT developers are adopting this approach: *Google Translate* and Microsoft *Bing* are well-known examples of this SMT type, to name just a few.

In the factored-based SMT model (Koehn & Hoang, 2007), linguistic information is integrated either in the pre-processing or post-processing phase of SMT in order to improve results and in particular to overcome data sparseness problems caused by limited training data.

The factored-based model foresees linguistic annotation at word level which can carry morphological, semantic or syntactical information. For instance, morphological information can be very useful for translation from or to morphological rich languages. In this way, words become vector of factors that represent different levels of annotation.

The translation process in factored-based SMT implies separate processing of the linguistic information after having translated the lemmas of the input text. In phrase-based factored SMT, phrase decomposition is performed by means of a mapping process which takes place in several steps. For instance, this model is applied in *Moses*, where mapping is performed simultaneously on source and target phrases and it is called a synchronous factored model, as exemplified in the following translation of the one-word phrase *Häuser* into English. The representation of *Häuser* in German is: surface-form *häuser* | lemma *haus* | part-of-speech NN | count plural | case nominative | gender neutral.

The three mapping steps in the morphological analysis and generation model provide the following applicable mappings:

Translation: Mapping lemmas

- *haus -> house, home, building, shell*

Translation: Mapping morphology

- *NN|plural-nominative-neutral -> NN|plural, NN|singular*

Generation: Generating surface forms

- *house|NN|plural -> houses*
- *house|NN|singular -> house*
- *home|NN|plural -> homes*
- ...

These mappings are used to expand the input phrase into a list of translation options which reflect language ambiguities. The German *häuser|haus*|NN|plural-nominative-neutral may be expanded as follows:

Translation: Mapping lemmas

- { ?|house|?|?, ?|home|?|?, ?|building|?|?, ?|shell|?|? }

Translation: Mapping morphology

- { ?|house|NN|plural, ?|home|NN|plural, ?|building|NN|plural, ?|shell|NN|plural, ?|house|NN|singular,... }

Generation: Generating surface forms

- { houses|house|NN|plural, homes|home|NN|plural,
- buildings|building|NN|plural, shells|shell|NN|plural, house|house|NN|singular, ... }

In this model the training parallel corpus is annotated with linguistic information and aligned on the basis of the word surface form or on lemmas or stems, or on any other factor. Therefore, translation and generation tables are extracted from the word-aligned parallel corpus on the basis of scoring methods that help to choose from ambiguous mappings.
Translation phrase mappings are scored on relative counts and word-based translation probabilities whereas generation tables are learned on a word-for-word basis and scored against monolingual data, i.e. the language model.

3.4. Hybrid Machine Translation (HMT)

Another current trend is represented by the integration of traditional linguistic approaches with data driven approaches, taking advantage of the benefits offered by the various technologies, unified in a hybrid MT system.

The advantages of SMT are represented by the possibility of fast development at low costs and, if properly trained with large parallel and monolingual corpora, by a relative fluency of the output. However, (pure) SMT needs large amounts of data which are not always available, especially for under-resourced languages.

The advantages of rule-based MT are that its linguistic resources provide a more precise description of linguistic phenomena and therefore produce less noise than statistical analysis and, furthermore, they can be easily checked, corrected and exploited for other NLP applications such as

electronic dictionaries, text mining and dialog systems. The disadvantages of RBMT are mainly slow development cycles at high costs and lack of robustness if input is incorrect.

There are mainly three types of architectures (Thurmair, 2009): coupling of systems (serial or parallel), architecture adaptations (integrating novel components into SMT or RMT architectures, either by pre/post-editing, or by system core modifications), and genuine hybrid systems, combining components of different paradigms.

In the first type, different MT engines are put together either serially or in parallel, but they are not really merged into one system. A typical example of the serial approach is statistical post-editing (SPE) of the output of an RBMT system, where the best translation is selected on the basis of a bilingual training module. This type of approach outperforms RBMT in MT quality as proven by Schwenk et al. (2009): the output tends to be more grammatical and lexical selection quality is also improved (Dugast, Sellenart, & Koehn, 2007).

In the parallel approach, translations are selected from the outputs produced by different MT systems, used in parallel. The selection is performed in two different ways: either the best translations are extracted from a list of *n*-best translations (Hildebrand & Vogel, 2008) or they are generated on a word or phrase level on the basis of the available MT outputs using confusion networks. This latter approach does not show a significant improvement over the use of one single MT system, either RBMT or SMT (Callison-Burch, Koehn, & Monz, 2009), and furthermore it is difficult to use mainly due to computational resources and availability of MT systems (Thurmair, 2009).

In the second HMT type, the original MT architecture (either knowledge- or data-driven) is adapted or extended by

integrating modifications from different MT approaches. These modification can be implemented either in the pre- or post-processing phase or can regard the core of the MT process. In RBMT, pre-editing by data-driven extensions mainly relate to the automatic update of the lexical base of the MT system by using term-extraction technologies, or the automatic extraction of grammar rules from corpora, both monolingual and bilingual. In particular, automatic extraction of lexical data pertains MWUs, which, since they have an internal linguistic structure, need to be lemmatised and annotated (Dugast, Senellart, & Koehn, 2009; Eisele et al., 2008). Both automatic term-extraction and rule-learning seem not to be particularly beneficial to RBMT since they both lead to significant combinatorial problems and have unexpected side-effects (Thurmair, 2009).

RBMT core extensions deals with the application of probabilistic information to parsing, mainly in the transfer phase, to select the most frequent translation of a given word. Relevant results have been achieved when contextual factors are considered and therefore computed for choosing the best translation candidate (Thurmair, 2009; Kim, Chang, & Zhang, 2002).

In SMT, pre-editing knowledge-driven extensions concern the annotation with linguistic (morphological, semantic, syntactic) information of words or phrases to improve alignment, reduce data sparseness and improve word order. With regard to morphology, POS and lemmatisation tagging are used and for these purposes, also compounding and decompounding techniques whereby complex word sequences are split.

Syntactic information is only used to select the well-formed phrases for the phrase table. SMT core extensions are obtained by the extension of the Phrase Table, rule-based

control of the Language-Model-based generation and factored translation. Extension of the phrase table with linguistic information (terms and phrases) derived by RBMT parsing has been proposed for instance by Eisele et al. (2008). Several proposals have been made for using target grammars in the decoding process, especially in the context of hierarchical translation, to improve the quality of the SMT output.

Finally, factored SMT has already been described in the previous section and is considered to work efficiently for linguistic phenomena such as NP agreement and compounding.

In genuine HMT, the different approaches are combined in a completely new system whose main components are: identification of SL chunks (words, phrases or equivalents thereof), transformation of the chunks into the TL by means of a bilingual resource and generation of a TL sentence.

The different approaches can be combined in different ways: in some systems, such as, for instance, METIS, parsing can be rule-based whereas language generation can be data-driven or, in others, analysis is data driven and generation is rule-based. Other forms of hybrid systems integrate elements of EBMT or TM.

METIS (Vandeghinste et al. 2006) investigates the combination of rule-based and data-driven methods to overcome the problems posed by these two approaches considered singularly: i.e. lack of sophisticated linguistic resources and/or of large parallel corpora. The system uses basic available NLP tools such as taggers, chunkers, lemmatisers, a transfer module based on bilingual dictionaries of single and multi-word terms, consisting basically of lemma and POS in SL and TL and, finally, a generation module with a language model centred on a

tokenised and tagged English corpus (the British National Corpus). This type of system shows similar results to SMT but worse results when compared with a RBMT like *Systran* (Vandeghinste et al., 2008).

Carbonell et al. (2006) propose a data driven approach integrated by a bilingual dictionary and a n-gram indexed target language corpus.

In conclusion, since SMT and RBMT are in some respect complementary, research in this area is attempting to reduce the distance between these two approaches and in the past few years, interest in hybridisation and system combination has significantly increased.

The combination of data-driven approaches and linguistic ones seems to offer a considerable potential to improve MT quality and efficiency and it is to this end that the present dissertation proposes the adoption of a linguistic methodology that identifies and translates MWUs which can be incorporated into a new generation HMT system.

Chapter 4 – Multi-word units

MWUs represent a significant challenge in the field of NLP and their processing has a crucial relevance for many NLP applications as proved by the annual workshops that have been held on this topic since 2001 in conjunction with major NLP conferences.[38] Growing interest by the research community can also be seen in the numerous papers, books and PhD theses devoted to various aspects of MWU processing, in particular in the field of MT,[39] which have been published in recent years.

In this chapter, we provide a broad yet not exhaustive discussion of the foundations, definitions, properties and current research trends in MWU treatment.

MWUs are very numerous and frequent: as noted by Biber et al. (1999), they account for between 30% and 45% of spoken English and 21% of academic prose, Jackendoff (1997) claims that the estimated number of MWUs in a lexicon is equivalent to its number of single words, Gross & Senellart (1998) established that more than 40% of all tokens in a one-year corpus of the French journal *Le monde* belong to MWUs, De Mauro in the GRADIT (1999-2007: XV) stated that out of 360,000 lemmata and sub-lemmata approximately 130,000 are MWUs. More recent theoretical estimations show that specialised lexica may contain

38 http://multi-word.sourceforge.net/PHITE.php?sitesig=CONF

39 A list of papers that analyse MWUs in connection with MT is given in *The Machine Translation Archive* (http://www.mt-archive.info/srch/ling-10.htm) under the following headings: Multiword expressions, Nouns and Noun phrases, Prepositional phrases, Verbs and verb phrases.

between 50% and 70% of this type of lexical unit (Sag et al., 2002). Lately, these estimations were confirmed by Ramisch, Villavicenzio, & Boitet (2010) who found that 56.7% of the terms annotated in the *Genia* corpus consist of two or more words and this is an underestimation since it does not include general-purpose MWUs such as phrasal verbs and fixed expressions.

MWU processing has to cope with many problems due to the peculiar properties of the MWUs which will be described in the next sections. In particular, MWUs have different degrees of compositionality and, in many cases, opaque meanings, i.e., the meaning of the unit is not given by the simple addition of the meanings of the individual constituents that make up the unit. This means that translations of MWUs are very often unpredictable and that a word-for-word translation may result in severe mistranslation.

In addition, their morpho-syntactic properties allow, in some cases, a certain number of formal variations with the possibility of dependencies of elements even when distant from each other in the sentence. Non-compositionality, numerosity and morpho-syntactic variations act as decisive factors in the choice of the effective processing approach of MWU translation.

This chapter presents a general overview of MWUs. Section 4.1. discusses the different definitions of MWUs since there is still no general consensus. Section 4.2. analyses the properties which characterise MWUs. Section 4.3. refers to the different and numerous classifications of MWUs proposed so far. Finally, Section 4.4. analyses the MWU issue in the framework of the Lexicon-Grammar theory.

4.1. Definition

MWUs are lexical elements composed of more than one word which have a particular structural and semantic internal cohesion, act as single lexical units and belong to different lexical categories. MWUs can be verbal structures: *to look at*; nominal structures: *heavy water, arsenic water*; adverbial structures: *as soon as possible*; prepositional structures: *in order to*; proverbs: *walls have ear* and finally conjunctions: *even though.*

They are constructions half way between morphology and syntax since they have a very similar structure to phrases, but present distribution and cohesion characteristics which are very close to words. This term is used to designate linguistic phenomena such as collocations, phrases, idiomatic expressions, proverbs. Jackendoff (1997), for instance, includes memorised poems, familiar phrases from TV commercials such as *to infinity and beyond* (Toy Story) or *to boldly go where no one has gone before* (Star Trek) as MWUs whereas Fillmore (2003) includes grammatical constructions, listable word configurations and frequent sequences as in the word *copy of* in *They gave me a copy of the book.*

MWUs have been an increasingly important concern for natural language processing scholars and are considered a ‘‘pain in the neck for NLP’’ (Sag et al., 2002) because of the many difficulties they raise. To begin with, there is no universally agreed definition or term for the concept of MWU. Concurrent terms of MWU are multi-word, multi-word expression, fixed expression, idiom, compound word and collocation used by many authors from different theoretical schools or following distinct natural language processing approaches.

For Firth (1957), MWUs are habitual recurrent words, combinations of everyday language. Choueka (1998) defines MWUs as sequences of words "whose exact and unambiguous meaning or connotation cannot be derived from the meaning or connotation of its components". For Manning & Schütze (1999), who emphasise conventionality in the use of MWUs, a collocation is an expression consisting of two or more words that correspond to a conventional way of saying things. Wray (1999) underlines how MWUs are any kind of linguistic unit that has been considered formulaic in any research field. Fillmore (2003) describes them as any linguistic expression involving more than one word that requires an interpreter – human or machine – to have more than the abilities of an Innocent Speaker-Hearer (ISH) who has only knowledge of (i) unitary words and (ii) word-to-word relations. For Fillmore, Kay, & O'Connor (1988), MWUs introduce a distinction between what a speaker can compute automatically from language and what he must explicitly know. Calzolari et al. (2002), define these units as "different but related phenomena [. . .]. At the level of greatest generality, all of these phenomena can be described as a sequence of words that acts as a single unit at some level of linguistic analysis".

Sag et al. (2002) propose a formal definition of the term "multi-word" that has been largely adopted by the natural language processing community: "Multi-word expressions are lexical items that can be decomposed into multiple lexemes and display lexical, syntactic, semantic, pragmatic, and/or statistical idiomaticity".

According to this definition, *decomposability* and *idiomaticity* are the basic requirements of MWUs. With regard to the concept of *idiomacity*, it is defined by Kim & Baldwin (2010) as the degree and kind of deviation of the

properties of an MWU from those of its component words which applies at the lexical, syntactic, semantic, pragmatic and/or statistical levels. According to these scholars, l*exical idiomaticity*, refers to the cases when one or more lexical components of an MWE are not part of the lexicon of the language in question. For example, the Latin expressions *ad hoc* and *per se* that are used in standard English, are made up of at least one component that does not belong to the vocabulary of a given language.

Lexical idiomaticity usually implies *semantic idiomaticity* (Kim & Baldwin, 2010), i.e., the meaning of the expression as a whole cannot be directly deduced from the meaning of its components. Semantic idiomaticity is closely related to the figurative use of language when a particular MWU has a metaphoric (like the idiomatic expression *to take the bull by the horns* in English), hyperbolic (as in the English expression *to be not worth the paper it's printed on*) or metonymic meaning (*to lend a hand*).

Syntactic idiomaticity refers to the cases where the syntax of the MWUs is not derived directly from the syntax of its word components (Kim & Baldwin, 2010). For example, the expression *by and large* in English, which in itself works as an adverb, is made up of a conjunction of one preposition (*by*) and one adjective (*large*).

Pragmatic idiomaticity occurs when a given MWU is associated with particular situations and cannot be used or fully understood when uttered out of context (Kim & Baldwin, 2010). Thus, its interpretation is strictly linked to the situational context it appears in. Examples in English include *good morning* and *welcome back* which work as greetings.

Finally, *statistical idiomaticity*, refers to specific combinations of words occurring with notably higher

frequency than alternative phrases of the same expressions (Kim & Baldwin, 2010). For example, the pairs of words *flawless logic* and *spotless condition* are correct and commonly used in English whereas *spotless logic* and *flawless condition*, even if grammatically correct, are not commonly used.

In earlier Lexicon Grammar Theoretical Frameworks, established by Maurice Gross (1975 and 1981), the most essential features of what we call MWU were non-compositionality and semantic opaqueness. Gross (1986) uses the term compound word to refer to a string composed of several words whose meaning cannot be computed from its elements. De Mauro (2000) describes it as a group of words with a single meaning which cannot be inferred from the meanings of the individual words that are part of it, both in the current usage of language and in special languages.

Recently, the significance of compositionality has changed and the term MWU has evolved in such a way that it can also refer to non-idiomatic units, being now used to refer to various types of linguistic entities, including idioms, compounds, phrasal verbs, light or support verb constructions, lexical bundles, etc.

4.2. Properties

MWUs are characterised by a series of properties that assure their semantic and syntactic cohesion. These properties have been discussed by several scholars including Manning & Schütze (1999), Sag et al. (2002), Moszczyński (2007) and Guenthner & Blanco (2004) among others.

These properties can be summarised as follows:

- **Non-substitutability**: one element of the MWU cannot be replaced without a change of meaning or without obtaining a non-sense (*in deep water → in hot water; gas chamber → *gas room*);

- **Non-expandability**: insertion of additional elements is not possible (*get a head start → *get a quick head start*);

- **Non-reducibility**: the elements in the MWU cannot be reduced and pronominalisation of one of the constituents is also not possible (*take advantage → *what did you take? advantage*; *Did you take it?;*

- **Non-translatability**: the meaning cannot be translated literally as is the case for many idioms and proverbs (En. *It's raining cats and dogs* → It. **Sta piovendo cani e gatti*), as well as other types of MWUs (It. *compilare un modulo* → En. **Compile a module*);

- **Invariability**: Invariability can affect both the morphological and the syntactic level. Inflectional variations of the constituents of the MWUs are not always possible. Invariability affects both the head elements and its modifiers (*fish out of water → *fishes out of water*; *dead on arrival → *dead on arrivals*; *in high places → *in high place*); syntactical variations inside an MWU may also not be acceptable (*credit card → *card of credit)*;

- **Non-displaceability**: displacement and a different order of constituents are not possible (*wild card → *is wild this card?*) - (*back and forth → *forth and back*);

- **Institutionalisation of use**: certain word units, even those that are semantically and distributionally "free", are used in a conventional manner. The Italian expression *in tempo reale* (a loan translation of the English expression *in real time*) is an example of this feature since its antonym **in tempo irreale (*in unreal time*) seems to be unmotivated and not used at all.

These features are not always present at the same time in an MWU since different MWU types may present different characteristics according to their degree of variability of co-occurrence. For instance, in proverbs and idioms, i.e. MWUs without any variability of co-occurrence among words, almost all features are present since we cannot replace any element in this type of MWU with a synonym, reduce, expand or displace it, they are invariable and what is more, their literal translations originate odd meanings in the target language as we have seen in the English idiomatic expression *it's raining cats and dogs*.

Prepositional constructions such as *per effetto di* (*under*), *in preda a* (*in the grip of*), or conjunctions such as *in modo che* (*so that*), *al fine i* (*in order to*), which are also MWUs without any variability of co-occurrence among words, share the same features listed above as proverbs and idioms.

The same applies to compound words with non-compositional meanings, which are MWUs without any variability of co-occurrence among words like the Italian *berretto verde* (officers of the Guardia di Finanza, an Italian police force under the authority of the Minister of Economy and Finance), *teste di cuoio* (members of a special anti-terrorist police team), *casa chiusa* (brothel), with the exception of morphological invariability since they allow for

inflections (e.g.: *berretti verdi*, *case chiuse*).

On the contrary, compound nouns with a compositional meaning which are MWUs with no or almost no variability of co-occurrence among words, i.e. combinations with fixed internal distribution, tend to be less frozen from a morphological point of view like the Italian *Stato membro* (*member state*) which can be inflected, i.e. *Stati membri*, but also from a semantic point of view since in some cases, as in the nominal construction we are analysing, one of the elements can be replaced by a synonym (*paese membro*) or the Italian compound noun *carta di credito* (*credit card*), for which expansions to specify different types of credit cards are possible, i.e. *carta di credito prepagata* (*prepaid credit card*), *carta di credito rateale o rotativo* (*revolving credit card*), etc.

Verbal constructions tend in general to be more variable, both on a morphological and syntactic level, but here too, the presence of different features depends on the internal cohesion of the MWU. For instance, in the Italian verbal construct *tirare le cuoia*, we cannot replace one of the elements, expand, reduce or displace it, nor translate it literally (the English equivalent is *Kick the bucket*), but morphological variations are allowed: *tirò le cuoia*. The same situation does not apply to the Italian verbal construct *fare luce* (→ En. *shine a light on*), where *luce* can be replaced by *chiarezza* (and takes on the meaning of En. *to clarify*) without a significant change in meaning and some expansions are possible, *fare un po' di luce* (→En: *to shine a little light on).*

MWUs with a high or limited degree of variability of co-occurrence among words and with a limited degree of variability may show only some of the features listed above.

As a consequence of the presence/absence of these

various features, these particular lexical constructions are quite difficult to identify and classify. A further element of complexity in this sense, especially in view of a computational disambiguation and translation of MWUs, is represented by the fact that some MWUs with no or almost no variability of co-occurrence among words can also be used as MWUs with a high degree of variability of co-occurrence among words: for instance, in Italian, the MWU with limited variability of co-occurrence among words and non-compositional meaning *appendere al chiodo* as in the sentence *Il calciatore ha appeso le scarpette al chiodo* means in English *to retire* (En. *The football player retired*) but it has also a compositional meaning, if it is not used in a figurative way as in the sentence: *Ha appeso il quadro al chiodo/alla parete/al muro/...* → En. *He hung the painting on the hook/the wall/...* As it is clear from this simple example, the different use of this expression (literal vs idiomatic or compositional vs non-compositional) has relevant consequences on the choice of the correct translation equivalent as well (*hang up* vs *retire*).

In some cases, the context and the co-text can help to identify the correct meaning, for instance the Italian compound noun *tiro a segno* can be translated in English as *shooting gallery* (It. *Sono stato al tiro a segno* → En. *I was at the shooting gallery;* It. *Ho sparato nel tiro a segno* → En. *I shot in the shooting gallery*) whereas in the Italian sentence *Ho mandato anche questo tiro a segno,* it takes on a compositional meaning. In the first sentence the co-occurrence of a locative preposition before *tiro a segno* may help to disambiguate it as a compound noun whereas in the second sentence, the co-occurrence of the verb *mandare* may help to identify the expression *mandare il tiro a segno* which means *to accomplish the shot.*

4.3. Classification of multi-word units

Classification of MWUs poses various problems since it can be approached from different points of view. Here too, no consensus has been reached with regard to a unique MWU taxonomy.

A formal linguistic classification was set up by Fillmore, Kay and O'Connor (1988) in the framework of a generative English grammar. They divided multi-word lexemes into four set of categories characterised by a binary opposition of properties:

- *encoding* vs. *decoding* MWUs: this opposition is based on the possibility that native speakers of a given language understand an unknown lexeme with complete confidence on the basis of prior experience (encoding lexeme) or not (decoding lexeme);
- *grammatical* vs. *extra-grammatical* MWUs: grammatical MWUs follow the grammatical rules (*spill the beans*) of a given language whereas extra-grammatical MWUs violate them (e.g. *by and large* or *at hand*);
- *substantive* vs *formal* MWUs: in a substantive or lexically filled MWU all elements are fixed whereas formal or lexically open MWUs are determined by a fixed structure which can be filled by the usual range of words appropriate to that structure (e.g., *the more ..., the X-er* as in *The more you practice, the easier it will get*);
- MWUs *with vs without pragmatic point:* MWUs with a pragmatic point are used in specific pragmatic contexts such as the formulaic expression *Good*

morning. Many other idioms do not have a specific pragmatic context such as the adverbial expression *all of a sudden*.

On the basis of the familiarity, i.e. the predictability of an MWU both with respect to standard syntactic and semantic compositionality, Fillmore et al. (1988) classify MWUs according to three categories:

- *Unfamiliar pieces unfamiliarly combined*: this class contains idiomatic MWUs which are idiosyncratic both from a semantic and syntactic point of view to such an extent that it may include, for instance, words that appear in a specific idiom (*ad hoc*, *with might and main*) or very specialised syntactic configurations that do not occur anywhere else in language (*the more*, *the merrier* and more generally, expressions of the type *the X-er*, *the Y-er*).
- *Familiar pieces unfamiliarly combined*: these syntactically and semantically idiosyncratic MWUs require rules for their interpretation even if the lexical elements of the multi-word are familiar ones. Examples are *all of a sudden*, *stay at home* and constructions of the type *first cousin twice removed*.
- *Familiar pieces familiarly combined*: MWUs are not idiosyncratic on the lexical, semantic and syntactic level. However, they have an idiomatic meaning as in *pull someone's leg* and *tickle the ivories*.

A different and quite complex MWU classification was presented by Brundage et al. (1992) and is based on a study

of approximately 300 English and German MWUs which were classified on the basis of their syntactic structure and the transformations they can undergo.

Sag et al. (2002) propose a classification based on a semantic and syntactic variability degree and identify two broad categories: *Lexicalised phrases* and *Institutional phrases*.

- *Lexicalised phrases* have at least partially idiosyncratic syntax or semantics, or contain words which do not occur in isolation; they can be further broken down into *fixed expressions*, *semi-fixed expressions* and *syntactically-flexible* expressions. *Fixed* expressions are fully lexicalised and can neither be varied morphosyntactically nor modified internally. Examples of fixed expressions are: *in short*, *by and large*, *every which way*. *Semi-fixed expressions* are invariable concerning word order and composition, but they can undergo some morphological and syntactical variation such as inflection, variation in reflexive form and determiner selection. Non-compositional idioms (*kick the bucket*), compound nouns (*car park*) and proper names (*the San Francisco 49ers*) belong to this category. *Syntactically-flexible expressions* are syntactically variable and occur in the form of non-compositional idioms (*sweep under the rug*), verb-particle constructions (*mix up*) and light verbs (*make a mistake*).

- *Institutionalised phrases* are syntactically and semantically compositional, but occur with markedly high frequency (in a given context), for example, *salt and pepper, traffic light.*

MWUs can also be classified in terms of compositionality, as proposed by Kim & Baldwin (2010) who identify two major classes: *compositional* and *non-compositional* MWUs:

- *Compositional MWUs* are lexical units whose meanings are directly related and predictable from the meanings of their component words. *Collocations* are a particularly important sub-class of compositional MWUs, given that their use is very widespread and that they must be mastered by second language learners in order to achieve fluency in their target language.
- *Non-compositional MWUs*, also known as *idioms*, are on the other hand, lexical units whose meanings cannot be deduced from the meanings of their component words.

A further possibility of classification is linked to the internal structure of MWUs as proposed by Dias et al. (1999) i.e.: *contiguous*, *non-contiguous* and *flexible multi-word lexical units*:

- contiguous multi-word lexical units are uninterrupted sequences of words such as single market or official languages.
- non-contiguous multi-word lexical units consist of fixed sequences of words interrupted by one or several gaps filled in by interchangeable words. For instance, the _____ European Council is a non-contiguous multi-word lexical unit where the gap is likely to be filled in by names such as Lisbon or Luxembourg.

- flexible multi-word lexical units correspond to free sequences of words. For example, to be responsible for is a flexible multi-word lexical unit since it can be found in text in the form to be successfully responsible for or to be for a long time responsible for.

In the theoretical framework of the Meaning-Text theory (MTT), Mel'čuk, Clas, Polguère (1995) suggested a classification expressed in terms of semantic compositionality to which the following classes belong: *complete phrasemes, semi-phrasemes* and *quasi-phrasemes*.

- *Complete phraseme*: fully non-compositional MWUs whose meaning cannot be deduced by the composition of the meanings of the constituent words in the unit such as *kick the bucket* and *Achilles' hill.*
- *Semi-phraseme*: partially compositional MWUs in which the overall meaning of the unit is based on the meaning of at least one of the constituent words and is not the result of the composition of the meanings of the different elements of the unit. Examples include collocations with support verbs such as *to do a favour*, intensifiers like *heavy smoker* and causative verbs such as *to get in a panic* among other types of collocations.
- *Quasi-phrasemes*: MWUs in which all the words keep their original meanings but an extra element of meaning is included due to the co-occurrence of the constituent elements in these units. Examples are *bacon and eggs* →dish consisting of raw eggs fried in a particular manner, and fried slices of bacon, or

shopping centre –> group of various types of shops built as a whole in a separate area, thus constituting a centre for shopping.

4.4. Lexicon-Grammar and multi-word units

Lexicon-Grammar (LG) is the linguistic formal analysis framework developed by Maurice Gross (1968, 1989), a French linguist who devoted his research work to the description of idiosyncratic properties of lexical elements in the late 60s. LG develops its theoretical foundations on specific mathematical models of language (Harris, 1982; Gross, Halle, & Schutzenberger 1973) and its main goal is to describe syntax by formalising all mechanisms of word combinations.

The basic concept of LG is that the lexicon cannot be separated from syntax, i.e. since any lexical element is part of a simple sentence, it takes with it a part of grammar. The grammatical properties of lexical elements are inalienable and combined with the grammatical properties of other lexical elements on the basis of co-occurrence and selection restriction rules. The analysis of word co-occurrence, distribution and selection restriction observed in simple sentences[40] by means of predicates syntactic-semantic properties represents the core of LG methodology.

Unlike other well-known formal analysis of natural languages and in particular Chomsky's transformational

40 In LG, simple sentences are defined as the minimal linguistic meaning contexts in which co-occurrence, selection restriction and distribution can be analysed. More specifically, a simple sentence is a context formed by a unique predicative element (a verb, but also a noun or an adjective) and all the necessary arguments selected by the same predicate in order to obtain an acceptable, grammatical sentence. For more information on simple sentence definition, see Gross (1968).

grammar (Chomsky, 1957; 1965)where language description is mainly grounded on an analysis of the systematic relations between syntactic structures, in the LG approach the formal description of natural language is deeply rooted in the empirical examination of the lexicon and the combinatory behaviours of lexical elements, encompassing both syntax and lexicon.

In the wake of the research work of Maurice Gross for the French language, the LG analytical method, based on Zellig Harris' concepts of Operator-Argument Grammar (Harris, 1982), and transformational rules (Harris, 1964), has produced empirical and exhaustive linguistic descriptions by means of large data sets consisting of tables of syntactic-semantic properties of thousands of lexical entries (mainly verbs, nouns and adjectives) for many languages (French, Italian, Portuguese, Spanish, English, German, Norwegian, Polish, Czech, Russian, Bulgarian, Greek, Arabic, Korean, Malagasy, Chinese, Thai).

LG scholars have been studying MWUs for years now and also in this case LG research is indebted to the structuralist approach of Harris (1946 and 1970), who analysed the combination of morphemes in more complex linguistic units. In his work *From Morpheme to Utterances* (Harris, 1946), he mentions the concept of free sequences of simple words with a unique overall meaning for the first time in contemporary linguistics and identifies morpheme distributional classes according to which words and sequences of simple words are classified. In this respect, simple words and sequences of simple words are analysed using the same methodology. Sequences of morpheme classes which are found to be substitutable in virtually all environments or some single morpheme classes will be equated to that morpheme class: AN=N means that "good boy", for example can be substituted for "man" anywhere.

Another seminal concept developed by Harris is the *co-occurrence likelihood* (Harris, 1968), i.e. some words are more

likely to occur together and their meanings are determined to a large extent by their collocational patterns.

The transformational and distributional concepts developed by Harris represent the pillars of LG theoretical reflections. Gross adopts and further develops both concepts of linguistic transformation and simple sentence in the framework of a formal grammar of natural languages. Furthermore, the LG analysis encompasses all the different types of MWUs.

D'Agostino & Elia (1998), Italian heirs of the theory developed by the French linguist, consider MWUs part of a continuum in which combinations can vary from a high degree of variability of co-occurrence of words (combinations with free distribution), to the absence of variability of co-occurrence. They identify four different types of combinations of phrases or sentences, namely (i) with a high degree of variability of co-occurrence among words, i.e. combinations with free internal distribution, compositional and denotative meaning such as in *dirty water*, *or clean water*; (ii) with a limited degree of variability of co-occurrence among words, i.e. combinations with restricted internal distribution such as in *natural water*, or *mineral water*; (iii) with no or almost no variability of co-occurrence among words, i.e. combinations with fixed internal distribution such as in *heavy water*; and (iv) with no variability of co-occurrence among words, i.e. proverbs such as *all good things come to he who waits*.

The several degrees of variability or invariability can be seen in compounds, as in the illustrated *water* compounds, but also in other types of MWUs. As demonstrated in (Barreiro, 2008), MWUs have been divided into three main categories: lexical units (with all the compounds), frozen and semi-frozen expressions (including phrasal verbs (*show up*), support verb constructions (*give a (big) hug to*) and proverbs) and lexical bundles (*I think that*; *Would you mind if*). Descriptions and examples of all the different types of MWUs can be found in the same work. Some MWUs do not fit into any of these three

major types.

Each type of MWU may need to follow a different formalisation method. There is the morphological aspect of MWUs (i.e., the morphology of composition) that weights considerably for morphologically-rich languages and remains a highly challenging task. From a lexicographical point of view, MWUs with a specific grammatical function and an autonomous meaning should be registered in dictionaries in a systematic way, i.e. as autonomous lemmata and not, as often is the case in traditional dictionaries, as examples of use of head nouns or adjectives.

As far as lemmatisation is concerned, a clear distinction between MWUs with a high degree of variability of co-occurrence among words and those with a limited or no variability of co-occurrence among words (compound words, idiomatic expressions, proverbs) should be made.

This is one of the most critical issues in the description of natural languages. For example, there is a significant difference in Italian between *colletto bianco* (with the meanings of "white collar" and "white collar worker") and *colletto rosso* ("red collar"). The first has to be lemmatised since it has the specific meaning of "employee" with distinctive morpho-grammatical and lexical properties, i.e. (i) it is singular masculine compound word with the meaning of "human being", with *colletti bianchi*, as its masculine inflected form; (ii) it does not allow for expansions since it does not accept any insertion of additional words, like for instance **colletto molto bianco* (**very white collar worker*).

On the contrary, *colletto rosso* does not have these characteristics, being a free nominal group, therefore not necessarily lemmatisable. This is quite a simple example of the difference between opposite poles in the continuum.

Sometimes, however, MWUs are much more difficult to classify and describe. For example, the Italian MWU *editto bulgaro* (Bulgarian edict), taken from the political language and

referring to a decision by the Italian prime minister Berlusconi in 2002 about some journalists and their banishment from the Italian Broadcasting Service, and *elezione bulgara* (Bulgarian elections) verge between the status of compound words and that of free nominal groups. This is a problem that occurs most frequently with compound words.

Another important level of analysis of MWUs concerns their morpho-syntactic classification which can be performed inside simple sentences and on a distributional basis. For example, compound words can be identified and therefore lemmatised also on the basis of their morpho-syntactic properties.

Lemmatisation of MWUs that belong to classes with limited or no variation of distribution (semi-frozen or frozen expressions) such as technical MWUs, idioms and proverbs, has important consequences in NLP, text automatic analysis, terminology, the structure of the semantic web and computer-aided translation.

In particular, the correct identification of MWUs has important effects on the quality of translations as we already discussed in the previous sections. For example, the famous English idiom: *It's raining cats and dogs*, cannot be translated literally into Italian as *Sta piovendo cani e gatti*.

Adaptation of the concept to the Italian language is required so that the expression *Sta piovendo a catinelle* (literally: *It's raining from jars*) is understood as *it's raining very hard*. The same property can be applied to other types of MWUs. For example, the English literal translation of the Italian verbal expression *compilare un modulo* (*compile a module*) does not convey the correct meaning. The correct translation is to *fill in a form*.

The main linguistic resources developed by LG researchers concerning MWUs are:

- *matrix tables* describing the syntactic-semantic properties of predicates;

- morphologically and semantically tagged *electronic dictionaries*;
- *local grammars* in the form of Finite State Automata (FSA)[41] and Finite State Transitions (FST)[42].

LG *matrix tables* describe the syntactic properties of predicates: each row corresponds to a predicate and each column represents a formal property. Rows may describe both distributional and transformational properties, using the sign "+" or "-" the presence of which means that the predicate can or cannot accept a specific property, respectively.

With regard to MWUs, matrix tables have been developed, for instance, by the Italian LG research group with reference to Support Verb constructions and Idiomatic expressions.

Table 1 illustrates an example of a matrix table for the Support verb structures with *essere* followed by a frozen or semi-frozen prepositional group such as *essere in ansia*, *essere in ballo* (Vietri 1996) which have been classified in thirteen LG matrix tables according to the number of arguments and the internal structure.

---

[41] **Finite-State Automata (FSA)** are a special case of Finite-State Transducers that do not produce any result (i.e. they have no output). NooJ's users typically use FSA to locate morpho-syntactic patterns in corpora and extract the matching sequences to build indices, concordances, etc.

[42] **Finite-State Transducers (FSTs)** are graphs that represent a set of text sequences and then associate each recognized sequence with an analysis result. The text sequences are described in the input part of the FST; the corresponding results are described in the output part of the FST. Typically, a syntactic FST represents word sequences and then produces linguistic information (its phrasal structure, for example).

N0=:umano	N0:=-umano	N0=Che F		Prep		Vsup=:stare	Vsup=:rest.,rim.	Vsup=:diventare	Vsup=:vivere	Vsup=:andare	Vsup=:entrare	Vsup=:mandare	Vsup=:metter	Vsup=:rendere	Vsup=:ridurre
-	+	-	essere	in	abolizione	+	+	-	-	-	-	-	-	-	-
-	+	-	essere	in	abrogazione	+	+	-	-	-	-	-	-	-	-
+	-	-	essere	In	allerta	+	+	-	+	+	-	+	+	-	-
-	+	-	essere	in	allestimento	+	+	-	-	+	-	-	+	-	-
+	-	-	essere	in	azione	+	+	-	-	+	+	+	+	-	-
+	+	-	essere	in	ballo	+	+	-	-	-	+	-	+	-	-
+	-	-	essere	in	ballottaggio	+	+	-	-	+	+	+	+	-	-
+	-	-	essere	in	calore	+	+	-	-	+	+	+	-	-	-
+	-	-	essere	in	castigo	+	+	-	+	+	-	+	+	-	-

Table 1 - Example of LG matrix table for the Vsup *essere* (Vietri 2008:59)

LG *electronic dictionaries* are part of the DELA[43] system, a homogeneously structured lexical database in which the morphogrammatical characteristics of lexical entries (gender, number and inflection) are formalised by means of distinctive, non-ambiguous alphanumeric tags. This system consists of Simple-Word Electronic Dictionaries (DELAS-DELAF) and Compound-Word Electronic Dictionaries (DELAC-DELACF) which include lexical meaning units such as *nursing home*, and *rocking chair*, i.e. MWUs composed of two or more simple words and characterised by a global meaning which may also be non-compositional.

Each entry in the dictionaries is given a consistent ontological description, being coherently tagged with reference to the knowledge domain(s) in which it is commonly used (i.e., in which it has a terminological unambiguous meaning). For instance, the Italian compound word *acconto dividendo* (→ En. *interim dividend*) is marked

43 Acronym from Dictionnaire Électronique of LADL (Laboratoire d'Automatique Documentaire et Linguistique).

with the tag ECON which stands for Economics. As a further example, the Italian compound *massimizzazione del gettito fiscale* (→ En. *revenue maximisation*) is marked with two different tags: ECON and FISC (Tax Regulations), due to the fact that it is used in both knowledge domains.

The development and management of an electronic dictionary consist of three main steps:

- *Lexical acquisition.* During this ongoing phase, MWUs are extracted from corpora and/or certified glossaries and continuously updated.
- *Morpho-grammatical, syntactic and domain tagging*. Each lexical entry is given a coherent linguistic description consisting of (i) a morpho-grammatical and inflectional paradigm, (ii) the internal structure of the compound word, (iii) the domain. The same information is given to the corresponding translation, together with the syntactic function of the terminological compound word (both in the source and the target language). In the following entry extracted from the English-Italian bilingual dictionary

  macchia/bianca,.NA:fs-+;MED/ =white/spot,.AN:s+/N

  the Italian compound noun *macchia bianca* is followed by the tag "NA:fs-+" which indicates the morphologic and grammatical pattern of the compound noun, i.e., the compound consists of a noun (N) followed by an adjective (A), it is feminine singular (fs), it does not have a masculine form (-) but a feminine plural form (+); the tag "MED" (for Medicine) refers to the domain that the entry belongs

to. The English translation *white spot* which follows after the equal sign is given the same consistent ontological description. Finally, at the end of the string, the tag "N" indicates the syntactic function of the compound noun, both in Italian and in English.
Examples of different possible morpho-syntactic subcategories are provided in Table 2.

N° of constituents in the lexical unit	POS tags	Example
bi-gram	NA NN …	aborto spontaneo (MED) interfaccia utente (INF) …
tri-gram	NPN NPN NPN …	capacità del disco (INF) cassa di risparmio (ECON) morbo di Crohn (MED) …
fourth-gram	NAPN …	disturbo respiratorio del sonno (MED) …
fifth-gram	NPNPN …	disturbo da deficit di attenzione (MED) …

Table 2 - Morpho-syntactic subcategories of MWUs

- *Testing on corpora*. The dictionary is used to automatically analyse and process large corpora.

As a sample, we provide a short excerpt from the Italian Electronic Dictionary of Medicine:

macchia/bianca,.NA:fs-+;MED/ =white/spot,.AN:s+/N
macchia/blu,.NA:fs-+;MED/ =blue/spot,.AN:s+/N
macchia/blu,.NA:fs-+;MED/ =macula/cerulea,.NA:s+/N
macchia/corneale,.NA:fs-+;MED/ =aglia,.N:s+/N
macchia/cribrosa,.NA:fs-+;MED/ =lamina/cribrosa,.NA:s+/N
macchia/di/Bier,.NPN:fs-+;MED/ =Bier/spots,.NN:s+/N

macchia/di/Bitot,.NPN:fs-+;MED/ =Bitot's/spots,.NPN:s+/N
macchia/di/Brushfield,.NPN:fs-+;MED/
=Brushfield's/spots,.NPN:s+/N
macchia/di/De Morgan,.NPN:fs-+;MED/ =De
Morgan's/spot,.NPN:s+/N
macchia/di/Filatov,.NPN:fs-+;MED/ =Filatov's/spots,.NPN:s+/N
macchia/di/Flindt,.NPN:fs-+;MED/ =Flindt's/spots,.NPN:s+/N
macchia/di/Koplik,.NPN:fs-+;MED/ =Koplik's/sign,.NPN:s+/N
macchia/di/Koplik,.NPN:fs-+;MED/ =Koplik's/spots,.NPN:s+/N
macchia/di/Maurer,.NPN:fs-+;MED/ =Maurer's/cleft,.NPN:s+/N
macchia/di/Maurer,.NPN:fs-+;MED/ =Maurer's/doft,.NPN:s+/N
macchia/di/Michel,.NPN:fs-+;MED/ =Michel's/flecks,.NPN:s+/N
macchia/di/Michel,.NPN:fs-+;MED/ =Michel's/spots,.NPN:s+/N
macchia/di/Mueller,.NPN:fs-+;MED/ =Mueller's/spots,.NPN:s+/N
macchia/di/Mueller,.NPN:fs-+;MED/ =vitiligo/iridis,.NN:s+/N
macchia/di/Roth,.NPN:fs-+;MED/ =Roth's/spots,.NPN:s+/N
macchia/di/Soemmering,.NPN:fs-+;MED/
=Soemmering's/spot,.NPN:s+/N

At present, 180 different domain tags are included in the DELAC/DELAF data-base The most important dictionaries are: Computing/IT (approx. 54,000 entries), Medicine (approx. 46,000 entries), Law (approx. 21,000 entries) and Engineering (approx. 19,000 entries). Subset tags are also provided for domains that include specific subsectors. This is the case for Engineering for which a generic tag ING is used whereas nine more explicit tags are used for Acoustic Engineering (ING ACUS), Aeronautics and Aerospace Engineering (ING AER), Chemical Engineering (ING CHIM), Civil Engineering (ING CIV), Mechanical Engineering (ING MECC), Mining Engineering (ING MIN), Naval Engineering (ING NAV), Nuclear Engineering (ING NUCL) and Oil Engineering (ING PETROL). The same formalisation method has been used for Physics which has been given a generic tag FIS plus more specific tags for Atomic Physics (FIS ATOM), Nuclear Physics (FIS NUCL), Physics of Plasma (FIS PLASMA), Solid-State Physics (FIS SOL) and Subnuclear Physics (FIS SUBNUCL).

Each dictionary has been created and verified under the

supervision of domain experts. In Monti et al. (2011), we illustrated how these MWU dictionaries are particularly relevant in all phases of the translation process (from the analysis phase to the revision phase) and how they can be used in applications not typically related to the translation process such as text mining and information retrieval, which, if integrated into translation workspaces, help to improve the documentary competence of translators in order to process unstructured (textual) information and make the information on the web or in texts accessible to translators.

Finally, *local grammars* are grammars that only account for certain grammatical features in a given language; they are used to parse texts on the basis of the syntactic information they describe and essentially encompass transformational rules and distributional behaviours (Harris, 1957). Local grammars are constructed in the form of FSA/FST[44], i.e. either deterministic or non-deterministic oriented graphs in which specific formalisms are used to first recognise and subsequently disambiguate, tag and rewrite sets of text sequences. FSTs/FSAs are useful for automatically recognising and parsing any kind of text. A detailed description of these types of grammars in connection with MWU processing will be given in Section 6.2.2.

In the framework of the LG approach, Salkoff specifically addresses the translation problems related to different types of MWUs in a contrastive French-English grammar (1999) and subsequently in an unpublished work about MT, i.e. *Loquatur!* (forthcoming). In this latter work he adopts a rule-based MT approach in which rules are written in the form of a string

[44] An FST has an input part in which the text sequences to process and an output part in which processing results are given are included. On the contrary, an FSA can be defined as a special case of FST that does not produce any result (i.e. it has no output) (Silberztein, 1993 and 2002). FSAs are typically used to locate morph-syntactic patterns in corpora; they can also extract matching sequences in order to construct indices, concordances, etc.

grammar, as defined by Harris (1962). The RBMT system developed by Salkoff relies on two linguistic modules:

- a French string grammar which contains a list of structures of the French language;
- a translation module which contains the list of comparative schema in English.

Salkoff analyses the translation problems concerning different types of MWUs and suggests that different types of MWUs should undergo different treatments. In particular, he analyses support verb constructions and frozen expressions (frozen prepositions, frozen adverbs and compound nouns). With regard to support verbs, he suggests that these expressions should be treated both in the lexicon and the grammar, i.e. the relationship between the support Verb *Vsup* and the supported noun, *Npred*, must appear in the lexicon and an appropriate rule should be included. With reference to frozen expressions, i.e. idioms, he distinguishes between completely frozen expressions (*take the bull by the horns*, *kick the bucket, ...*) and partially frozen expressions (*pull the wool over N1poss eyes*, *wear one's heart on one's sleeve*, …) and suggests the following processing methodology:

- all the frozen expressions should be listed in a lexicon in such a way that all words in the idioms are listed in the entry together with their function in the string containing the idiom and the equivalent translation of the idiom;
- on the basis of the lexicon, a pre-processor scans the texts to recognise words and particular sequences of words (idioms) and fixed combinations (compound

nouns) and delivers the lexical entries detected to the parser;

- the parser places the words in the idioms in a specific syntactic context.

In his view, totally frozen expressions should be treated as unique sequences that have a single lexical entry as in the French conjunctions *afin de (*En. *in order to)* and *afin que (*En. *in order that)*, which can each be given a single lexical entry.

On the contrary, for semi-frozen expressions, the different parts should be separated by an intercalated adjunct since there are ambiguous sequences which can be analysed either as non-compositional expressions belonging to a unique grammatical category or as compositional expressions, made up of individual words carrying their own meaning. The example proposed by Salkoff is the French expression *au moins*, which is either an idiomatic adverb (*in order that*) or a sequence of words *à le moins* (*to the least*).

A more recent study on MT processing of MWUs has been proposed by Váradi (2006) who focuses on a specific typology of MWU which are partially fixed and partially productive. The experiment carried out by Váradi for the Hungarian language is based on the use of local grammars to capture the productive regularity of MWUs and its outcome is uniform processing implementation in the NooJ tool, which will be illustrated in more detail in Section 6.2.1. Based on the assumption that MWUs are particularly frequent when viewed in a multilingual setting, the Hungarian scholar analyses common phrases such as *a twenty year old woman*, which generally is not viewed as an MWU until one analyses the syntactic/semantic and translational constraints involved in its structure (e.g. *year old woman).

His contribution highlights that the use of local grammars in a multilingual setting can provide the flexibility required to cover the phenomena of partially productive MWUs which

form a continuum between frozen MWUs and open-ended productive phrases defined by syntactic rules sensitive to part of speech categories only.

Chapter 5 - Multi-word unit processing in MT

The importance of the correct processing of MWUs in MT and computer-aided translation has been stressed by several authors.

Thurmair (2004) underlines how translating MWUs word-by-word destroys their original meanings. Villavicenzio et al. (2005) underline how MT systems must recognise MWUs in order to preserve meaning and produce accurate translations. Váradi (2006) highlights how MWUs significantly contribute to the robustness of MT systems since they reduce ambiguity in word-for-word MT matching and proposes the use of local grammars to capture the productive regularity of MWUs. Hurskainen (2008) states that the main translation problems in MT are linked to MWUs. Rayson et al. (2010) underline the need for a deeper understanding of the structural and semantic properties of MWUs in order to develop more efficient algorithms.

Different solutions have been proposed in order to guarantee proper handling of MWUs in an MT process. Diaconescu (2004) stresses the difficulties of MWU processing in MT and proposes a method based on Generative Dependency Grammars with features. Lambert & Banchs (2006) suggest a strategy for identifying and using MWUs in SMT, based on grouping bilingual MWUs before performing statistical alignment. Barreiro (2008) describes where and why MT engines are unsuccessful at handling the translation of support verb constructions and finds a method based on paraphrases to overcome the machine's inability to

translate them. Moszczyński (2010) explores the potential benefits of creating specialised MWU lexica for translation and localisation applications.

The most critical problem in MWU processing is that the MWUs often have unpredictable, non-literal translations; they are numerous and not all included in dictionaries; they may have different degrees of compositionality (from free combinations to frozen MWUs, as in the English noun phrase *round table*) and their morpho-syntactic properties allow, in some cases, a certain number of formal variations with the possibility of dependencies of elements even when distant from each other in the sentence.

These problems result in mistranslations by MT systems since not all approaches are capable of processing them correctly. In addition, they can have an opaque meaning, i.e., the meaning of the unit is not given by the meaning of the individual constituents that make up the unit and a literal translation is often not understandable and incorrect.

The problem of MWU processing and translation in MT has been discussed from several viewpoints according to the different MT modelling approaches, i.e. rule-based MT, example-based MT or statistical MT. The aim of this chapter is therefore to present an overview of the state-of-the art of the various MT approaches to MWU processing, focusing on the identification task. State-of-the-art MWU processing techniques represent the starting point for the methodology proposed in Chapter 6. Information in this chapter allows this work to be contextualised within the MT community.

Section 5.1 describes MWU processing in RBMT and in particular a specific approach in the framework of the various linguistic approaches to MWU processing represented by the so-called SEMTAB rules in the *Openlogos* MT system, an open-source version of the former

famous commercial *Logos* MT system. The next sections describe the empirical approach to MWUs and recent experiments conducted by different scholars in the fields of EBMT and SMT.

5.1. Multi-word unit processing in RBMT

In RBMT, the identification of MWUs is mainly based on two different approaches: a lexical approach and a compositional approach. In the lexical approach, MWUs are considered as single lemmata and lemmatised as such in the system dictionaries. In the compositional approach, MWU processing is obtained by means of tagging and syntactic analysis of the different components of an MWU.

One of the most interesting processing approaches to MWU in RBMT is performed by the former Logos system, now *Openlogos*. *Logos* was one of the first commercial general purpose fully automatic MT systems , based on the transfer approach.

The MT system is based on SAL (Semantico-syntactic Abstraction language), an abstract hierarchical tree structure language in which the system translates every natural language string before parsing. It is grounded on the scientific belief that the syntactic structure on which a RBMT system is based should be essentially merged with the semantic structure. In other words, semantic information is available at every point of the process to help resolve ambiguities at every linguistic level (lexical, syntactic or semantic).

The key element of the SAL lies in the semantico-syntactic description of the verbs which are the main means for the production and comprehension of natural languages.

The *Logos* model is set up on different phases through which natural language ambiguities can be simplified and reduced in an incremental way. At the end of the process, an abstract, formal and semantico-syntactic SAL representation of the source language is obtained which is subsequently translated into the target language. The main linguistic knowledge bases of the *Openlogos* system are dictionaries, syntactic rules (analysis, transfer and generation) and SEMTAB rules. The SEMTAB rules have an important role in the processing of MWUs with a limited degree of variability of co-occurrence among words (Scott, 2003; Scott and Barreiro, 2009; and Barreiro et al., 2011) since they analyse, formalise and translate words in context.

SEMTAB rules disambiguate the meaning of words in the ST by identifying the semantic and syntactic structures underlying each meaning and provide the correct equivalent translation in the TL. In *OpenLogos*, they are invoked after dictionary look-up and during the execution of source and/or target syntactic rules (TRAN rules) at any point in the transfer phase in order to solve various ambiguity problems: (i) homographs such as *bank* which can be a transitive and intransitive verb or a noun; (ii) verb dependencies such as the different argument structures, [*speak to*], [*speak about*], [*speak against*], [*speak of*], [*speak on*], [*speak on N*(radio, TV, television, etc.)], [*speak over N1*(air) *about N2*], for the verb *speak*; (iii) MWUs of a different nature.

In order to explain the nature of this type of rule and how it operates, we will discuss it using the English phrasal verb *mix up* as an example. This verb assumes different meanings according to the words and the nature of the words it occurs with. In (1), it means to change the order or arrangement of a group of things, especially by mistake or in a way that you do not want. In (1), it means to prepare something by

combining two or more different substances. In (3), it means to think wrongly that somebody/something is somebody/something else and in (4), it means to be in a state of confusion.

*(1) try not to **mix up** all the different problems together.*
*(2) **mix up** the ingredients in the cookie mix.*
*(3) Tom **mixes** John **up** with Bill.*
*(4) I'm all **mixed up**.*

All these different meanings of mix up represented in (1)-(4) correspond, obviously, to different translations in Italian or any other language. Table 3 illustrates the SEMTAB rules comment lines written for the English-Italian language pair. These rules comprehend the different semantico-syntactic properties of each verb (also called linguistic constraints).

Semantic table (SEMTAB) rules	Italian Transfer
mix up(vt) in	mescolare in
mix up(vt) N in	mescolare N in
mix up(vt) N with	confondere N con
mix up(vt) N (human) in	confondere N in
mix up(vt) N (ingredient)	mescolare N
mix up(vt) N (medicine)	preparare N
mix up(vt) with	confondere con
mix up(vt) N (human,info) with	confondere N con
mix(vt) up (part)	confondere

Table 3 - SemTab rules comment lines for the verb *mix up*

For example, the SEMTAB rule number 8 describes the meaning (iii) of the verb *mix up*, by generalising to an abstract level of representation the nature of its direct object and classifying it under the Information or Human noun superset of the Semantico-syntactic Abstract Language (SAL) ontology. SAL is the *OpenLogos* representation language, containing over 1,000 concepts (expandable), organised in a hierarchical taxonomy consisting of Supersets, Sets and Subsets, distributed over all parts of-speech. In SAL, both meaning (semantics) and structure (syntax) are merged. This type of abstraction allows coverage of a number of different sentences in which different types of human nouns occur, as illustrated in (5)

(5) *Tom* ***mixed*** *John/him/the brother/the man/the buyer/the Professor, …* ***with*** *Bill.*

In order to properly disambiguate MWUs, a much wider context than the simple word level must be considered and context-sensitive semantico-syntactic rules applied.

An unusually powerful aspect of SEMTAB is that the rules are conceptual, deep structure rules, meaning that each rule can apply to a variety of surface structures, regardless of word order, passive/active voice construction, etc., approaching Chomsky theoretical assumptions concerning the universality of language. The same rule can apply to different surface structures, e.g., the *mixing up of languages, mix up the languages, languages mix up*, etc.

These very simple examples show how an adequate identification and analysis of MWUs in the source language by means of hand-drafted semantico-syntactic rules can influence the performance of an MT system with reference to different language pairs. Linguists can create rules that are more or less general or they can create very specific rules,

depending on the type of MWU. SEMTAB comment lines are written by a linguist, but the rules are built automatically using an appropriate tool (SEMANTHA or SEMTAB rule editor). The *OpenLogos* approach is thoroughly described in Scott (2003) and Barreiro, Scott, Kasper, & Kiefer (2011).

5.2. Multi-word unit processing in EBMT

EBMT relies on the analogy principle and therefore re-uses translations already stored in the system to translate MWUs.
MWU processing in EBMT has been discussed by several scholars over the last decades (Sumita et al. 1990 and 1991; Nomiyama, 1992; Franz et al., 2000; Gangadharaiah & Balakrishnan, 2006 and very recently Anastasiou, 2010). Basically, the EBMT approach to MWUs uses examples of possible translations of MWUs, integrated in many cases by linguistic rules. This is the case in Franz et al. (2000), Gangadharaiah and Balakrishnan (2006). The work by Anastasiou (2010) presents an exhaustive study to idiom processing in EBMT and a concrete application within the data-driven METIS-II system. The idiom linguistic resources used in the system are:

- a dictionary, consisting of 871 German idioms together with their translations into English;
- a corpus assembled from a subset of the EUROPARL corpus, a mixture of manually constructed data and examples extracted from the Web and, finally, a part of the DWDS, a digital lexicon of the German language;

- a set of rules to identify continuous and discontinuous idioms.

Idiom processing in Metis-II is divided into five stages: SL analysis, dictionary look-up, syntactic matching rules to identify idioms as a lexical unit, use of Expander, a tool that formalises the German sentence into the corresponding English target sentence by changing its word order, use of a ranking tool, Ranker, to choose the most appropriate target translation and, finally, a stage in which the systems generate the target sentence.

PRESEMT, on the other hand, another EBMT approach to MWUs proposed by Tambouratzis et al., (2012) relies on the use of a large monolingual and small parallel bilingual corpus with a few hundred sentences aligned at sentence level to identify sub-sentential segments in both SL and TL and therefore transfer structural information between languages.

Alignment is therefore a crucial aspect for EBMT. Alignment is an unsupervised methodology, i.e. a methodology that uses raw (un-annotated) input data to extract correspondences from large parallel corpora. Originally, the alignment process was used in translation memories and took place at sentence level in order to provide translators with ready solutions extracted from previous translations stored in the TM database. TMs either return to translators sentence pairs with identical source segments (exact matches) or sentences that are similar, but not identical to the sentence to be translated (fuzzy matches).

First generation TM systems, based on sentence alignment, showed severe shortcomings since the full repetition of a sentence only occurs in a very limited number of texts, i.e. technical documents, texts with related content or text revisions. In order to overcome these limitations,

research in this area is now addressing the possibility of alignment on a sub-sentential level.

Several scholars have focused their research on the possibility of automatically producing sub-sentential alignments from parallel bilingual corpora both to recover text chunks which have a higher occurrence probability than the sentence, but also to efficiently cope with the problem of translating MWUs.

In Groves et al. (2004), for instance, the methodology foresees the development of an automatic algorithm that aligns bilingual context-free phrase-structure trees at sub-structural level and its application to a subset of the English-French section of the HomeCentre corpus and more recently in Ozdowska, (2006), where syntactic information is used in a heuristics-based method that expands anchor alignment using a set of manually defined syntactic alignment rules.

Sub-sentential alignment seems to be a more suitable solution for the alignment of MWUs, especially if it takes into account the divergences between languages which can occur on the lexical, syntactic and semantic level, i.e. if the method adopted is able to cope with the asymmetries between languages which concern the translation of MWUs. For instance, if we take the English collocation *act contrary to law*, the Italian translation is *contravvenire alla legge* and it is immediately clear that a one-to-one word mapping between the two text segments is not possible and that a different solution should be found.

Recently, Barreiro et al. (forthcoming) address this problem by proposing a set of linguistically informed and motivated guidelines for aligning multilingual texts. The guidelines are based on the alignment of bilingual texts of the test set of the Europarl corpus covering all possible combinations between English, French, Portuguese and

Spanish. This contribution specifically analyses and propose guidelines which take into account MWUs and semantico-syntactic unit alignments. In particular, it offers alignment solutions for four different classes : lexical and semantico-syntactic, (MWUs, including support verb constructions, compound verbs and prepositional predicates), morphological (lexical versus non-lexical realisation such as articles and zero articles, the pro-drop phenomenon including subject pronoun dropping and empty relative pronoun, and contracted forms), morpho-syntactic (free noun adjuncts), and semantico-discursive (emphatic linguistic constructions such as pleonasm and tautology, repetition and focus constructions).

Other types of MWUs have also been taken into account with reference to alignment problems, and in particular (i) bilingual terminology by Claveau (2009), whose method relies on syntax to extract patterns such as Noun-Verb, Adjective-Noun, Prepositional Noun Phrase, etc; (ii) collocations by Seretan (2009) through bilingual alignments where POS-tags are equivalents or close (even with distant words). With regard to collocations, Segura &Prince (2011) propose an alignment process between pairs of sentences, strongly based on syntax. It relies on an *alignment memory*, consisting of a learnt set of good alignments as well as a rule-based process that asynchronously combines alignment constraints in order to maximise coverage.

5.3. Multi-word unit processing in SMT

In SMT, which evolved from the IBM word-based models (Brown et al., 1988, 1990) to phrase-based models (Zens et al., 2002; Koehn et al., 2003; Tillmann and Xia, 2003), the

problem of MWU processing is not specifically addressed.
The traditional approach to word alignment following IBM Models (Brown et al., 1993) shows many shortcomings related to MWU processing, especially due to their inability to handle many-to-many correspondences. Since alignment is performed only between single words, i.e. one word in the source language only corresponds to one word in the target language, these models are not able to handle MWUs properly.

The phrase-based alignment approach also does not take into account the problem of MWUs since, even if it considers many-to-many alignments as I have shown in section 3.3.2, some combinations of words or *n*-grams have no linguistic significance (e.g., *the war*) while others are linguistically meaningful (e.g., *cold war*). In SMT, phrases are therefore sequences of contiguous words not linguistically motivated and do not implicitly capture all useful MWU information.

In the state-of-the-art PB-SMT systems, the correct translation of MWUs occurs therefore only on a statistical basis if the constituents of MWUs are marked and aligned as parts of consecutive phrases (*n*-grams) in the training set and it is not generally treated as a special case where correspondences between source and target may not be so straightforward, i.e. it does not consist of consecutive many-to-many source-target correspondences.

MWU processing and translation in SMT started being addressed only very recently and different solutions have been proposed so far, but basically they are considered either as a problem of automatically learning and integrating translations or as a problem of word alignment as already described for EBMT.

The most used methodology is the following:

- Identification of possible monolingual MWUs. This phase can be accomplished using different approaches,(i) by means of morpho-syntactic patterns (Okita et al., 2010; Dagan & Church, 1994); (ii) statistical methods (Vintar & Fišer, 2008) and finally (iii) hybrid approaches (Wu & Chang, 2004; Seretan & Wehrli 2007; Daille, 2001; Boulaknadel, Daille, & Aboutajd, 2008).
- Alignment to extract and attribute the equivalent translations of the identified monolingual MWUs according to the different alignment methodologies.

Recently, increasing attention has been paid to MWU processing in SMT since it has been acknowledged that large scale applications cannot be created without proper handling of MWUs of all kinds. Current approaches to MWU processing move towards the integration of phrase-based models with linguistic knowledge and scholars are starting to use linguistic resources, either hand-crafted dictionaries and grammars or data-driven ones, in order to identify and process MWUs as single units.

A first possible solution is the incorporation of machine-readable dictionaries and glossaries into the SMT system, for which there are several straightforward approaches. One is to introduce the lexicon as phrases in the phrase-based table. Unfortunately, the words coming from the dictionary have no context information.

A similar approach is to introduce them to substitute the unknown words in the translation, but this poses the same problem as before. Okuma (2008) presents a more sophisticated approach where the lexicon words are introduced in the training corpus to enlarge their corpus. The criterion that they use is basically a Name Entity Recognition

classification which allows them to substitute the named entity in the original corpus with any named entity from their lexicon. Note that their lexicon contains only proper nouns but it could be extended to any word, given the appropriate tagging of the original corpus.

To deal with out-of-vocabulary words, Aziz et al., (2010) use entailment rules, in this case obtained from WordNet, and scored by different methods, including distributional similarity. The different scores are combined in an active learning fashion and the expert model is applied/learnt in such a way that it never harms the performance of the original model.

Another solutions for overcoming translation problems in MT and in SMT in particular is based on the idea that MWUs should be identified and bilingual MWUs should be grouped prior to statistical alignment (Lambert and Banchs, 2006). They adopted a method in which a bilingual MWUs corpus was used to modify the word alignment in order to improve the translation quality. In their work, bilingual MWU were grouped as one unique token before training alignment models. They showed on a small corpus, that both alignment quality and translation accuracy were improved. However, in their further study, they reported even lower BLEU scores after grouping MWUs by part-of-speech on a large corpus (Lambert and Banchs, 2006).

More recently, Ren et al. (2009) have underlined that experiments show that the integration of bilingual domain MWUs in SMT could significantly improve translation performance. Wu et al. (2008) propose the construction of phrase tables using a manually-made translation dictionary in order to improve SMT performance. Korkontzelos & Manandhar (2010) highlight that knowledge about MWUs leads to an increase of between 7.5% and 9.5% in the

accuracy of shallow parsing and finally Bouamor et al. (2011) affirm that integration of contiguous MWUs and their translations in Moses improves translation quality and propose a hybrid approach for extracting contiguous MWUs and their translations in a French-English parallel corpus.

Other solutions try to integrate syntactic and semantic structures (Chiang, 2005; Marcu et al., 2006; Zollmann & Venugopal, 2006), in order to obtain better translation results, but the solutions undoubtedly vary according to the different degrees of compositionality of the MWU.

Very recently identification and disambiguation of MWUs, as we already mentioned before, are being considered as a problem of Word Sense Disambiguation (WSD), i.e. the identification and the selection of the proper meaning of a word in a given context when it has multiple meanings, and several approaches to integrate WSD in SMT have been proposed.

The problem is here to select the most appropriate translation in TL to a given lexical unit in the SL. Some scholars refer to this problem also as word translation disambiguation (WTD), such as for instance Yang and Kirchoff (2012).

Methods in this research area range from supervised methods, that make use of annotated training corpora, to semi-supervised or minimally supervised methods, that rely on small annotated corpora as seed data in a bootstrapping process, or word-aligned bilingual corpora, and finally unsupervised methods that work directly from raw un-annotated corpora. Lately, there are a few papers which address inaccurate lexical choices in SMT from a WSD perspective and in particular Carpuat & Wu (2007) investigate a new strategy for integrating WSD into an SMT system, that performs fully phrasal multi-word disambiguation. They define the WSD task in such a way as

to match the exact same phrasal translation disambiguation task faced by phrase-based SMT systems.

Carpuat and Diab (2010), for instance, conducted an English- Arabic translation pilot study for task-oriented evaluation of MWUs in SMT using manually defined WordNet MWUs and a dictionary matching approach to MWU detection. They proposed two different integration strategies for monolingual MWU in SMT, considering different degrees of MWU semantic compositionality, i.e. (i) a static integration strategy that segments training and test sentences according to the MWU vocabulary, and (ii) a dynamic integration strategy that adds a new MWU-based feature in SMT translation lexicons. The first strategy allows a source text to be segmented in such a way that MWU are recognised and frozen as single lexical units. In this way during the training and decoding phases, MWUs are handled as distinct words regardless of their compositionality. In the dynamic strategy, the SMT system decides at decoding time how to segment the input sentence and it attempts to translate compositional MWU on the basis of a count feature in the translation lexicon that represents the number of MWUs in the input language phrase. On the basis of the positive outcome of their pilot study Carpuat and Diab conclude that it would be interesting to use more general MWU definitions such as automatically learned collocations (Smadja, 1993) or verb-noun constructions (Diab & Bhutada, 2009) on a larger scale.

In the wake of this latter study, different scholars have analysed this problem in more depth from different points of view.

Pal et al., (2010) show how single-tokenisation of two types of MWUs, namely named entities (NE) and compound verbs, as well as their prior alignment can boost the

performance of PB-SMT. (4.59 BLEU points absolute, 52.5% relative improvement on an English—Bangla translation task). This model is further implemented in Pal et al. (2011), who propose to pre-process a parallel corpus to identify noun-noun MWUs, reduplicated phrases, complex predicates and phrasal prepositions. Single tokenisation of noun-noun MWUs, phrasal preposition (source side only) and reduplicated phrases (target side only) provide significant gains (6.38 BLEU points absolute, 73% relative improvement) over the PB-SMT baseline system on an English- Bengali translation task.

Finally, Green et al. (2011) show that simple parsing models can effectively identify MWUs of arbitrary length, and that Tree Substitution Grammars achieve the best results. Their experiments based on the French Treebank (Abeillé et al., 2003) produced a 36.4% F1 absolute improvement for French over an *n-gram* surface statistics baseline, currently the predominant method for MWU identification.

Chapter 6 - Multi-word units processing: linguistic resources and tools for English-Italian MT

The previous chapter illustrated the state-of-the-art concerning MWU processing according to the different MT approaches and provided a bibliographic review of past and present research in this area. This chapter presents the methodological framework on which the research work in this dissertation is based and discusses a possible solution to identification and translation problems concerning MWU using a knowledge-based approach that adopts different strategies according to the different types of MWUs.

This approach relies on the use of linguistic resources, namely an electronic E-I MWU dictionary, containing different MWU typologies and a set of grammars. We will address two specific MWU typologies, i.e. terminological compound words and collocations with various degrees of compositionality.

The solutions suggested are obtained using FSTs, FSAs, RTNs and CFGs within the NLP tool NooJ, developed by Max Silberztein (Silberztein, 2005; Silberztein et al. 2007).

Section 6.1 illustrates some examples of mistranslations and presents a first research paper (Monti et al. 2011) which shows that a knowledge driven approach gives better results compared with an empirical one.

Section 6.2. details the methodology that can be applied to the identification and translation of MWUs.

6.1. MWU processing: better to *give up*?

In this section we illustrate the problems related to the processing of different types of MWU, namely compound nouns and collocations. In fact, MT, and especially SMT, still presents many translation problems related to these different linguistic aspects as is clear from the following examples taken from the MT translations of posts powered by Bing Translation (Microsoft) in the social network Facebook:

Figure 9- Translation of a post by Bing Translation: example n.1

The Italian translation **sta per i cani* is a typical example of mistranslation of an idiomatic expression, since the correct equivalent for the English expression *going to the dogs* should have been *sta andando in malora*.

Figure 10 - Translation of a post by Bing Translation: example n. 2

In the above example, the English expression *to get married* is also translated word-for-word and the Italian translation **di ottenere sposato* is completely wrong since it should have been *di sposarsi.*

Figure 11- Translation of a post by Bing Translation: example n. 3

In Figure 11, a last example of a machine translated post shows two different linguistic problems related to MWU processing, the first one is the Italian translation of the English compound noun *Combat photographer* which is rendered with **combattere editor* instead of *reporter di guerra* and the second problem is given by Bing's inability

to flawlessly translate the English verbal expression *to take pictures* with *fotografare*.

This is only a very small sample of a wide range of translation shortcomings with MWUs.

Monti et al. (2011), in a preliminary study of this dissertation, discuss and compare several examples of lexical ambiguities concerning MWUs in the translations performed by an SMT system, namely *Google Translate* (GT), and an RBMT system, i.e. the *OpenLogos* (OL) system and highlight, analyse and discuss how two MT systems of a different conceptual nature perform with regard to the different types of MWUs.

The comparison is based on a small corpus of non-specialised texts of about 300 sentences (approximatively10,000 words) containing MWUs extracted from the Web using two different tools: *Webcorp LSE*[45], developed by the Research and Development Unit for English Studies (RDUES), based in the School of English at Birmingham City University and *Web as Corpus*[46], developed by Bill Fletcher. The corpus was used to study the outputs of the abovementioned MT systems with reference to the translation of MWUs with the word *up*. This word is listed in the dictionary as a verb, adverb, noun, preposition and adjective and occurs in many different MWUs such as in the phrasal verbs *to mix up, to come up, to call up* or in expressions such as *to be up to something/someone*, *up and down*, and so on. In the following table, we present some of the results of the test we performed by comparing the *Google Translate* (GT) and *OpenLogos* (OL) outputs.

[45] http://www.webcorp.org.uk/webcorp_linguistic_search_engine.html
46 http://178.63.122.132/wac/

en_{SRC}	Why did these questions never **come up**?
it_{GT}	Perché mai queste domande **salire**?
it_{OL}	Perché queste domande non **si sono mai poste**?
en_{SRC}	and travels to some of the **world's trouble spots**
it_{GT}	e viaggia ad alcuni dei **problemi del mondo spot**
it_{OL}	e viaggia a alcuni dei **punti caldi del mondo**
en_{SRC}	**...** this year the Europeans **stood up for** freedom of speech
it_{GT}	… quest'anno gli europei **si alzò in piedi** per la libertà di parola.…
it_{OL}	… gli Europei hanno **sostenuto** la libertà del discorso.

Table 4- Comparison of MWU translation between an SMT and an RBMT systems

From this comparison it is clear that the linguistic approach of OL performs better than the statistical approach of GT. The translations by GT highlight inadequate MWU processing which heavily affects the understandability and correctness of the TTs compared with a general better performance by OL. In the first sentence in Table 4, GT is not able to select the appropriate translation of the verb *come up* in relation to the context. whereas the OL system takes into consideration the co-text of the sentence and analyses the verb *come up* in connection with the noun *questions*, thereby selecting the correct Italian translation: *porre delle domande.* In sentence 2, the multi-word unit *world's trouble spots* is not recognised as such by GT whereas it is translated correctly by the OL system as *punti caldi del mondo*. Finally, in sentence 3 the phrasal verb *stand up for* is translated literally by GT as *alzare in piedi*, selecting the wrong meaning in this context. On the contrary, the OL system produces an acceptable translation in Italian. The correct translation for the multi-word unit [*stand up for N/PRON*] where N/PRON is a non-animate noun or pronoun, is *difendere* or *lottare per*.

As a conclusion to this small experiment, Monti et al.

(2011) propose the use of Lexicon-Grammar lexical resources and semantic-syntactic rules (following the methodology adopted for the SEMTAB rules in *OpenLogos*) as a possible solution to overcome MT limitations with regard to the automated processing and translation of MWUs.

The main assumption of the methodology proposed in this dissertation is therefore that the proper treatment of MWUs calls for a computational approach which must be, at least partially, knowledge-based, and in particular should be grounded on an explicit linguistic description of MWUs, both using a dictionary and a set of rules.

Empirical approaches bring interesting complementary robustness-oriented solutions but taken alone, they can hardly cope with this complex linguistic phenomenon for various reasons. For instance, statistical approaches fail to identify and process non high-frequent MWUs in texts or, on the contrary, they are not able to recognise strings of words as single meaning units, even if they are very frequent.

Furthermore, MWUs change continuously both in number and in internal structure with idiosyncratic morphological, syntactic, semantic, pragmatic and translational behaviours.

The hypothesis is that a linguistic approach can complement probabilistic methodologies to help identify and translate MWUs correctly since hand-crafted and linguistically-motivated resources, in the form of electronic dictionaries and local grammars, obtain accurate and reliable results for NLP purposes.

In the next section we present the methodology adopted for this research work which is mainly based on the following elements:

- an accurate linguistic description that accounts for the description of the different types of MWUs and their semantic properties by means of well-defined steps: identification, interpretation, disambiguation and finally application.
- an NLP environment which allows the development and testing of linguistic resources.

6.2. MWU processing: a knowledge-based approach

This section presents a detailed description of the methodology for MWU processing that this research work is based on. Subsection 6.2.1 provides a general overview of the main principles that underlie the development of the methodology and the corresponding implementation. The second subsection 6.2.2. provides a detailed description of NooJ, the NLP tool used for development of the linguistic MWU resources, both dictionaries and grammars, and the subsequent analysis on a corpus collection, containing various MWU typologies. Section 6.2.3 illustrates the MWU dictionary and section 6.2.4., the local grammars used for the experiments.

6.2.1. NooJ: an NLP environment for the development and testing of MWU linguistic resources

NooJ is a freeware linguistic-engineering development platform used to develop large-coverage formalised descriptions of natural languages and apply them to large corpora, in real time.

The knowledge bases used by this tool are: electronic

dictionaries (simple words, MWUs and frozen expressions) and grammars represented by organised sets of graphs to formalise various linguistic aspects such as semi-frozen phenomena (local grammars), syntax (grammars for phrases and full sentences) and semantics (named entity recognition, transformational analysis). It integrates a broad spectrum of computational technology – from finite-state automata to enhanced/recursive transition networks.

NooJ is also used as a corpus processing system: it allows users to process thousands of sets of text files. Typical operations include indexing morpho-syntactic patterns, frozen or semi-frozen expressions (e.g. technical expressions), lemmatised concordances and performing various statistical studies on the results.

NooJ is a very flexible tool which can be used for many different purposes, not only as a linguistic-engineering development platform or corpus processor, but also as an information-extraction system, a terminology extractor and a machine-translation development tool, as well as for teaching Linguistics and Computational Linguistics.

Modules for several languages are currently available for free download: Arabic, Armenian, Bulgarian, Catalan, Chinese, Croatian, English, French, German, Hebrew, Hungarian, Italian, Polish, Portuguese and Spanish. Several other modules are under development.

NooJ's linguistic engine includes several computational devices used both to formalise linguistic phenomena and parse texts such as FSTs, FSAs, Recursive Transition Networks (RTNs),[47] Enhanced Recursive Transition

47 **Recursive Transition Networks (RTNs)** are grammars that contain more than one graph; graphs can be FST or FSA, and also include references to other embedded graphs; these latter graphs may in turn contain other references to the same or to other graphs. Generally, RTNs are used in NooJ

Networks (ERTNs),[48] Regular Expressions (RegExs),[49] Context Free Grammars (CFGs).[50]

NooJ is an annotation system which allows the annotation of any level of grammar to formalise various linguistic phenomena and apply the corresponding grammars in cascade. This means that the parsing approach in NooJ is bottom-up, i.e. it parses texts starting from the lowest levels of linguistic analysis (the character level) up to the most complex ones, including syntactic transformations and translations. During the parsing, NooJ automatically annotates a text with a Text Annotation Structure (TAS) on the basis of large dictionaries and extensive grammars.

The output of the parsing is therefore a text in which each recognised linguistic unit is associated to an annotation. NooJ adds annotations to the TAS at various stages of the analyses on the basis of the linguistic resources used. It can annotate morphological, lexical and syntactic linguistic

---

to build libraries of graphs from the bottom-up: simple graphs are designed; they are then re-used in more general graphs; these are in turn re-used, etc.

48 **Enhanced Recursive Transition Networks (ERTNs)** are RTNs that contain variables; these variables typically store parts of the matching sequences and are then used to perform operations with them (e.g. put their content in the plural, etc.), and then produce the resulting output. Because variables can be duplicated, inserted and/or displaced in the output, ERTNs give NooJ the power to perform linguistic transformations on texts. Examples of transformations include negation, passivisation, nominalisation, etc.

49 **Regular Expressions (RegExs)** represent a way to perform simple queries without having to build specific grammars. When the sequence to be located consists of a few words, it is much quicker to enter these words directly into a regular expression.

50 **Context-Free Grammars (CFGs in general)** constitute an alternative means to entering morphological or syntactic grammars. For instance, NooJ includes an inflectional/derivational module that is associated with its dictionaries so that it can automatically link dictionary entries with their corresponding forms in the corpora

phenomena. In the TAS, all unsolved ambiguities (Silberztein 2007) are kept.

Figure 12 shows the TAS of the English sentence: *For the foreground, I mix up burnt umber and deep violet for the winter scene and varying shades of green for summer*, having applied a simple word English dictionary together with a small dictionary containing the various occurrences of the verb *mix up* together with a local grammar for processing phrasal verbs.

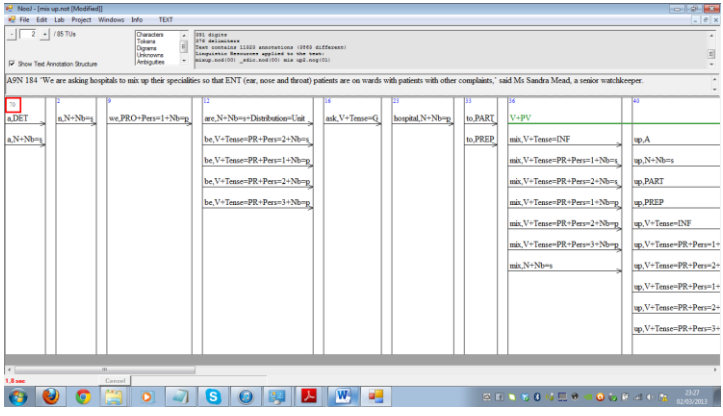


Figure 12 – Text Annotation Structure (TAS) in NooJ

NooJ is a tool that is particularly suitable for processing different types of MWUs and several experiments have already been carried out in this area: for instance, Machonis (2007 and 2008) analysed discontinous phrasal verbs using a phrasal verb dictionary containing over 1,200 entries and a local phrasal verb grammar; Anastasiadis, Papadopoulou & Gavriilidou (2011) used NooJ to automatically identify and translate Greek frozen expressions using a Greek frozen expression dictionary of 5,000 entries as well as a set of graphs created for their processing and automatic translation in French; Aoughlis (2011) developed a French-English MT

system for Computer Science Compound Words and finally Vietri (2008) presents a translation experiment from Italian to English using NooJ in the area of terminological compound words used in Economics. These are only a few examples of the various analysis performed in the last few years on MWU using NooJ as an NLP development and testing environment.

A powerful feature of NooJ is that it processes simple words and MWUs in a unified way, i.e. they are stored in the same dictionaries and their inflectional and derivational morphology and annotations are formalised in the same way as those of simple words. In order to do this, NooJ uses standard dictionaries and standard syntactic grammars at runtime. The following sections describe the linguistic resources developed and used for MWU identification and translation.

6.2.2. Linguistic Resources: MWU dictionary and grammars

The linguistic resources developed for the purposes of this dissertation are (i) a Dictionary of the English-Italian MWUs (EIMWU) and (ii) a set of grammar rules.

6.2.2.1 Dictionary of English-Italian multi-word units (EIMWU.dic files)

EIMWU.dic is a dictionary used to represent and recognise various types of MWUs.

This dictionary is based on a contrastive English-Italian analysis of continuous and discontinuous MWUs with different degrees of variability of co-occurrence among word compositionality and different syntactic structures. The main part of the dictionary consists of phrasal verbs, support verb

constructions, idiomatic expressions and collocations which we have already discussed in the previous chapters.

It includes only a few compound words of different types. Of these MWUs, collocations are the most frequent. These MWUs have specific properties such as arbitrariness and cohesion as lexical clusters (Smadja, 1999; McKeown and Radev,1999) and account for the many translation mistakes that can be found in MT outputs.

Indeed, the translation of MWUs requires the knowledge of the correct equivalent in the target language which is hardly ever the result of a literal translation. Given their arbitrariness, MT has to rely on the availability of ready solutions in both languages in order to perform an accurate translation process (McKeown and Radev, 1999).

Each entry of the dictionary is given a coherent linguistic description consisting of:

- the grammatical category for each constituent of the MWU: noun (N), Verb (V), adjective (A), preposition (PREP), determiner (DET), adverb (ADV), conjunction (CONJ);
- one or more inflectional and/or derivational paradigms (e.g. how to conjugate verbs, how to nominalise them), preceded by the tag +FLX;
- one or more syntactic properties (e.g. "+transitive" or +N0VN1PREPN2);
- one or more semantic properties (e.g. distributional classes such as "+Human", domain classes such as "+Politics");
- the translation into Italian.

Here are some examples of entries extracted from the English-Italian multi-word bilingual dictionary:

> ask,V+FLX=ASK+JM+FXC+Intrans+PREP="about"+IT="informa
> rsi su"
> at,PREP +JM+FXC+N="present"+IT="attualmente"
> robust,ADJ+JM+FXC +N="pace"+IT="andatura sostenuta"

In the dictionary all linguistic information, i.e. all the syntactic, semantic, morphological properties of an entry, is coded in the form of features preceded by the character"+" and associated with a value. The feature/value relationship is written in the form +name of feature=value.

For instance for the English MWU *ask about*, the lexical entry *ask* is followed by the following tags: (i) "V" which indicates its grammatical category, i.e. verb, (ii) "+FLX=ASK", which indicates that the inflectional paradigm of the word *ask* is "ASK", i.e. it takes the inflection pattern of the verb *ask*, as stored in the English Inflectional Description files (.nof)[51], (iii) "+FXC" which means that it is a frozen expression compound, (iv) "+Intrans" which means that the verb is used in its intransitive form (v) PREP= "about" which means that the verb *ask* collocates with the preposition *about* (vi) "+IT= "informarsi su" for the Italian translation of the verbal MWU *ask about.*

The same feature of an entry can be repeated as many times as necessary to indicate alternative possibilities for that specific feature. For instance, in the entry:

> acquire,V+FLX=LIVE+JM+FXC+Trans+N1="knowledge"+N1="e
> xperience"+N1="skills"+IT="acquistare N1"

[51] Inflectional Description files contain the inflection patterns of the words.

the feature “N1” is repeated three times to indicate that the verb *acquire* can collocate with the nouns *knowledge*, *experience*, and *skills* and it always takes the Italian translation “acquistare N1”.

Different types of semantic features such as +Conc”, “+Abstr”,”+Hum” can also be assigned.

When a word is associated with different set of properties, i.e. different syntactic or distributional information, the word is duplicated and the corresponding form is processed as ambiguous. If we consider the verb *act*, we have as many entries as necessary to describe the different meanings of the verb and its translations (see Figure 13), such as for instance En. *act as if* →It. *agire come se*; En. *act for N2* → It. *rappresentare N2*, En. *act in interest of* → It. *agire nell’interesse di*; En. *act contrary to law* → It. *contravvenire alla legge*; En. *act out* → It. *rappresentare*, and so on.

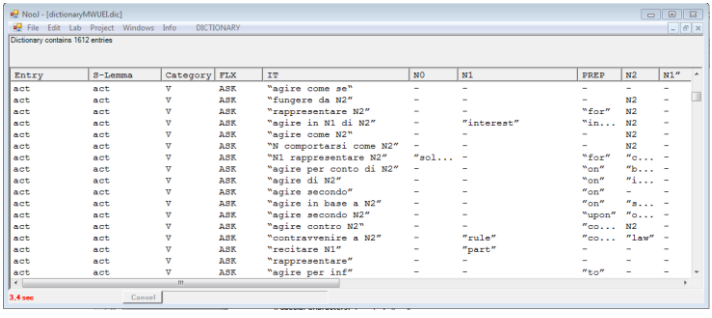


Figure 13- Dictionary entries for the English verb *act*

In order to analyse texts, NooJ needs dictionaries that contain and describe the words in a text and a mechanism to link these lexical entries to all the corresponding inflected and/or derived forms that occur in texts.

The inflection features “FLX” contains the value needed

to inflect the lexical entry according to the appropriate inflection pattern. For instance if we consider the following entries:

ask,V+FLX=ASK+JM+FXC+Intrans+PREP="about"+IT="informarsi su"
abound,V+FLX=ASK+JM+FXC+Intrans+PREP="in"+IT="essere ricco di"

they are both associated with the same conjugation class, i.e., ASK, stored in the English inflectional description file for verbs compiled by NooJ into a Finite State Transducer (Verb.nof):

ASK=<E>/INF | <E>/PR+1+2+s | <E>/PR+1+2+3+p | s/PR+3+s | ed/PT+1+2+3+s+p | ed/PP | ing/G;

The Verb.nof file is therefore an inflectional grammar used to represent the inflection (e.g. conjugation) properties of verbal lexical entries entered in the form of rules as in the example above.

The ASK class is defined using:

1. the following special operators:

<B>: keyboard Backspace
<D>: Duplicate current character
<E>: Empty string
<L>: keyboard Left arrow
<N>: go to end of Next word form
<P>: go to end of Previous word form
<R>: keyboard Right arrow
<S>: delete/Suppress current chararacter
Arguments for commands <B>, <L>, <N>, <P>, <R>, <S>:
xx number: repeat xx times
W: whole word

2. the following ## Inflectional Codes:

Singular: s

Plural: p
First Person: 1
Second Person: 2
Third Person: 3
Infinitive: INF
Present: PR
Preterit: PT
Past Participle: PP
Gerundive: G

In this way the inflectional class of ASK can be described as follows:

<E>/INF = Infinitive: ask
<E>/PR+1+2+s = Present simple: 1,2 person singular: ask,
<E>/PR+1+2+3+p = 1,2,3, person plural: ask
s/PR+3+s = 3 person singular: asks,
ed/PT+1+2+3+s+p = Preterit: asked
ed/PP= Past Participle: asked
ing/G = Gerundive: asking

This paradigm states that if we add an empty string to the lexical entry *ask* we get the infinitive form of the verb (*to ask*), the first person (*I ask*) or the second person singular (*you ask*), or any of the plural forms (*we ask*) of the Present simple. If we add an "s" to the entry we obtain the Present simple, third person singular (*he asks*). If we add "ed" we obtain the past participle form (*asked*) and any of the preterit forms (*asked*). If we add "ing" we obtain the gerundive form (*asking*).

This inflection class associated to the verb *abound* generates the correct conjugation pattern:

<E>/INF = Infinitive: abound
<E>/PR+1+2+s = Present simple: 1,2 person singular: abound,
<E>/PR+1+2+3+p = 1,2,3, person plural: abound
s/PR+3+s = 3 person singular: abounds,
ed/PT+1+2+3+s+p = Preterit: abounded

ed/PP= Past Participle: abounded
ing/G = Gerundive: abounding

The EIMWU.dic contains different types of MWU POS patterns. The main part of the dictionary consists of different types of verb entries. In the next paragraphs of this section, the main verb structures are explained with examples extracted from the *British National Corpus,* from the Internet by means of the *WebCorp LSE* application or with our own examples together with the Italian translations. Finally, the corresponding dictionary entry for each example of MWU POS pattern is provided.

[VIntrans +N0]
This category encompasses all intransitive verbs which collocate:

1. with a specific noun with subject function (N0):

   (1) En. *The **storm broke** at five o'clock.* [BNC] → It. *La **tempesta scoppiò** alle cinque.*

   This structure is formalised in the EIMWU dictionary as follows:

   break,V+FLX=SPEAK+JM+FXC+Intrans+N0="storm"+IT="N0 scoppiare"

2. with a series of specific nouns with subject function (N0):

   (2) En. *The **problem arises** when more common, everyday names are available.* [BNC] → It. *Il **problema sorge** quando sono disponibili nomi comuni di uso corrente.*
   (3) En. *The **question arises** as to why there is so little official action to combat soil erosion.* [BNC] → It. *La **questione sorge** rispetto al perché c'è una così limitata*

azione ufficiale per combattere l'erosione del suolo.

This structure is formalised in the EIMWU dictionary as follows:

arise,V+FLX=RISE+JM+FXC+Intrans+N0="problem"+N0= "question"+IT="N0 sorgere"

3. with a generic class of nouns with subject function, like in the following example where the verb *bruise* collocates with any human noun (N0Hum):

   (4) En. *If the number of platelets in your blood goes down **you** may **bruise** easily.* [WebCorp] → It. *Se il numero di piastrine del tuo sangue scende, **ti** puoi **coprire di lividi** molto facilmente.*

   This structure is formalised in the EIMWU dictionary as follows:

   bruise,V+FLX=LIVE+JM+FXC+Intrans+N0Hum+IT="N0 coprirsi di lividi"

4. with any noun with subject function (N0) and with an Italian translation represented by an MWU:

   (5) En. *Ramsey **bicycled** over to Wordsworth Grove to see if there were any letters.* [BNC] → It. *Ramsey **andò in bicicletta** verso Wordsworth Grove per vedere se c'erano delle lettere.*

   This structure is formalised in the EIMWU dictionary as follows:

   bicycle,V+FLX=LIVE+JM+FXC+Intrans+N0+IT="N0 andare in bicicletta"

[VIntrans +N0+ADJ]

This category encompasses all intransitive verbs that collocate with a specific adjective, like in:

(1) En. *Do not allow the patient to **lie flat**.*[WebCorp] → It. *Non permettere al paziente di **sdraiarsi**.*

These structures are formalised in the EIMWU dictionary as follows:

lie,V+FLX=LIE+JM+FXC+Intrans+N0+ADJ="flat"+IT= "sdraiarsi"

[VIntrans+N0+PART]

This category consists of phrasal verbs and encompasses all intransitive verbs that collocate with a particle and:

1. a generic noun with subject function (N0):

   (2) En. *We need to **bear down** and go right on into the future.* [BNC] → It. *Dobbiamo **avanzare** e andare dritti verso il futuro.*

   This structure is formalised in the EIMWU dictionary as follows:

   bear,V+FLX=BEAR+JM+FXC+Intrans+N0+PART="down"+IT=" avanzare"

2. with a specific noun with a subject function (N0):

   (3) En. ***Clouds would bank up** about midday, and showers fall* → It. ***Le nuvole si addensavano** a mezzogiorno e c'erano degli acquazzoni.*

This structure is formalised in the EIMWU dictionary as follows:

bank,V+FLX=ASK+JM+FXC+Intrans+PART=”up”+N0=”cloud”+ IT=“N addensarsi”

3. with a series of specific nouns with subject function (N0):

   (4) En. *It appears that the* ***recession*** *has* ***bottomed out****, and we are seeing an improvement in economic conditions.*→ It. *Sembra che la recessione* ***abbia toccato il fondo*** *e cominciamo a vedere un miglioramento della condizione economica.*

   (5) En. *The* ***market*** *would* ***bottom out*** *at around 920 points.* [BNC] → It. *Il mercato* ***toccava il fondo*** *a circa 920 punti.*

   This structure is formalised in the EIMWU dictionary as follows:

   bottom,V+FLX=ASK+JM+FXC+Intrans+PART=”out”+N0=“rate” +N0=”stock”+N0=”market”+N0=”profits”+N0=”recession”+IT= “N0 toccare il fondo”

4. with a generic class of nouns with subject function, like in the following example where the verb *blank* collocates with any noun that is a human noun (N0Hum):

   (6) En. *When I tried to remember my client's name, I just* ***blanked out*** → It. *Quando cercai di ricordare il nome del mio cliente* ***ebbi un vuoto di memoria***.

   This structure is formalised in the EIMWU dictionary as follows:

   blank,V+FLX=ASK+JM+FXC+Intrans+PART=”out”+N0Hum+IT =“N avere un vuoto di memoria”

[VIntrans +N0+PART+PREP+N2]

This category consists of phrasal verbs and encompasses all intransitive verbs which collocate that a particle, a specific preposition followed by:

1. a generic noun (N2):

   (7) En. *Iveco has also started to* ***branch out into eastern Europe***. [BNC] → It. *L'Iveco ha anche iniziato ad* ***espandersi in Europa orientale***.

   This structure is formalised in the EIMWU dictionary as follows:

   branch,V+FLX=ABOLISH+JM+FXC+Intrans+N0+PART="out" +PREP="into"+N2+IT="espandersi in N2"

2. a specific noun (N2):

   (8) En. *Sabbath means 'to cease' or* ***'to break off' from work*** → It. *Sabbath significa 'terminare' o* ***interrompere il lavoro***.

   This structure is formalised in the EIMWU dictionary as follows:

   break,V+FLX=SPEAK+JM+FXC+Intrans+N0+PART="off" +PREP="from"+N2="work"+IT="interrompere il lavoro"

[VIntrans +N0+PART+PREP+Ving]

This category consists of phrasal verbs and encompasses all intransitive verbs that collocate with a particle, a specific preposition and a verb in the –ing form:

(9) En. *At some point we will have to* ***break off sending*** *immortals south* [WebCorp] → It. *Ad un certo punto dovremo* ***smettere di inviare*** *gli immortali al sud.*

This structure is formalised in the EIMWU dictionary as follows:

break,V+FLX=SPEAK+JM+FXC+Intrans+N0+PART="off" +PREP="from"+Ving+IT="smettere di Vinf"

[VIntrans +N0+PREP+N2]

This category encompasses all intransitive verbs that collocate with a specific preposition and:

1. a generic noun (N2):

(10) En. *Logic needs to* ***account for logical relations*** *among sentences.* [BNC] → It. *La logica* ***deve spiegare le relazioni logiche*** *tra le frasi.*

This structure is formalised in the EIMWU dictionary as follows:

account,V+FLX=ASK+JM+FXC+Intrans+N0+PREP="for"+N2+IT ="spiegare N2"

2. a specific noun (N2):

(11) En. *The prince* ***acceded to the throne*** [Freedict] → It. *Il principe* ***è salito al trono***.

This structure is formalised in the EIMWU dictionary as follows:

accede,V+FLX=LIVE+JM+FXC+Intrans+N0+PREP="to"+N2= "throne"+IT="salire al trono"

3. a series of specific nouns:

(12) En. *The emir refused to* ***accede to*** *Iraq's financial*

demands. [WebCorp] → It. *L'emiro si è rifiutato di aderire alle richieste finanziarie dell'Iraq.*

(13) En. *A third party may claim the right to* ***accede to*** *a* ***treaty*** *in accordance with its terms.* [BNC] → It. *Una terza parte può reclamare il diritto di* ***aderire ad*** *un* ***trattato*** *in conformità delle sue clausole.*

These structures are formalised in the EIMWU dictionary as follows:

accede,V+FLX=LIVE+JM+FXC+Intrans+N0+PREP="to"+N2= "treaty"+N2="demand"+IT="aderire a N2"

4. with a generic class of nouns with subject function, like in the following example where the verb *allow* collocates with any human noun (N0Hum) and the preposition *for* followed by any noun (N2):

(14) En. *But within that framework* ***he allowed for*** *as much flexibility as possible.* [BNC] → It. *Ma nell'ambito di quel contesto* ***ha tenuto conto della*** *massima flessibilità possibile*.

This structure is formalised in the EIMWU dictionary as follows:

allow,V+FLX=ASK+JM+FXC+Intrans+N0Hum+PREP="for"+N2 +IT="tener conto di N2"

[VTrans +N0+N1]

This category encompasses all transitive verbs that collocate:

1. with a specific N1 with an object function:

(15) En. *They did not* ***advance any reason*** *for the differences*

which they identified. [WebCorp] → It. *Non hanno esposto le ragioni delle differenze che hanno identificato.*

This structure is formalised in the EIMWU dictionary as follows:

advance,V+FLX=LIVE+JM+FXC+Trans+N0+N1="reason"+IT= "esporre N1"

2. with a series of specific nouns with object function (N1):

(16) En. *They **run this shop** in order to fund their small animal rescue charity* [WebCorp] → It. ***Gestiscono questo negozio** per finanziare la loro piccola organizzazione di beneficenza per animali.*

(17) En. *They also **run a restaurant** and cooking school.*[WebCorp] → It. ***Gestiscono** inoltre **un ristorante** e una scuola di cucina.*

This structure is formalised in the EIMWU dictionary as follows:

```
run,V+FLX=RUN+JM+FXC+Trans+N0+N1="hotel"
+N1"shop"+N1="restaurant"+IT="gestire N"
```

3. with a generic class of nouns, like in the following example where the verb *advise* collocates with any human noun with an object function (N1):

(18) En. *You **tip waiters** in restaurants, right?* [WebCorp] → It. *Si da la mancia nei ristoranti, è così?*

This structure is formalised in the EIMWU dictionary as follows:

```
tip,V+FLX=+JM+FXC+Trans+N0+N1hum+IT="dare la
mancia a N"
```

4. with any noun with an object function (N1) and with an Italian translation represented by an MWU:

   *(19) En. Women may be attacked because their customs and dress do not **fit gender** stereotypes.* [WebCorp]→It. *Le donne possono essere attaccate perché i loro usi ed i loro abiti non **si adattano agli stereotipi** di genere.*

This structure is formalised in the EIMWU dictionary as follows:

```
fit,V+FLX=ADMIT+JM+FXC+Trans+N0+N1+IT=
"adattarsi a N"
```

[VTrans +N0+ADJ+N1]

This category encompasses all transitive verbs that collocate with a specific adjective, like in:

(20) En. *His momentary surprise was enough to **break him free** of the killing impulse*. [WebCorp]→ It. *La sua sorpresa momentanea fu sufficiente a **liberarlo** dal suo impulso omicida.*

This structure is formalised in the EIMWU dictionary as follows:

```
break,V+FLX=SPEAK+JM+FXC+Trans+N0+N1+ADJ=
"free"+IT="liberare N1"
```

[VTrans +N0+PART +N1]

This category consists of phrasal verbs and encompasses all transitive verbs which collocate with a:

1. with a specific N0 with a subject function:

*(21) En. Camus carefully manipulates the plot to **bring up the question** of innocent suffering.* [BNC] → It. *Camus manipola abilmente la trama per **sollevare il problema** della sofferenza innocente.*

This structure is formalised in the EIMWU dictionary as follows:

bring,V+FLX=BRING+JM+FXC+Trans+N0+PART="up"+
N1="question"+IT="sollevare N1(problema)"

2. with a series of specific nouns:

(22) En. *The initiative is thus handed to the opposition, which can then **bring forward evidence** of the missing sovereignty.* [WebCorp] →It. *L'iniziativa è così consegnata all'opposizione, che può quindi **presentare la prova** della mancata sovranità.*

This structure is formalised in the EIMWU dictionary as follows:

bring,V+FLX=BRING+JM+FXC+Trans+PART="forward"+
N1="evidence"+N1="argument"+N1="proposal"+IT="presentare N1"

3. with a generic class of nouns, like in the following example where the verb *bring* collocates with any human noun:

(23) En. *The Mary White […]was able to **bring off** seven of the American crew* [BNC] → It. *La Mary White fu in grado di **salvare** sette persone dell'equipaggio americano.*

This structure is formalised in the EIMWU dictionary as follows:

bring,V+FLX=BRING+JM+FXC+Trans+PART="off"+N1Hum+IT ="salvare N1"

4. with any noun:

(24) En. *That way you will not **burn off** too many calories* [WebCorp] → It. *In questo modo non si **bruceranno** troppe calorie.*

This structure is formalised in the EIMWU dictionary as follows:

burn,V+FLX=BUILD+JM+FXC+Trans+PART="off"+N1+IT= "bruciare"

[VTrans +N0+PART +N1+PREP+N2]

This category consists of phrasal verbs and encompasses all transitive verbs which collocate with a particle and a specific preposition followed by:

1. a specific noun (N2):

(25) En. *Can I **bring back from memory** all or most of what I had learned before?* [WebCorp] → It. *Posso **richiamare alla mente** tutto o quasi tutto ciò che ho imparato in precedenza?*

This structure is formalised in the EIMWU dictionary as follows:

bring,V+FLX=BRING+JM+FXC+Trans+PART="back"+N1+ PREP="from"+N2="memory"+IT="richiamare a N2(mente)"

2. a generic noun (N2):

(26) *En. The U.S. interim administration in Baghdad is scheduled to **hand over power to a transitional***

government [WebCorp] → It. E*' stato stabilito che l'amministrazione provvisoria statunitense a Baghdad* ***consegnerà il potere ad un governo di transizione***.

This structure is formalised in the EIMWU dictionary as follows:

hand,V+FLX=+JM+FXC+Trans+PART="over"+N0+N1
+PREP="to"+N2+IT="consegnare N1 a N2"

3. a generic class of nouns:

(27) En. *The perception that immigrants* ***take away jobs from*** *the existing population[…]do not find confirmation in the analysis of data laid out in this report.* [WebCorp] → It. *La sensazione che gli immigrati* ***tolgano il lavoro alla*** *popolazione esistente (…) non trova conferma nell'analisi dei dati presentati in questo rapporto.*

This structure is formalised in the EIMWU dictionary as follows:

take,V+FLX=TAKE+JM+FXC+Trans+PART="away"+N0+N1
+PREP="from"+N2Hum"+IT="togliere a N"

[VTrans +N0+N1+PREP+N2]

This category consists of transitive verbs which collocate with a:

1. with a specific N1 with an object function:

(28) En. *Breaking bad news to someone is never a pleasant task.* [WebCorp] → It. ***Comunicare*** *cattive* ***notizie*** *a qualcuno non è mai un compito piacevole.*

This structure is formalised in the EIMWU dictionary as follows:

break,V+FLX=SPEAK+JM+FXC+Trans+N0+N1="news"+PREP="to"+N2Hum+IT="comunicare N1 a N2"

2. with a series of specific nouns with an object function:

(29) En.:*Israel seeks to renew cooperation with Palestinians* [WebCorp] → It. *Israele cerca di* ***riprendere la cooperazione con*** *i Palestinesi.*

(30) En.*Israeli Arab group calls on Abbas to* ***renew dialogue with*** *Hamas* [WebCorp] → *It. I gruppi arabo israeliani invitano Abbas* ***a riprendere il dialogo c****on Hamas.*

This structure is formalised in the EIMWU dictionary as follows:

renew,V+FLX=ASK+JM+FXC+Trans+N0+N1="talk"+N1="discussion"+N1="negotiation"+N1="summit"+N1="dialogue"+N1="cooperation+PREP="with"+IT="riprendere"

3. with a generic class of nouns, like in the following example where the verb *accustom* collocates with any human noun with an object function:

(31) En. *Let no one think that it is nothing, to* ***accustom people to*** *give a reason for their opinion* [WebCorp] →It. *Che nessuno pensi sia una sciocchezza* ***abituare le persone a*** *dare una motivazione alle loro opinion.*

This structure is formalised in the EIMWU dictionary as follows:

accustom,V+FLX=ASK+JM+FXC+Trans+N0+N1Hum+PREP="to"+N2+IT="abituare N1 a"

4. with any noun with an object function:

(32) En. *I* ***advise fellows on*** *the right to relax and enjoy the*

fun [WebCorp] → It. ***Consiglio*** *gli amici sul diritto a rilassarsi e a divertirsi.*

This structure is formalised in the EIMWU dictionary as follows:

advise,V+FLX=LIVE+JM+FXC+Trans+N0+N1+PREP="on"+N2+IT="consigliare N1 su N2"

5. with a specific noun (N2):

(33) En. *The corporation desires to* ***acquire land by purchase*** [WebCorp] → It. *La corporazione desidera* ***ottenere*** *la terra* ***mediante acquisto***.

This structure is formalised in the EIMWU dictionary as follows:

```
acquire,V+FLX=LIVE+JM+FXC+Trans+N0+N1+
PREP="by"+N2="purchase"+IT="ottenere  mediante
N2
```

[VTrans +N0+N1+PREP+Ving]

This category consists of transitive verbs that collocate with a noun as direct object and a preposition followed by the gerundive form of a verb:

(34) En. *Medical associations* ***bar doctors from*** *participating in executions* [WebCorp] → It. *Le associazioni mediche* ***impediscono ai medici di partecipare*** *nelle esecuzioni.*

This structure is formalised in the EIMWU dictionary as follows:

```
bar,V+FLX=ADMIT+JM+FXC+Trans+N1Hum+PREP="from
"+VG+IT="impedire a N1 di inf"
```

[VTrans +N0+N1+to+Vinf]

This category consists of transitive verbs which collocate with a noun as direct object and a preposition followed by the gerundive form of a verb:

(35) En. *And I **beg you to** explain why I should not go.* [WebCorp] → It. *E ti **chiedo di spiegarmi** perché non dovrei andare.*

This structure is formalised in the EIMWU dictionary as follows:

```
beg,V+FLX=BEG+JM+FXC+Trans+N1+PREP="to"+VINF+
IT="chiedere a N1 di inf"
```

The EIMWU dictionary consists of nominal, adjectival and prepositional MWUs as well. Nominal units are only very few, and concern wither frozen expressions like *perfect pitch.*

Figure 14 shows all the different entries listed in the dictionary for the adjective *open.*

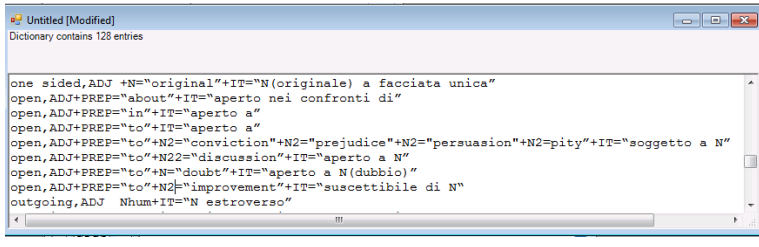


Figure 14 - Dictionary entries for the adjective *open*

Figure 15 shows all the different entries listed in the dictionary for the preposition *on.*

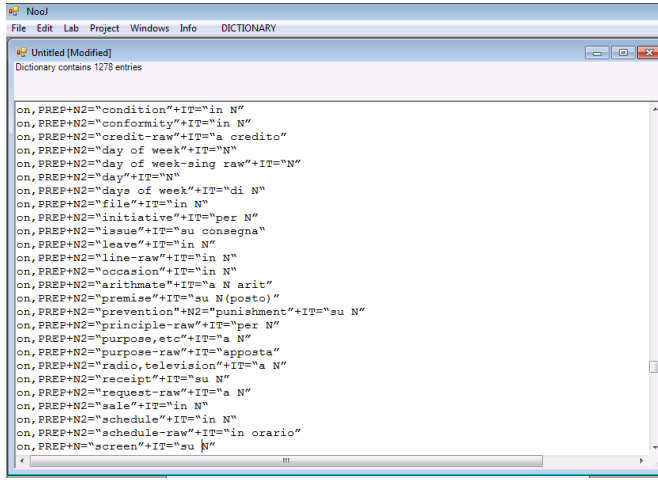


Figure 15 - Dictionary entries for the preposition *on*

6.2.2.2.Local Grammars (.nog files) for MWUs

In NooJ, syntactic or semantic grammars (.nog files) are used to recognise and annotate expressions in texts, e.g. to tag noun phrases, certain syntactic constructs or idiomatic expressions, extract certain expressions or interest (name of companies, expressions of dates, addresses, etc.), or disambiguate words by filtering out some lexical or syntactic annotations in the text.

These grammars recognise different types of MWU, such as frozen and semi-frozen units, and are particularly useful with discontinuous MWUs.

For instance, if we want to analyse the phrasal verb to *mix up* in the following sentences:

(1) *try not to **mix up** all the different problems together*
(2) ***mix up** the ingredients in the cookie mix*
(3) *Tom **mixes** John **up** with Bill*

in order to identify the verb in the sentences we need to apply the following local grammar:

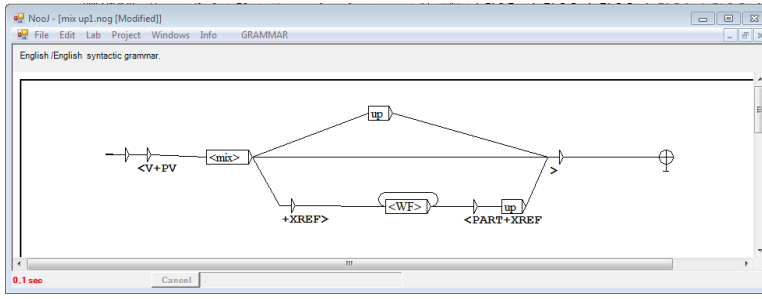


Figure 16 - *Mix up* local grammar

This grammar allows the phrasal verb *mix up* to be identified as a single lexical unit, consisting of a first component <V+PV> and a second one <PART> in all the different abovementioned sentences, also when it is discontinuous as in example (3), where the resulting annotation structure is as follows:

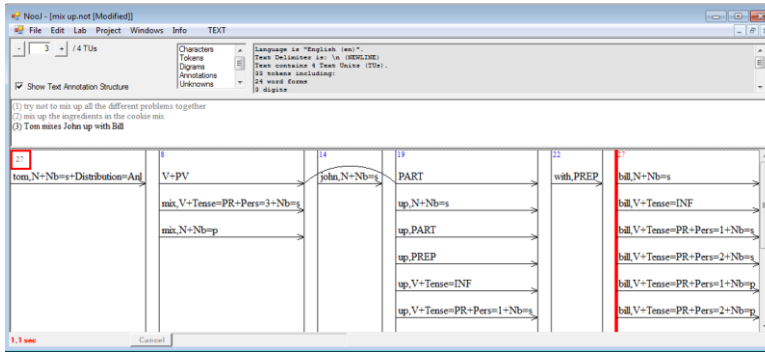


Figure 17 – TAS for discontinuous form of *mix up*

This annotation allows NooJ to identify and process *mix up*

as a single lexical unit but at the same time it keeps the information for the single words: *mix* as a Verb and *up* as a particle.

If we have a look at the concordances of *mix up* in the analysed text, NooJ is able to locate all the various occurrences of the verb *mix up*, as illustrated by Figure 18.

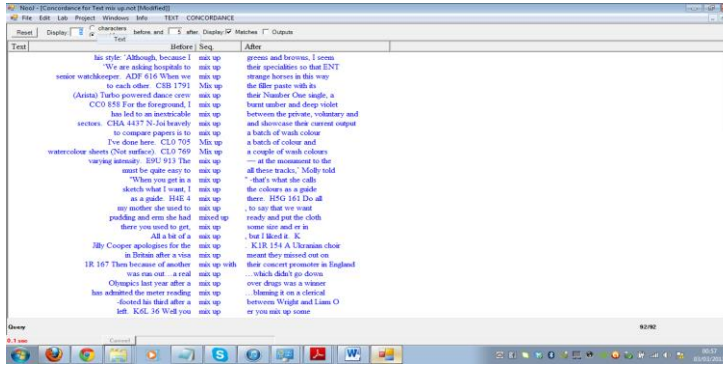


Figure 18 - Concordances of the verb *mix up* in NooJ

As we have seen in Section 3.1.2, the verb *mix up* takes different meanings to which correspond different translations. In NooJ it is the possible to develop a local grammar or graph, which used along with the EIMWU.dic dictionary, can identify occurrences of continuous and discontinuous phrasal verbs and show in the TAS all the translations of the verb. In this way it is therefore possible to concatenate verb and particle. Figure 19 shows the rule that is applied together with the dictionary in which all the different structures and correspondent translations of *mix up* are listed.

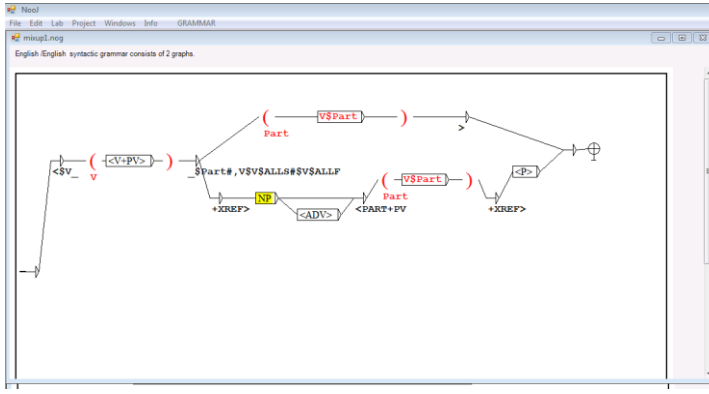


Figure 19 - Local grammar for phrasal verbs

If we apply this grammar together with the EIMWU.dic on a text, the corresponding TAS will include the concatenated form of *mix up* and all information associated with the verb, including all possible translations, as shown in the figure below.

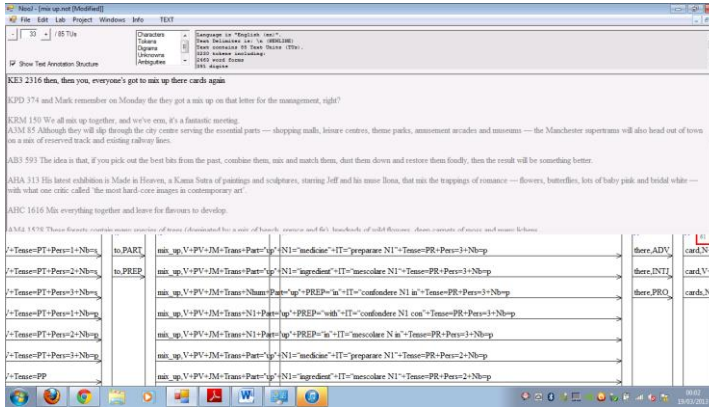


Figure 20 - TAS resulting from the interaction of a dictionary and a local grammar

In this way, TAS can be used to automatically tokenise all occurrences of *mix up,* both continuous and discontinuous ones, in texts as single units and, at the same time, to provide, for instance, to SMT the most appropriate translation in a given context.

The rule illustrated in Figure 20, is a very general rule that can be applied to identify, disambiguate and translate all phrasal verbs listed in the EIMWU.dic.

This is only an example of how rules can be applied in NooJ to disambiguate MWUs, future work will produce further grammars for MWU processing.

Chapter 7 - Conclusions and future work

This chapter summarises the work presented in previous chapters and describes the current and future directions of our research. It is divided into two sections: dissertation achievements (Section 7.1), and future perspectives (Section 7.2).

7.1. Dissertation achievements

After many years of research and improvements together with the adoption of different approaches in MT, MWUs still represent a critical area in current translation technologies. Due to their intrinsic morpho-syntactic and semantic properties, MWUs give rise to many ambiguities which seriously challenge the precision and quality of MT outputs.

In this dissertation, I have tried to show that a linguistic approach to MWUs, by means of a precise analysis and formalisation of their linguistic properties, can improve the MT processing task as far as MWU identification is concerned.

I began this research work by presenting its motivations and the obstacles posed to MT by this linguistic phenomenon which is very frequent both in the current usage of language and in languages for special purposes, and yet is very difficult to handle properly in NLP applications and particularly in MT, due to its characteristics, i.e. arbitrariness, heterogeneity, semantic/syntactic variability and translation idiosyncrasies.

The dissertation began with a brief historical overview of MT up to current trends in MT technologies to give an idea of the development of MT over almost seventy years and has attempted to explain the reasons for the recent spread of online MT services, from free online MT services to online applications where MT is integrated to perform different functions such as CLIR, IM and collaborative translation applications.

Significant improvements in MT quality have been achieved since its beginnings, but nevertheless, MWU treatment still presents important shortcomings. If MT intends to become a really useful tool adopted in multilingual everyday communication on the Internet, it has to tackle the problems posed by MWUs and provide an adequate processing approach to this ubiquitous lexical phenomenon which is statistically significant both in everyday and scientific texts. If it does not, it will fail to produce high quality natural output.

This work has presented the ongoing theoretical discussion concerning different aspects of MWUs such as their definition, properties and classification and illustrated the specific approach to MWUs which has been adopted as a basis for a proper linguistic formalisation of this particular linguistic phenomenon, i.e. Lexicon-Grammar. This linguistic approach provides the theoretical reference framework and foundational concepts for this work, having analysed MWUs and the difficulties they present to proper computational treatment in different types of NLP applications and for different languages since Gross's seminal paper on the representation of compound words (Gross, 1986).

Since different MWU processing methods have been used according to the different approaches in MT, one specific

chapter gives a broad and deep review of the different methodologies adopted in RBMT, EBMT, SMT and finally HMT.

The dissertation presents an MWU processing experiment based on linguistic knowledge which allows MWUs to be identified as single meaning units. Since the basic assumption of this work is that the integration of linguistic and probabilistic approaches can complement each other, the proposed method can be adopted either to improve MWU processing in SMT or become an important processing module in HMT.

Based on the Lexicon-Grammar theoretical framework, this experiment provides, on the one hand, an investigation of a broad variety of combinations of MWU types and an exemplification of their behaviour in texts extracted from different corpora and, on the other hand, a representation method that foresees the interaction of an electronic dictionary and a set of local grammars to efficiently handle different types of MWUs and their properties in MT as well as in other types of NLP systems.

This research work has therefore produced two main results in the field of MWU processing so far.

First of all, it has led to the development of a first version of an English-Italian electronic dictionary, specifically devoted to different MWUs types, as thoroughly described in Section 6.2.2. of this work.

Second, it has led to the analysis of a first set of specific MWU structures from a syntactic point of view and to the development of local grammars for the identification of continuous and discontinuous MWUs in the form of FST/FSA.

The whole work is based on a repeatable and extendable method based on linguistic resources that allow a deep

understanding of MWU lexical, syntactic and semantic structure in a translational setting. Probabilistic methods developed so far are not able to reach the same granularity as the one proposed in this work, in particular with respect to MWU with limited or no variability of co-occurrence among words.

A fine-grained linguistic analysis of all the different MWU types has a crucial role in developing effective processing methodologies that enable MT to be a true means of multilingual communication across the Internet for people speaking different languages. If next generation MT systems are able to produce more understandable and natural translations in the future, this will be thanks to a proper identification and translation of MWUs.

7.2. Future perspectives

For future work, we plan to further investigate MWUs from a Lexicon-Grammar perspective and in particular with respect to cross-linguistic asymmetries and translational equivalences.

Our long term goal is to integrate MWU treatment in either data-driven or hybrid approaches to MT in order to achieve high quality translation by combining probabilistic and linguistic information.

However, to achieve this goal, we must devise efficient strategies for representing deep attributes and semantic properties for MWUs in a cross-linguistic perspective.

Furthermore, we must consider both theoretical and practical aspects of the computational treatment of MWUs focusing on the new applicative settings in which MT is being used, i.e. social media such as Facebook, Twitter and

the like together with micro-blogs.

In conclusion, the focus of this research for the coming years will be to improve the results obtained so far and to extend the research work to provide a more comprehensive methodology for MWU processing in MT, taking into account not only the analysis phase but also the generation one.

Even if we are aware of the fact that it is unlikely that a computational method, whether it is data-driven or knowledge based, will be able to tackle this problem in all its complexity in the near future, nevertheless, we firmly believe that comprehensive and analytic linguistic resources will significantly improve current MWU processing strategies.

References

(1996). *The Language Engineering Directory – A resource guide to Language Engineering Organisations, products and services.* Madrid: Language and Technology, S.L.

Abeillé, A., Clément, L., & Toussenel, F. (2003). Building a treebank for French. In A. Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora.* Dordrecht: Kluwer.

Anastasiadis, M., Papadopoulou, L., & Gavrilidou, Z. (2011). Processing Greek frozen expressions with Nooj. In K. Vučković, B. Bekavac, & M. Silberztein, *Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference (Dubrovnik).* Newcastle: Cambridge Scholars Publishing.

Anastasiou, D. (2010). *Idiom treatment experiments in Machine Translation.* Newcastle: Cambridge Scholars Publishing.

Aoughlis, F. (2011). A French-English MT system for Computer Science Compound Words. In K. Vučković, B. Bekavac, & M. Silberztein, *Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference (Dubrovnik).* Newcastle: Cambridge Scholars Publishing.

Austermühl, F. (2001). *Electronic Tools for Translators.* Manchester: St. Jerome.

Austermühl, F. (2006). Training Translators to localize. In A. Pym, A. Perekrestenko, & B. Starink (eds.), *Translation*

Technology and its Teaching. Tarragona: Servei de Publicacions.

Aziz, W., Dymetman, M., Mirkin, S., & Specia, L. (2010). Learning an Expert from Human Annotations in Statistical Machine Translation: the Case of Out-of-Vocabulary Words. *Proceedings of EAMT.* Saint-Raphael, France.

Barreiro, A. (2008). *Make it simple with paraphrases. Automated paraphrasing for authoring aids and machine translation. PhD Dissertation.* Faculdade de Letras da Universidade do Porto, Oporto, Portugal.

Barreiro, A., Coheur, L., Luís, T., Costa, Â., Batista, F., Graça, J., & Trancoso, I. (forthcoming). "Multiword and Semantico-Syntactic Unit Alignment". *Language Resources and Evaluation*.

Barreiro, A., Monti, J., Elia, A., & Monteleone, M. (2010). Mixed up with Machine Translation: Multi-word Unit Disambiguation Challenge. *Translating and the Computer 32 ASLIB 18-19 November 2010.* London, England.

Barreiro, A., Scott, B., Kasper , W., & Kiefer, B. (2011). OpenLogos: Rule-Based Machine Translation: Philosophy, Model, Resources and Customization. *Machine Translation 25*, 107-126.

Bassnett-McGuire, S. (1980). *Translation studies.* London: Methuen.

Bayes, T., & Price, R. (1763). An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S.". Philosophical Transactions of the Royal Soci. *Philosophical Transactions of the Royal Society of London 53 (0)*, 370–418.

Biber, D., Conrad, S., & Reppen, R. (1999). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press.

Bouamor , D., Semmar , N., & Zweigenbaum, P. (2011). Improved statistical machine translation using multi-word expressions. *Proceedings of MT-LIHMT.* Barcelona, Spain.

Boulaknadel, S., Daille, B., & Aboutajd, D. (2008). A multi-term extraction program for arabic language. *Proceedings of LREC.* Marrakech, Morocco.

Britton, D. B., & McGonegal, S. (2007). *The Digital Economy Fact Book, Ninth Edition.* Washington, D.C.: The Progress & Freedom Foundation.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., . . . Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics, 16 (2)*, 79-85.

Brown, P. F., Cocke, J., Della Pietra, S., Della Pietra, V. J., Jelinek, F., Mercer, R. L., & Roossin, P. S. (1988). A statistical approach to language translation. *Coling 88: Proceedings of the 12th conference on Computational linguistics, volume 1*, (p. 71-76). Budapest, Hungry.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19(2)*, 263–312.

Brundage, J., Kresse, M., Schwall , U., & Storrer, A. (1992). *Multi-word lexemes: A monolingual and contrastive typology for NLP and MT. Technical Report IWBS 232.* Heidelberg: IBM Deutschland GmbH, Institut fur Wissenbasierte Systeme.

Callison-Burch, C., Koehn, P., & Monz, C. (2009). Findings of the 2009 Workshop on Statistical Machine Translation.

Proceedings 4th EACL Workshop on Statistical Machine Translation , (p. 1-28). Athens.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. *Proceedings of the 3rd International Conference on Language Resources and Evaluation LREC 2002*, (p. 1934-1940). Las Palmas.

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T., & Frey, J. (2006). Context-based Machine Translation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas.* Cambridge.

Carpuat, M., & Diab, M. (2010). Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. *HLT-NAACL 2010.*

Carpuat, M., & Wu, D. (2007). Improving statistical machine translation using word sense disambiguation. *Proceedings od the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, (p. 61-72).

Carson-Berndsen, J., Somers, H., Vogel, C., & Way, A. (2010). Integrated language technology as a part of next generation localization. *Localization Focus, 8 (1)*, 53-66. Retrieved 1st February, 2012 from http://www.localization.ie/resources/locfocus/LF_Vol%208%20Issue%201.pdf.

Cherry, C. (1966). *On human communication. (2nd ed.).* Cambridge, MA: The MIT Press.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. *Proceedings of Association of Computational Linguistics (ACL).*

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, Massachusetts: MIT Press.

Choueka, Y. (1998). Looking for needles in a haystack or locating interesting collocational expressions in large textual database. *Proceedings of the RIAO*, (p. 38-43).

Claveau, V. (2009). Translation of biomedical terms by inferring rewriting rules. In V. Prince, & M. Roche (eds.), *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration.* IGI - Global.

Corpas Pastor, G., & Varela Salinas (eds.), M. J. (2003). *Entornos informáticos de la traducción profesional: las memorias de traducción.* Granada: Atrio.

Dagan, I., & Church, K. W. (1994). Termight: Identifying and translating technical terminology. *Proceedings of the 4th Conference on ANLP*, (p. 34-40). Stuttgart, Germany.

D'Agostino, E., & Elia, A. (1998). Il significato delle frasi: un continuum dalle frasi semplici alle forme polirematiche. In AA.VV., *Ai limiti del linguaggio* (p. 287-310). Bari: Laterza.

Daille, B. (2001). Extraction de collocation a partir de textes. *Actes de TALN 2001 (Traitement automatique des langues naturelles).* Tours: ATALA, Université de Tours.

Danlos, L. (1992). Support Verb Constructions: Linguistic Properties, Representation, Translation. *French Language Studies, 2 (1)*, 1-32.

De Angelis, A. (1989). Nominalizations with the Italian support verb avere. *Lingvisticae Investigationes 13(2)*, 223-238.

De Mauro, T. (1999-2007). Introduzione. In *GRADIT, vol. 1º* (p. VII-XLII).

De Mauro, T. (2000). *Dizionario della lingua italiana.* Milano: Paravia.

Désilets, A. (2011). *Wanted: Best Practices for Collaborative Translation, TAUS Report.* Retrieved 1st February 2012 from: http://www.translationautomation.com/best-practices/wanted-best-practices-in-collaborative-translation.html.

Diab, M., & Bhutada, P. (2009). Verb Noun Construction MWE Supervised Token Classification. *Proceedings of ACL-IJCNLP 2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications.*

Diaconescu, S. (2004). Multiword Expression Translation Using Generative Dependency Grammar. *Advances in Natural Language Processing 4th International Conference, EsTAL 2004, October 20-22*, (p. 243-254). Alicante, Spain.

Dias, G., Guilloré, S., & Pereira Lopes, G. (1999). Multilingual aspects of multi-word lexical units. *Proceedings of the Workshop Language Technologies — Multilingual Aspects*, (p. 11-21). Ljubljana.

Dugast, L., Sellenart, J., & Koehn, P. (2007). Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. *Proceedings of the Second Workshop on Statistical Machine Translation* (p. 220-223). Prague: Association for Computational Linguistics.

Dugast, L., Senellart, J., & Koehn, P. (2009). Statistical Post Editing and Dictionary Extraction: Systran. *Proceedings of the Fourth Workshop on Statistical Machine Translation* (p. 110-114). Edinburgh: Association for Computational Linguistics.

Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., . . . Chen, Y. (2008). Hybrid Machine Translation Architectures within and beyond the

EuroMatrix project. *Proceedings 12th European Machine Translation Conference.* Hamburg.

Elia, A., D'Agostino, E., & Martinelli , M. (1985). Tre componenti della sintassi italiana: frasi semplici, frasi a verbo supporto e frasi idiomatiche. *Proceedings of the 17th International Conference of SLI* (p. 311-325). Roma: Bulzoni.

Elia, A., Monteleone, M., Monti, J., & Marano, F. (2011). Linguistically motivated knowledge management: exploitation of language resources for NLP applications. *Proceedings 30th International Conference on Lexis and Grammar.* Nicosia, Cypre.

Elia, A., Postiglione, A., Monteleone, M., Monti, J., & Guglielmo, D. (2011). CATALOGA®: a software for semantic and terminological information retrieval. *Proceedings of on Web Intelligence, Mining and Semantics (WIMS'2011).* Songdal, Norway.

Eskelsen , G., Marcus, A., & Fereee, K. W. (2008). *The Digital Economy Fact Book, Tenth edition.* Washington, D.C.: The Progress & Freedom Foundation.

Esselink, B. (2000). *A Practical Guide to Localization.* Amsterdam & Philadelphia: John Benjamins.

Fillmore, C. J. (2003). Multiword expressions, November. Invited talk at the Institute for Research in Cognitive Science (IRCS), University of Pennsylvania.

Fillmore, C., Kay, P., & O'Connor, C. (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of let alone. *Language 64*, 501–38.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1-32.

Franz, A., Horiguchi , K., Duan, L., Ecker, D., Koontz, E., & Uchida, K. (2000). An integrated architecture for Example-based Machine Translation. *18th COLING 2000*, (p. 1031-1035). Saarbrücken, Germany.

Gangadharaiah, R., & Balakrishnan, N. (2006). Application of Linguistic Rules to Generalized Example Based Machine Translation for Indian Languages. *In First National Symposium on Modeling and Shallow Parsing of Indian Languages, (MSPIL).* Mumbai, India.

Giry-Schneider, J. (1978). *Les nominalisations en français. L'opérateur faire dans le lexique.* Genève: Droz.

Giry-Schneider, J. (1987). *Les prédicats nominaux en français. Les phrases simples à verbe support.* Genève: Droz.

Giry-Schneider, J. (2005). Les noms épistémiques et leurs verbes supports. *Lingvisticæ Investigationes 27(2)*, 219-238.

Goutte, C., Yamada, K., & Gaussier, E. (2004). Aligning words using matrix factorization. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL '04), Main volume*, (p. 502-509). Barcelona, Spain.

Green, S., de Marneffe, M.-C., Bauer, J., & Manning, C. D. (2011). Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. *Empirical Method for Natural Language Processing (EMNLP'11).*

Gross, M. (1968). L'emploi des modèles en linguistique. *Langages 9*, 3-8.

Gross, M. (1975). *Méthodes en syntaxe.* Paris: Hermann.

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

Gross, M. (1984). Lexicon-Grammar and the Syntactic Analysis of French. *Proceedings of Coling*, (p. 275-282). Stanford.

Gross, M. (1986). Lexicon-Grammar. The representation of compound words. *Proceedings of COLING '86.* Bonn: University of Bonn, http://acl.ldc.upenn.edu/C/C86/C86-1001.pdf.

Gross, M. (1989). La construction de dictionnaires électroniques. *Annales des Télécommunications, 44 (1-2)*, 4-19.

Gross, M., & Senellart, J. (1998). Nouvelles bases statistiques pour les mots du français. *In 4`emes Journées internationales d'Analyse statistique des Donnees Textuelles (JADT'98)*, (p. 335-349). Nice.

Gross, M., Halle, M., & Schutzenberger (eds.), M.-P. (1973). *The Formal analysis of natural languages. Proceedings of the first international conference.* Paris: The Hague.

Groves, D., Hearne, M., & Way, A. (2004). Robust sub-sentential alignment of phrase-structure trees. *Proceedings of the 20th international conference on Computational Linguistics (COLING)*, (p. 1072-es). Geneva, Switzerland .

Guenthner, F., & Blanco, X. (2004). Multilexemic expressions: an overview. In C. Lèclere, E. Laporte, M. Piot, & M. Silberztein (eds.), *Syntax, Lexis, and Lexicon-Grammar* (p. 239-252). Amsterdam/New York: John Benjamins.

Han, S.-H. (2000). *Les prédicats nominaux en coréen: Constructions à verbe support hata.* PhD Thesis, University of Marne-la-Vallée.

Harris, Z. S. (1946). From Morpheme to Utterance. *Language, 22 (3)*, 161–183.

Harris, Z. S. (1957). Co-occurrence and transformation in linguistic structure. *Language, 33 (3)*, 283-340.

Harris, Z. S. (1962). *String Analysis of Sentence Structure.* Mouton: The Hague.

Harris, Z. S. (1964). Transformations in Linguistic Structure. *Proceedings of the American Philosophical Society, 108 (5)*, (p. 418-122).

Harris, Z. S. (1968). Mathematical Structures of Language. *Interscience Tracts in Pure and Applied Mathematics, 21*.

Harris, Z. S. (1970). *Papers in Structural and Transformational Linguistics.* Dordrecht/ Holland: D. Reidel.

Harris, Z. S. (1982). *A Grammar of English on Mathematical Principles.* New York: John Wiley & Sons.

Hatim, B., & Mason, I. (1990). *Discourse and the Translator.* London: Longman.

Hildebrand, A. S., & Vogel, S. (2008). Combination of Machine Translation Systems via Hypothesis Selection from Combined N-Best Lists. *MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas* (p. 254-261). Waikiki, Hawaii,: Association for Machine Translation in the Americas.

Hong, C.-S. (1991). Objet interne, verbe support et dictionnaire. *Proceedings of Computational Lexicography (Balatonfüred, Hungary, 1990)* (p. 93-102). Budapest: Hungarian Academy of Sciences.

Howe, J. (2006). The rise of crowdsourcing. *Wired, 14 (6)*.

Hurskainen, A. (2008). *Multiword Expressions and Machine Translation.* Technical Reports. Language Technology Report No 1.

Hutchins, J. (2005). Current commercial machine translation systems and computer-based translation tools: system

types and their uses. *International Journal of Translation, 7 (1-2)*, 5-38, http://www.hutchinsweb.me.uk/IJT-2005.pdf.

Hutchins, J. (2010). *Compendium of Translation Software. Commercial machine translation systems and computer-aided translation support tools.* European Association for Machine Translation. 15th edition.

Hutchins, J. (2010). *Compendium of Translation Software. Directory of commercial machine translation systems and computer-aided translation support tools.* Genève: European Association for Machine Translation.

Hutchins, J. W., & Somers, H. L. (1992). *An introduction to machine translation.* London: Academic Press.

Jackendoff, R. (1997). Twistin' the night away. *Language, 73*, 534-559.

Kelly, N., Ray, R., & De Palma, D. (2011). From crawling to sprinting: Community Translation goes mainstream. *Linguistica Antverpiensia, 10*, 75-96.

Kim, S. N., & Baldwin, T. (2010). How to pick out token instances of English verb-particle constructions. Lang. Res. & Eval. *Special Issue on Multiword expression: hard going or plain sailing, 44 (1-2)*, 97–113.

Kim, Y.-S., Chang, J.-H., & Zhang, B.-T. (2002). A Comparative Evaluation of Data-driven Models in Translation Selection of Machine Translation. *COLING '02 Proceedings of the 19th international conference on Computational Linguistics* (p. 1-7). Taipei: Association for Computational Linguistics.

Koehn, P. (2009). *Statistical machine translation.* Cambridge: Cambridge University Press.

Koehn, P., & Hoang, H. (2007). Factored translation models. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning (EMNLP-CoNLL)*, (p. 868-876).

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-based translation. *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, (p. 127-133). Edmonton, Canada.

Korkontzelos, I., & Manandhar, S. (2010). Can Recognising Multiword Expressions Improve Shallow Parsing? *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010).* Los Angeles, California.

Krenn, B., & Erbach, G. (1994). Idioms and Support Verb Constructions. In J. Nerbonne, K. Netter, & C. Pollard (eds.), *German in Head Driven Phrase Structure Grammar* (p. 365—395). Stanford CA: CSLI Publications.

Lambert, P., & Banchs, R. (2006). Grouping multi-word expressions according to Part-Of-Speech in statistical machine translation. *Proceedings of the EACL Workshop on Multi-word expressions in a multilingual context.* Trento, Italy.

Machonis, P. A. (2007). Look this up and try it out: an original approach to parsing phrasal verbs. *Actes du 26 Colloque international Lexique Grammaire, Bonifacio 2-6 octobre 2007.*

Machonis, P. A. (2008). NooJ: a practical method for Parsing Phrasal Verbs. *Proceedings of the 2007 International NooJ Conference.* (p. 149-161). Newcastle: Cambridge Scholars Publishing.

Macleod, C., Grishman, R., Meyers, A., Barret, L., & Reeves, R. (1998). NOMLEX: A Lexicon of Nominalizations. *Proceedings of EURALEX '98*. Liege, Belgium.

Macleod, C., Meyers, A., Grishman, R., Barrett, L., & Reeves, R. (1997). Designing a Dictionary of Derived Nominals. *Proceedings of Recent Advances in Natural Language Processing.* Tzigov Chark - Bulgaria.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, USA: MIT Press.

Marcu, D., & Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Marcu, D., Wei , W., Echihabi, A., & Knight, K. (2006). SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. *Proceedings of Empirical Methods in Natural Language Processing (EMNLP).*

Mel'čuk, I. (1998). Collocations and lexical functions. In P. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications* (p. 23-53). Oxford: Claredon Press.

Mel'čuk, I. A. (1996). 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* (p. 37-102). Amsterdam/Philadelphia: Benjamins.

Mel'čuk, I. A., Clas, A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire.* Louvain-la-Neuve: Duculot.

Monti, J. (2004). Dal sogno meccanico alla e-translation: la traduzione automatica è realtà. *Media Duemila – Mensile di Cultura informatica e ICT 219*, 60-67.

Monti, J. (2007). Localizzazione: il ruolo e i saperi della traduzione. In G. Marchesini, & C. Montella (eds.), *I saperi del tradurre. Analogie, affinità e confronti* (p. 173-197). Milano: Franco Angeli.

Monti, J. (2010). La E-translation da Google a Second Life: le più recenti applicazioni di Traduzione Automatica online. In G. Massariello Merzagora, & S. Dal Maso, *I luoghi della traduzione. Le interfacce* (p. 545-552). Roma: Bulzoni Editore.

Monti, J. (2012). Translators'knowledge in the cloud: the new translation technologies. *Proceedings of the International Symposium on Language and Communication: Research trends and challenges (ISLC 2012).* Izmir - Turkey.

Monti, J., Barreiro, A., Elia, A., Marano, F., & Napoli, A. (2011). Taking on new challanges in multi-word unit processing for Machine Translation. *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation.*, (p. 11-19). Barcelona, Spain.

Monti, J., Elia, A., Monteleone, M., Postiglione, A., & Marano, F. (2011). In search of knowledge: text mining dedicated to technical translation. *Proceedings of Translating and the Computer 33 ASLIB Conference - 18 November 2011.* London.

Moszczyński, R. (2007). A Practical Classification of Multiword Expressions. *ACL '07 Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop.* Association for Computational Linguistics .

Moszczyński, R. (2010). Towards a bilingual lexicon of information technology multiword units. *Proceedings of the XIV Euralex International Congress.* Leeuwarden, the Netherlands.

Nagao, M. (1984). A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In A.

E. (eds.), *Artificial and Human Intelligence* (p. 173-180). Amsterdam: North- Holland.

Newmark, P. (1981). *Approaches to translation.* Oxford: Pergamon.

Newmark, P. (1988). *A Textbook of Translation.* New York: Prentice Hall.

Nida, E. A., & Taber, C. R. (1969). *The theory and practice of translation.* Leiden: Brill.

Nirenburg, S., Somers, H., & Wilks , Y. (2003). *Readings in machine translation.* Cambridge, Mass.: The MIT Press.

Nomiyama, H. (1992). Machine Translation by Case generalisation. *14th COLING 1992*, (p. 714-720). Nantes, France.

O'Brien, S. (2002). Teaching post-editing: A proposal for course content. *6th EAMT Workshop Teaching Machine Translation* (p. 99-106). Manchester: Retrieved 1st February, 2012 from http://www.mt-archive.info/EAMT-2002-OBrien.pdf.

Okita, T., Guerra, A. M., Graham, Y., & Way, A. (2010). Multi-Word Expression Sensitive Word Alignment. *Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010*, (p. 26-34). Beijing.

Okuma, H., Yamamoto, H., & Sumita , E. (2008). Introducing a Translation Dictionary into Phrase-Based SMT. *IEICE Transactions 91-D (7)*, 2051-2057.

Orliac , B., & Dillinger, M. (2003). Collocation extraction for machine translation. *Proceedings of Machine Translation Summit IX*, (p. 292-298). New Orleans, Lousiana, U.S.A.

Ozdowska, S. (2006). *ALIBI, un système d'ALIgnement BIlingue à base de règles de propagation syntaxique.* Toulouse, France: Ph.D. thesis, Université de Toulouse.

Pal, S., Chakraborty, T., & Bandyopadhyay, S. (2011). Handling Multiword Expressions in Phrase-Based Statistical Machine Translation. *Machine Translation Summit XIII* , (p. 215-224). Xiamen, China.

Pal, S., Sudip, K. N., Pavel, P., Sivaji, B., & Way, A. (2010). Handling Named Entities and Compound Verbs in Phrase- Based Statistical Machine Translation. *Proceedings of the workshop on Multiword expression: from theory to application (MWE-2010) The 23rd International conference of computational linguistics (Coling 2010),Beijing, Chaina, pp. 46-54 (2010)*, (p. 46-54). Beijing, China.

Pym, A. (2003). Redefining Translation Competence in an Electronic Age. In Defense of a Minimalist Approach. 481-497. *Meta, 48 (4)*, 481-497.

Ramisch, C., Villavicenzio, A., & Boitet, C. (2010). Multiword expressions in the wild? the mwetoolkit comes in handy. *Proceedings of the 23rd COLING (COLING 2010) — Demonstrations*, (p. 57–60). Beijing, China.

Ranchhod, E. (1989). Predicative Nouns and Negation. *Lingvisticae Investigationes 13(2)*, 387-397.

Ranchhod, E. (1990). *Sintaxe dos predicados nominais com estar.* Lisbon: INIC.

Rayson, P., Piao, S., Sharoff, S., Evert, S., & Villada Moirón, B. (2010). Multiword expressions: hard going or plain sailing? *Journal of Language Resources and Evaluation. Lang Resources & Evaluation 44*, 1–5.

Ren, Z., Lü, Y., Cao, J., Liu, Q., & Zhixiang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, (p. 47-54). Singapore.

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multi-word expressions: A pain in the neck for NLP. *Proc. International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, (p. 1-5). Mexico City, Mexico.

Salkoff, M. (1999). *A French-English Grammar: A Contrastive Grammar On Translational Principles.* Amsterdam, Philadelphia: John Benjamins Publishing Company.

Salkoff, M. (forthcoming). *Loquatur! being a program for LOw Quality Automatic Translation of Unrestricted Range.*

Schwenk, H., Abdul-Rauf, S., Barrault, L., & Senellart, J. (2009). SMT and SPE machine translation systems for WMT'09. *Forth ACL Workshop on Statistical Machine Translation*, (p. 130-134).

Scott, B. (2003). The Logos Model: An Historical Perspective. *Machine Translation 18*, 1-72.

Segura, J., & Prince, V. (2011). Using Alignment to detect associated multiword expressions in bilingual corpora. *Tralogy [En ligne], Session 6 - Translation and Natural Language Processing / Traduction et traitement automatique des langues (TAL).* http://lodel.irevues.inist.fr/tralogy/index.php?id=144&format=print.

Seretan, V. (2009). Extraction de collocations et leurs équialents de traduction à partir de corpus parallèles ». *TAL, 50*, 305–332.

Seretan, V., & Wehrli, E. (2007). Collocation translation based on sentence alignment and parsing. *Actes de TALN 2007 (Traitement automatique des langues naturelles).* Toulouse: ATALA, IRIT.

Setiawan, H., Li, H., & Zhang, M. (2005). Learning phrase translation using level of detail approach. *Proceedings of*

the Tenth Machine Translation Summit (MT Summit X). Phuket, Thailand.

Shin, J. H., Han, Y. S., & Choi, K.-S. (1996). Bilingual knowledge acquisition from Korean-English parallel corpus using alignment. *Proceedings of the 16th International Conference on Computational Linguistics (COLING).*

Shin, K.-S. (1994). *Le verbe support hata en coréen contemporain : morpho-syntaxe et comparaison.* PhD thesis, University of Paris 7.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes.* Paris: Masson.

Silberztein, M. (2002). *NooJ Manual.* Available for download at: www.nooj4nlp.net.

Silberztein, M. (2005). NooJ's Dictionaries. *Proceedings of 2nd Language and Technology Conference.* Poznan: Poznan University.

Silberztein, M., Koeva, S., & Maurel (eds.), D. (2007). *NooJ"s Linguistic Annotation Engine. In Formaliser les langues avec l'ordinateur : de INTEX à NooJ. Cahiers de la MSH Ledoux.* Besançon: Presses Universitaires de Franche-Comté.

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

Sumita, E., & Iida, H. (1991). Experiments and prospects of Example-based Machine Translation. *29th Annual Meeting of the ACL 1991*, (p. 185-192). Berkley California.

Sumita, E., Iida, H., & Kohyama, H. (1990). Translating with examples: A new approach to Machine Translation. *3rd TMI 1990*, (p. 203-212). Texas, USA.

Tambouratzis, G., Troullinos, M., Sofianopoulos, S., & Vassiliou, M. (2012). Accurate phrase alignment in a bilingual corpus for EBMT systems. *Proceedings of the 5th Workshop on Building and Using Comparable Corpora (BUCC2012) [held in conjunction with LREC2012]*, (p. 104-111). Istanbul, Turkey.

Thurmair, G. (2004). Multilingual content processing. *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC).* Lisbon, Portugal.

Thurmair, G. (2009). Comparing different architectures of hybrid Machine Translation systems. *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, (p. 340-347). Ottawa, Ontario, Canada.

Tillmann, C. (2003). A projection extension algorithm for statistical machine translation. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, (p. 1-8).

Tillmann, C., & Xia, F. (2003). A Phrase-based Unigram Model for Statistical MachineTranslation. *Companian Vol. of the Joint HLT and NAACL Conference (HLT 03)*, (p. 106-108). Edmonton, Canada.

Torres del Rey, J. (2005). *La interfaz de la traducción. Formación de traducctores y nuevas tecnologías.* Granada: Comares.

Vandeghinste, V., Dirix, P., Schuurman, I., Sofianoupulos, S., Sofianopoulos, S., Vassiliou, M., . . . Schmidt, P. (2008). Evaluation of a Machine Translation System for Low Resource languages: METIS-II. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, (p. 449-456). Marrakech.

Vandeghinste, V., Schuurman, I., Markantonatou, S., & Badia, T. (2006). METIS-II: Machine Translation for Low Resource Languages. *Proceedings LREC.* Genoa.

Váradi, T. (2006). Multiword Units in an MT Lexicon. *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts* (p. 73-78). Trento, Italy: Association for Computational Linguistics.

Vietri, S. (1996). The Syntax of the Italian verb essere Prep. *Lingvisticae Investigationes*, 287-363.

Vietri, S. (2008). *Dizionari elettronici e grammatiche a stati finiti. Metodi di analisi formale della lingua italiana.* Cava dei Tirreni : Plectica.

Villavicencio, A. B. (2005). Introduction to the special issue on multiword expressions: having a crack at a hard nut. *Journal of Computer Speech and Language Processing, 19(4)*, 365-377.

Vinay, J. P., & Darbelnet, J. (1958). *Stylistique comparée du français et de l'anglais. Méthode de traduction.* Paris: Dider.

Vintar, Š., & Fišer, D. (2008). Harvesting multi-word expressions from parallel corpora. *Procceedings of LREC.* Marrakech, Morocco.

Wilks, Y. (2009). *Machine translation: its scope and limits.* New York: Springer.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching*, 213-231.

Wu, C.-C., & Chang, J. S. (2004). Bilingual collocation extraction based on syntactic and statistical analyses. *In Computational Linguistics*, 1-20.

Wu, H., Wang, H., & Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. *Proceedings of Conference on Computational Linguistics (COLING)*, (p. 993–1000).

Yamada, K., & Knight, K. (2001). A syntax-based statistical translation model. *Proceeding ACL '01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, (p. 523-530).

Yamamoto, K., Kudo, T., Tsuboi, Y., & Matsumoto, Y. (2003). Learning sequence-to-sequence correspondences from parallel corpora via sequential pattern mining. *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond.* Association for Computational Linguistics.

Yang, J., & Lange, E. D. (1998). Systran on AltaVista: A User Study on Real-Time Machine Translation on the Internet. In D. Farwell, L. Gerber, & E. Hovy (eds.), *Machine Translation and the Information Soup* (p. 275-285). Berlin: Springer.

Yang, J., & Lange, E. D. (2003). Going Live on the Internet. In H. Somers (ed.), *Computers and Translation: A Translator's Guide* (p. 191-210). Amsterdam: John Benjamins.

Zens, R., Och, F. J., & Ney, H. (2002). Phrase-based statistical machine translation. *Proceedings of the German Conference on Artificial Intelligence (KI 2002).*

Zhang, Y., & Vogel, S. (2005). Competitive grouping in integrated phrase segmentation and alignment model. *Proceedings of the ACL workshop on Building and Using Parallel Texts* (p. 159-162). Association for Computational Linguistics.

Zhang, Y., Vogel, S., & Waibel, A. (2003). Integrated phrase segmentation and alignment algorithm for statistical machine translation. *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE '03).* Beijing, China.

Zhao, B., & Vogel, S. (2005). A generalized alignment-free phrase extraction. *Proceedings of the ACL Workshop on Building and Using Parallel Texts* (p. 141-144). Association for Computational Linguistics.

Zollmann, A., & Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. *In Proceedings of the Workshop on Statistical Machine Translation, HLT/NAACL.*