Translation asymmetries of multiword expressions in machine translation

An analysis of the TED-MWE corpus

Johanna Monti[1], Mihael Arcan[2] and Federico Sangati[1]
[1]Università degli Studi di Napoli "L'Orientale" / [2]Insight Centre for Data Analytics

Machine Translation (MT) is now extensively used both as a tool to overcome language barriers on the internet and as a professional tool to translate technical documentation. The technology has rapidly evolved in recent years thanks to the availability of large amounts of data in digital format and in particular parallel corpora, which are used to train Statistical Machine Translation (SMT) tools. The quality of MT has considerably improved but the translation of multiword expressions (MWEs) still represents a big and open challenge, both from a theoretical and a practical point of view (Monti, 2013). We define MWEs as any 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* (soul mate), *carta di credito* (credit card), *acqua e sapone* (water and soap), *piovere a catinelle* (rain cats and dogs). The persistence of mistranslation of MWEs in MT outputs originates from their lexical, syntactic, semantic, pragmatic but also translational idiomaticity. Therefore, there is a need to invest in further research in order to achieve significant improvements MT and translation technologies. In particular, it is important to develop resources, mainly MWE-annotated corpora, which can be used for both MT training and evaluation purposes (Monti and Todirascu, 2016).

This work focuses on the translation asymmetries between English and Italian MWEs, and how they affect the SMT performance. By translation asymmetries we mean the differences which may occur between an MWE in a source language and its equivalent in the target language, like in many-to-many word translations (En. *to be in a position to* → It. *essere in grado di*), many-to-one (En. *to set free* → It. *liberare*) and finally one-to-many correspondences (En. *overcooked* → It. *cotto troppo*). This chapter describes the evaluation of mistranslations caused by translation asymmetries concerning multiword expressions detected in the TED-MWE corpus (http://tiny.cc/TED_MWE), which contains 1,500 sentences and 31,000 EN tokens. This corpus is a subset of the

https://doi.org/10.1075/ivitra.24.02mon

TED spoken corpus (Monti et al., 2015) annotated with all the MWEs detected during the evaluation process. The corpus contains the following information: (i) the English source text, (ii) the Italian human translations (from the parallel corpus), and (iii) the Italian SMT output. All the annotators were Italian native speakers with a good knowledge of the English language and with a background in linguistics and computational linguistics. They were asked to identify all MWEs in the source text together with their translations in approximately 300 random sentences each and to evaluate the automatic translation correctness. The identified MWEs and the evaluation of both the human and the machine translation are also recorded in the corpus. This chapter will discuss (i) the related work concerning the impact of anisomorphism (the absence of an exact correspondence between words in two different languages) and the consequent translation asymmetries on MWEs translation quality in MT, (ii) the corpus, (iii) the annotation guidelines, (iv) the methodology adopted during the annotation process (Monti et al., 2015), (v) the results of the annotation and finally (vi) the evaluation of translation asymmetries in the corpus and ideas for future work.

Keywords: machine translation, translation asymmetries, multiword expressions, TED-MWE corpus

1. Introduction

Multiword expressions, i.e. groups of two or more words that convey a single, usually non-compositional meaning, such as *credit card*, *get off*, *European Union*, *pay attention*, still represent a true bottleneck in Natural Language Processing (NLP), Machine Translation (MT) and Translation Technology (TT), despite the remarkable advances achieved in these fields in recent years. MWEs are very frequent and productive linguistic phenomena both in everyday language and in language for special purposes. In addition, they 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. MWEs represent, therefore, a very frequent source of mistranslations in MT because of intrinsic ambiguities, structural complexity, lexical asymmetries between languages and, finally, cultural differences (Monti, 2014).

Processing and translating MWEs is a crucial task in many NLP applications such as multilingual terminology extraction, machine translation (MT), cross-lingual information retrieval (CLIR) and cross-language information extraction (CLIE) among others. In particular, CLIR and CLIE success in retrieving relevant information relies on the quality of MT (Fu et al., 2009) and therefore inaccurate or incorrect translations may cause serious problems.

Even the dominant paradigm, SMT, and also the more recent neural machine translation (NMT) technology face several difficulties in translating these types of constructions, since they tend to translate on a word-by-word basis and are not able to reconstruct the intended meaning, as it can be easily verified using the available online MT systems. For instance, if we translate the English sentence, "*Every kid in the world is the apple of their parents' eye*." into Italian with Google Translate, which is now based on the neural approach, the result (https://translate.google.it/?hl=it as of June 2018) is the following: "*Ogni bambino al mondo è la mela dell'occhio dei loro genitori*". Here, the meaning of the idiomatic MWE *to be the apple of someone's eye(s)* is non-compositional and corresponds to the Italian idiomatic MWE *essere la luce degli occhi di qualcuno*, but the translation system is not able to translate it correctly.

MT has enormously improved in the last decades, but processing and translating MWE still represents one of the most important challenges. The traditional word-based alignment approach, following IBM Models (Brown et al., 1993), shows many shortcomings related to MWE processing, especially due to its 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 MWEs properly. Figure 1 presents a typical MWE misalignment in a word-based SMT system, namely Giza++ (Och and Ney, 2003).

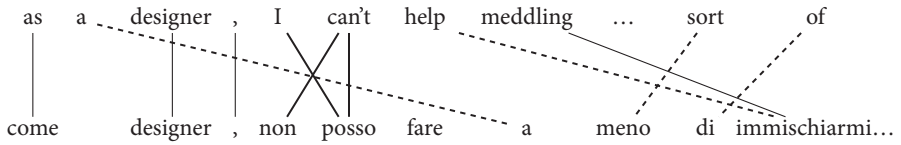

Figure 1. Example of a GIZA++ misalignment between the English MWE *I can't help* and its Italian MWE translation *non posso fare a meno di* (lit. not can do to less than). Dotted lines are indicating incorrect alignments, and tick lines (both continuous and dotted) are those adjacent to MWE tokens in the source or target sentence

The phrase-based (PB) alignment approach (Koehn et al., 2003) is better at dealing with MWEs as it considers many-to-many word alignments. However, many combinations of words or n-grams have no linguistic significance (*the war*), while others are linguistically meaningful (*cold war*). In the widely used PB-SMT systems, phrases are sequences of contiguous words, which are not linguistically motivated and do not implicitly capture all useful MWE information, although they are able to translate contiguous MWEs and sometimes also discontiguous ones. The correct translation of MWEs occurs on a statistical basis if the constituents of MWEs are aligned as parts of consecutive phrases (n-grams) in the training set.

Furthermore, MWEs are not generally treated as a special case when correspondences between source and target language do not consist of consecutive many-to-many source-target correspondences. MWE processing and translation within SMT started being addressed only very recently and different solutions have been proposed so far, but they are basically considered either as a problem of automatically learning and integrating translations into an SMT system or as a problem of word alignment. The most used methodologies are identification of possible monolingual MWEs. This phase can be accomplished using different approaches, by means of (i) morpho-syntactic patterns (Okita and Way, 2010; Dagan and Church, 1994), (ii) statistical methods (Vintar and Fišer, 2008) and finally (iii) hybrid approaches (Wu and Chang, 2004; Seretan and Wehrli, 2007; Boulaknadel et al., 2008; Daille, 2001).

Furthermore research was performed on the extraction of the equivalent translations of the identified monolingual MWEs according to the different alignment methodologies. Current approaches to MWE processing integrate phrase-based models with linguistic knowledge, such as hand-crafted dictionaries and grammars or data-driven ones, in order to identify and process MWEs as single units.

Finally, the new neural approach to MT in which a large neural network is trained by deep learning techniques is still in its pioneering stage and little has been reported about the improvements it can bring to MWE processing and translation.

One of the main problems in translating MWEs is represented by their translation idiomaticity, i.e. it is not usually possible to translate MWEs literally. In addition to that, their internal structure may greatly vary from one language to another one. This property, which goes under the name of non-literal translatability, means that an MWE cannot be translated from one language to another on a word-for-word basis (Sag et al., 2002; Barreiro, 2008; Monti, 2012), and is characteristic of the majority of MWEs, in particular those with limited or no variation of distribution of their internal constituents. This is the case for idioms (e.g. *it's raining cats and dogs!* → It. **sta piovendo cani e gatti*), but also of many collocations (e.g. *heavy rain* → It. **pioggia pesante*), fixed expressions (e.g. *by and large* → It. **da e largo*), proverbs (e.g. *there's no such thing as a free lunch* → It. **non esiste una cosa come un pranzo gratuito*) and phrasal verbs (e.g. *bring somebody down* → It. **portare qualcuno giù*) amongst others.

The anisomorphism between languages leads to translation asymmetries, i.e. the differences which may occur between a MWE in its source language and its translation, like in many-to-many translations (En. *to be in a position to* → It. *essere in grado di*) but also in many-to-one (En. *to set free* → It. *liberare*) and one-to-many (En. *overcooked* → It. *cotto troppo*) correspondences.

MWEs are sometimes discontinuous, i.e. it is possible to insert an element between the constituents of a multiword. As an example, it is possible to insert an NP into the verbal MWE *take into account* as in *take something into account*.

Translation asymmetries are one of the main sources of mistranslations in MT and one of the possible solutions to this problem is to develop large linguistic resources, mainly MWE-annotated corpora, which can be used both for MT training and evaluation purposes (Monti and Todirascu, 2016).

This chapter presents the results of the English-Italian TED-MWE project and in particular: (i) the related work, (ii) the MWE-TED corpus, the annotation guidelines and methodology, (iii) the results of the experiment and finally (iv) the evaluation of the translation asymmetries and the mistranslation in the TED-MWE corpus.

2. Related work

Studies on translation asymmetries and their impact on MT quality are underrepresented in recent NLP studies. The definition of Translation asymmetries in MT can be dated back to Pause's paper on *Interlingual strategies in translation* (1997), but the concept of a source language structure translated with a different structure in the target language was already discussed in Dorr (1994), who classified human and MT *divergences* in six different types. Dorr identifies a specific class for MWE translation, namely *Conflational or Inflational Divergence*. A conflational divergence is when two or more words in the source language are translated by one word in the target language. The inflational divergence, instead, arises when one word in the source language is translated by two or more words in the target language.

This classification has been used in Lin et al. (2005), Mahesh et al. (2005) and lately by Kauffmann (2013), who devotes a few words to the problem of MWEs in conflational divergences. According to Kauffman, large-scale monolingual lexicons of multi-word expressions (and collocations) and bilingual lexicons that record their translations represent a possible solution to the processing of such divergences in MT. MWE multilingual lexicons as well as parallel corpora annotated with MWEs represent invaluable linguistic resources for MWE processing and translation, but recent surveys (Constant et al., 2017 and Losnegaard et al., 2016) have highlighted that these types of resources are still lacking and this fact may hinder research both on the translation of MWEs across languages and NLP involving two or more languages.

Translation asymmetries represent an important clue as to the occurrence of MWEs in parallel corpora and are at the heart of a few studies which aim to detect MWEs using unsupervised or semi-unsupervised methods. Melamed (1997) develops a method for the discovery of MWE on the basis of their translational entropy in parallel corpora. A statistically-driven alignment-based approach to MWE identification in technical corpora, including parallel corpora, is shown in Caseli et al.

(2009); they examine how a second language can provide relevant clues for this tasks and extract sequences of length 2 or more in the source language that are aligned with sequences of length 1 or more in the target (m:n alignments). Bouamor et al. (2012) address non-compositional contiguous MWE sequences and present a method combining linguistic and statistical information to extract and align MWEs in a French-English parallel corpus. The extracted bilingual MWEs are integrated into MOSES to show that MT quality can be improved by the use of such units. In recent years, different approaches have been adopted with reference to MWE identification from the translational asymmetries (misalignments) in parallel corpora, such as Lambert (2005), who use an asymmetry-based approach and focus on alignment sets in which source-to-target links proposed by Giza++ are different from target-to-source alignments, or Tsvetkov and Wintner (2010), who focus on misalignments to develop an unsupervised algorithm for identifying MWEs in (small) bilingual corpora, using automatic word alignment extraction of MWEs of various types, lengths along with their translations. Other works are based on extraction of bilingual MWEs, such as Thurmair and Aleksić (2012), who extract terms and lexicon entries directly from SMT translation models, or Arcan et al. (2017), who propose a framework for extracting bilingual terms from a post-edited corpus and using them to enhance the performance of an SMT system embedded in a collaborative CAT environment. Moirón and Tiedemann (2006) focus on Dutch expressions and their English, Spanish, and German translations in the Europarl corpus (Koehn, 2005). MWE candidates are ranked by the variability of their constituents' translations. To extract the candidates, they use syntactic properties (based on full parsing of the Dutch text) and statistical association measures. Sangati and van Cranenburgh (2015) focus on identification and extraction of MWEs from a large set of recurring syntactic fragments from a given treebank. They use these fragments to identify MWEs as a parsing task (in a supervised manner) and compare various association measures in re-ranking the expressions underlying these fragments in an unsupervised fashion.

3. The TED-MWE corpus

Annotated parallel corpora are a very important resource for MT, but to present there are only very few small-sized corpora, containing, aligned sentences representative of a specific type of MWE and for a limited number of language pairs, which are also very difficult to reuse in research settings different from the original ones (Monti and Todirascu, 2016). To our knowledge, none of the corpus resources developed so far encode multiword expressions of all different types in a parallel corpus. Therefore we developed the TED-MWE corpus, which is based on the

web inventory named WIT3 (https://wit3.fbk.eu), a collection of transcribed and translated talks (Cettolo et al., 2012). The core of WIT3 is the TED Talks corpus that basically redistributes the original content published by the TED Conference website. Since 2007, the TED Conference posted all video recordings of its talks together with subtitles in English: almost all talks have been translated by volunteers into more than 80 languages and the translated talks range in number from several hundred (e.g. such as for the Dutch, German, Hebrew, Romanian languages) to just one (e.g. for Hausa, Hupa, Bislama, Ingush, Maltese). The WIT3 corpus re-purposes the original content in a way that is more convenient for MT researchers. For our experiments, we used the 2014-released WIT3 TED data set for the English-Italian language pair, which contains the training data of 190,000 parallel sentences, needed to build an SMT system. In addition, we used the 2014 TED development set (1,000 sentences) and the 2010/2011/2012 test sets (1,500 sentences each).

The TED-MWE corpus is the result of the annotation of the English-Italian WIT3 TED data set with MWEs of different types. Besides the WIT3 English-Italian parallel corpus, the TED-MWE corpus also contains the Italian output for the English source sentences obtained using the statistical translation toolkit Moses (Koehn et al., 2007), where the word alignments were built with the GIZA++ toolkit (Och and Ney, 2003). The IRSTLM toolkit (Federico et al., 2008) was used to build the 5-gram language model. The parameters within the SMT system were optimised on the development data set using MERT (Clark et al., 2011; Bertoldi et al., 2013).

The TED-MWE corpus is available for download at: http://tiny.cc/TED_MWE. In the next sections we describe the guidelines used for the annotation, the methodology adopted for the annotation process and the results of the annotation process.

4. The annotation guidelines

The judgement of whether an expression should qualify as an MWE relies on the annotation guidelines, which are based on (i) the PARSEME MWE template and (ii) the testing of MWE properties.

The PARSEME MWE Template (Savary et al., 2015) was designed to provide information and examples for MWEs in different languages along comparable dimensions of classification. These dimensions are: syntactic structures (e.g. nominal, verbal, adjectival, prepositional and clausal MWEs), the fixedness/flexibility of MWE parts (such as passivisation or modification), the different levels of idiomaticity (lexical, syntactic, semantic, pragmatic, statistical) and finally the rhetoric relations within an MWE.

In addition to the template, annotators were provided with a set of tests (Monti, 2013) to be used to assess whether a certain group of words can be considered as a MWE on the basis of the following properties:

- **Non-substitutability**: one element of the MWE 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 MWE 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-literal translatability**: the meaning cannot be translated literally. The difficulty of a literal translation across cultural and linguistic boundaries is mainly a property of MWEs with limited or no variation of distribution, such as idioms (e.g. *it's raining cats and dogs* → It. *sta piovendo cani e gatti), but also of many collocations (e.g. *heavy rain* → It. *pioggia pesante), fixed expressions (e.g. *by and large* → It. *da e largo), proverbs (e.g. *there's no such thing as a free lunch* → It. *non esiste una cosa come un pranzo gratuito), phrasal verbs (e.g. *bring somebody down* → It. *portare qualcuno giù);
- **Invariability**: Invariability can affect both the morphological and the syntactic level, whereby the inflectional variations of the constituents of the MWEs are not always possible. Invariability affects the head elements as well as 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 MWE 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.

In order to consider a certain word unit as an MWE it is sufficient that it shows at least one of the above-mentioned properties. Nevertheless, during the annotation process, the property which turned out to characterise the majority of MWEs was the non-literal translatability.

5. The annotation methodology

The annotation was organised in three distinct phases: individual annotation, inter-annotation check and validation.

Individual annotation

During the first phase, thirteen annotators with linguistic background in Italian and English were asked to annotate 1,529 sentences in the TED-MWE corpus. The sentences were organised in a spreadsheet (see Figure 2) containing the following information: (i) the English source text, (ii) the Italian manual translations (from the parallel corpus) and finally (iii) the Italian SMT output.

SNT #	Source (EN)	MANUAL Manual Translation (IT)	AUTO Automatic Translation (IT)	MWE				
				SOURCE TEXT	MANUAL TEXT	MANUAL CHECK (Y/N)	AUTO TEXT	AUTO CHECK (Y/N)
369	people sort of think i went away between "titanic" and "avatar" and was buffing my nails someplace, sitting at the beach.	la gente pensa quasi che me ne sia andato tra "titanic" e "avatar" e che mi stessi girando i pollici seduto su qualche spiaggia.	persone come pensare partii tra "titanic" e "avatar" e fu buffing mie unghie da qualche parte, seduto in spiaggia.	buffing my nails	girando i pollici	Y	buffing mie unghie	N

Figure 2. Annotation Phase 1 – Individual annotation

The annotators were asked to identify all MWEs in the source text together with their translations in approximately 300 random sentences each and to evaluate the correctness of the automatically translated MWE. If the manual or the SMT generated translations were wrong, the annotators were asked to specify the correct translations. The annotation work was organised in such a way that each sentence was annotated by at least two annotators. The annotation took into account all MWE types detected in the source text with no restrictions to a particular type of MWE and in particular, both continuous and discontinuous MWE types were recorded in the dataset. The MWEs identified during the annotation process were recorded as sequences of tokens with no further information about their internal syntactic structure or semantic features.

Inter-annotation validation

In the second phase, each annotator was confronted with the anonymised annotations of the other annotators on his/her annotation subset, in order to decide about his/her choices, i.e. to confirm or change the annotations for each source text/manual/SMT set.

Evaluation

Finally, we have randomly selected about half of the annotated sentences (801) and asked the annotators to integrate and resolve the possible annotation conflicts (see Figure 3).

SNT #	Source (EN)	MANUAL Manual Translation (IT)	AUTO Automatic Translation (IT)	ANN #	SOURCE TEXT	MANUAL TEXT	MWE MANUAL CHECK (Y/N)	AUTO TEXT	AUTO CHECK (Y/N)
26	"don," i said, "just to get the facts straight, you guys are famous for farming so far out to sea, you don't pollute."	"don", gli ho detto "tanto per capire bene, voi siete famosi per fare allevamento così lontano, in mare aperto, che non inquinate."	"non", ho detto, "per ottenere i fatti dritto, siete famosa per coltivare così lontano in mare, non inquinante."						
				3	to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N
				9	just to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N
				13	get… stright	capire bene	Y	per ottenere… dritto	N
				FINAL	just to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N

Figure 3. Annotation Phase 3 – Validation

6. The results of the annotation process

Based on the annotation process, out of 1,529 annotated sentences, 541 (35.9%) showed a good inter-annotation agreement, i.e. at least two annotators completely agreed on the annotations. In total we have collected 2,484 English MWEs types out of which 2,391 (96%) are contiguous and 93 (4%) are discontinuous. At least two annotators agreed for the 27% (671) of the MWEs and in 45% of them (1,115) at least two annotators showed an agreement (at least one word in common).

As a final step we have randomly selected about half of the annotated sentences (800) and asked the annotators to integrate and resolve the possible annotation conflicts. This resulted in a total of 799 English MWE types (931 tokens), of which 729 (91%) are contiguous and the 9% (70) are discontinuous.

Most MWEs have length of 2 (515) and 3 (261), but there are MWEs up to the length of 8. In 52% of the cases (471) the annotators have evaluated the automatic translation to be incorrect. Out of the 729 continuous MWEs, 253 occur only once in the whole English corpus and are therefore excluded from the final data set used for the experiments, which contains the remaining 476 English MWEs.

7. Translation asymmetries and mistranslations in the TED-MWE corpus

The fact that translation asymmetries (or divergences) between a source language and a target language may cause mistranslations is a well-known problem, but to the best of our knowledge it has not been systematically studied in the context of Machine Translation, in particular, with reference to the English-Italian language pair.

In our study, we have analysed all the asymmetries (about 250 different ones) which occur in the TED-MWE corpus and evaluated their impact on the quality of the MT output. In order to do so, a further annotation step was required: all the MWEs found in the corpus together with the correct Italian manual translation and their incorrect MT generated were annotated with POS information, as shown in Figure 4:

	A	B	C	D	E	F	G	H	I	J	K	L
499	no, i actually fell out of love with this fish because, i swear to god, after that conversation, the fish tasted like chicken.	no, in effetti mi disamorai di questo pesce perché, ve lo giuro dopo quella conversazione, il pesce sapeva di pollo.	no, ho cadevano d' amore con questo pesce perché, lo giuro su dio, dopo quella conversazione, i pesci assaggiato di pollo.	fell out of love	mi disamorai	Y		cadevano d' amore	N	V Part Prep N	Pron V	V Prep N
500	i was imagining a "march of the penguins" thing, so i looked at miguel.	immaginavo una cosa tipo la marcia dei pinguini così guardai miguel.	immaginai "marcia dei pinguini" cosa, così ho analizzato miguel.	looked at	guardai	Y		ho analizzato	N	V Prep	V	V
501	with pivot, you can drill into a decade.	con pivot si può osservare un decennio.	con perno, puoi bucare in un decennio.	drill into	osservare	N	instillare	bucare	N	V Prep	V	V
502	was coming into our living rooms with his amazing specials that showed us animals and places and a wondrous world that we could never really have previously	entrava nei nostri salotti con i suoi fantastici documentari che ci mostravano animali e luoghi e un mondo meraviglioso che prima non avremmo neanche potuto	stava arrivando nei nostri salotti con i suoi fantastici specialità che ci ha mostrato animali e luoghi e un mondo meraviglioso che potremmo mai immaginato prima	was coming into	entrava	Y		stava arrivando	N	V Prep	V	V

Figure 4. Annotation with POS information

A sample of about 500 MWEs incorrectly translated into Italian are taken into account. The mistranslations occur mainly with nominal and verbal MWEs. Discontinuous MWEs are mainly verbal ones and account for about 10% of translation errors. Examples of wrong translation correspondences for discontinuous MWEs are:

- [Verb … Adjective] as in *Not even the truth will **set** them **free** → Nemmeno la verità li **renderà libero*** (instead of *Neanche la verità riesce a **liberarli***)
- [Verb … Noun] as in *Is there any chance that politicians, that the country generally, would take a finding like that seriously and **run** public **policy** based on it? → C'è una possibilità è che i politici, che il paese generalmente, vorrebbe una scoperta simile seriamente e **correre politica** pubblica basato su?* (instead of *Esiste la possibilità che i politici, e la nazione in generale, possano prendere una scoperta come quella seriamente e **portare avanti** una **politica** pubblica basata su di essa?*)
- [Verb … Particle] as in *I'll **get** my sleeve **back**. → **Prenderò** mia manica* (instead of ***Tiro su** la manica.*)

The Table below shows the most mistranslated MWEs, in absolute terms.

Table 1. Translation errors per source MWE

Source MWE	#
Noun Noun	98
Verb Particle	86
Adjective Noun	54
Verb Preposition	36
Verb Noun	21
Verb Adverb	12
Verb … Noun	12

On the other hand, if we take into account the correspondences between source and target MWEs the picture changes. There are 262 different types of source-target MWE correspondences in the selected corpus and the most frequent mistranslations concern the following translation asymmetries:

Table 2. Translation errors per source-target MWE correspondences

Source MWE	Target MWE	#
Verb Particle	Verb	54
Noun Noun	Noun Preposition Noun	50
Adjective Noun	Noun Adjective	25
Noun Noun	Noun Adjective	25
Noun Noun	Noun	17
Verb Preposition	Verb	14

The one-to-many correspondences produce incorrect translations only in very few cases and concern the following structures [Verb → Verb Noun], [Adjective → Adjective Adverb], [Noun → Noun Adjective], and [Verb → Verb Preposition Noun].

Mistranslation due to many-to-one correspondences are numerous (153): the majority (140) are due to 2:1 correspondences and include 35 different types of correspondences among which the ones which produce the highest number of translation errors are:

– [Verb Particle → Verb] correspondence (54 translation errors), such as in *We **put out** a lot of carbon dioxide every year → Abbiamo **messo fuori** un sacco di anidride carbonica ogni anno*. (instead of *noi **emettiamo** molta co2 ogni anno)*
– [Noun Noun → Noun] correspondence (17 translation errors), such as in *I decided I was going to become a **scuba diver** at the age of 15. → Ho deciso che sarei diventato un **tuffatore bombole** all'età di 15 anni*. (instead of *Ho deciso che sarei diventato un **sommozzatore** all'età di 15 anni*.)
– [Verb Preposition → Verb] correspondence (14 translation errors), such as […] *You are not going to **get to** the correct answer. → […] Non vanno a **raggiungere** la risposta giusta* (instead of *Non potrete **ottenere** la risposta corretta*.)

The occurrence of mistranslation in many-to-many correspondences is shown in Figure 5. These types of correspondences represent the widest group with 378 translation errors in the corpus.

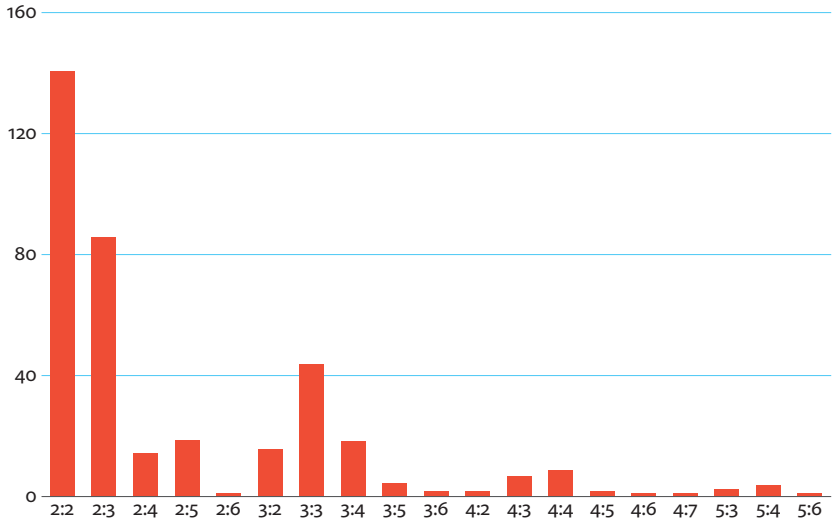

Figure 5. Translation errors per many-to-many correspondences

The main sources of errors are represented by:

- 2:2 correspondences with 141 translation errors, among which the [Adjective Noun → Noun Adjective] correspondence is the most problematic case (25 translation errors) as in *It struck me how much this dive, these deep dives, was like a **space mission**. → Mi colpì quanto questa immersione, queste immersioni profonde, era come uno **spazio missione**.* (instead of *Sono rimasto fulminato da quelle immersioni profonde, era come una **missione spaziale**.*)
- 2:3 correspondences with 86 translation errors, among which the [Noun Noun → Noun Preposition Noun] correspondence causes 50 errors, such as in *It's a **fish farm** in the southwestern corner of Spain. → È un **pesce fattoria** in un angolo sudovest della Spagna* (instead of *È un **allevamento di pesci** nell'angolo sudoccidentale della Spagna*)
- 3:3 correspondences with 44 translation errors: this translation asymmetry shows a high grade of variability with 40 different correspondences. An example is the [Noun Noun Noun → Noun Noun Adjective] asymmetry, as in *5.93 million years ago was when our earliest **primate human ancestors** stood up. → 5.93 milioni di anni fa era quando i nostri primi **primate antenati umani** si alzò.* (instead of *5.93 milioni di anni fa fu il periodo i nostri **antenati primati umani** si alzarono in piedi*).
- 3:4 correspondences with 19 translation errors, among which the [Adjective-Noun Noun → Noun Preposition Noun Adjective] one represent the most troublesome class as in *I hope that you will agree with me that gamers are a human resource that we can use to do **real-world work** → Spero che sarete d'accordo con me che giocatori sono una risorsa umana che possiamo usare per fare **funzionare reale*** (instead of *Spero che siate d'accordo con me che i giocatori abituali sono una risorsa umana che possiamo utilizzare per fare del **lavoro nel mondo reale***)
- 2:5 correspondences with 19 translation errors, among which there are nine different correspondences. An example is the [Verb Particle → Verb Preposition Determiner Adjective Noun] correspondence as in *If you're getting queasy, **look away** → Se vi steste queasy, **guarda**.* (instead of *Se vi sentite male **guardate da un'altra parte**.*)
- 3:2 correspondences with 16 translation errors. An example is the [Verb Particle Noun → Verb … Noun] correspondence, as in *He **set up** a **camera** in front of gamers. → Così **ha creato** una **telecamera** davanti ai giocatori mentre erano giocare.* (instead of ***Ha messo** una **telecamera** di fronte ai giocatori*).
- 2:4 correspondences with 15 translation errors. An example is [Noun Noun → Noun Preposition Noun Adjective], as in *The average young person today in a country with a strong **gamer culture** will have spent 10,000 hours playing online → A media oggi giovani in un paese con un forte **giocatore culture** avranno speso 10.000 ore davanti giochi online dall'età di 21 anni.* (instead of *Il tipico giovane medio oggi giorno in un paese con una forte **cultura di giocatore abituale**, avrà passato 10.000 ore giocando online, all'età di 21 anni.*)

7. Conclusions and future work

In this chapter, we have dealt with the concept of translation asymmetries of multi-word expressions in Machine Translation with reference to the English-Italian language pair. The study is based on the analysis of the TED-MWE corpus, containing MWE-annotated sentences of an English-Italian parallel corpus, complemented and compared with an Italian MT output also annotated with MWEs. The MT output has been further analysed in terms of translation divergences, looking at the correspondence patterns between the two languages under examination.

The rationale for taking on translation asymmetries is to observe the cases where structures of both source and target language are divergent, and where these divergences are the cause of mistranslations. This analysis might prove to be useful in relation to better MWE processing and translation, since it conducts a thorough analysis of the patterns which may create problems to an accurate and fluent MT output.

Future work will concern a more fine-grained analysis of the types of errors that occur for different translation asymmetries, which cause a one of the largest translation error class. This analysis will help researchers to understand whether specific translation asymmetries are related to specific error typologies.

References

Mihael, A., Turchi, M., Tonelli, S., & Buitelaar, P. (2017). Leveraging bilingual terminology to improve machine translation in a CAT environment. In *Natural Language Engineering*, 23(5), 763–788. https://doi.org/10.1017/S1351324917000195

Barreiro, A. (2008). *Make it simple with paraphrases: Automated paraphrasing for authoring aids and machine translation* (PhD Thesis, Universidade do Porto).

Bertoldi, N., Haddow, B., & Fouet, J.B. (2009). Improved minimum error rate training in MOSES. *Prague Bull. Math. Linguistics*, 91(1), 7–16.

Bertoldi, N., Cettolo, M., & Federico, M. (2013). Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of Machine Translation Summit* XIV. Nice, France.

Bouamor, D., Semmar, N., & Zweigenbaum, P. (2012). Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey

Boulaknadel, S., Daille, B., & Aboutajdine, D. (2008). A multi-word term extraction program for arabic language. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., & Mercer R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2), 263–311.

Caseli, H., Villavicencio, A., Machado, A., & Finatto, M. J. (2009). Statistically-driven alignment- based multiword expression identification for technical domains. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications* (pp. 1–8). Singapore: Association for Computational Linguistics.

Cettolo, M., Girardi, C., & Marcello, F. (2012). Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)* (pp. 261–268). Trento, Italy.

Clark, J., Dyer, C., Lavie, A., & Smith, N. (2011). Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*-Volume 2 (pp. 176–181). Association for Computational Linguistics.

Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Rasmich, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: a survey. *Computational Linguistics*, 43(4), 837–892. https://doi.org/10.1162/COLI_a_00302

Dagan, I., & Church, K. (1994). Termight: Identifying and translating technical terminology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing* (pp. 34–40). Association for Computational Linguistics. https://doi.org/10.3115/974358.974367

Daille, B. (2001). Extraction de collocation à partir de textes. In *TALN 2001 (Traitement automatique des langues naturelles*).

Dorr, B. J. (1994). Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4), 597–663.

Marcello, F., Bertoldi, N., & Cettolo, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association* (pp. 1618–1621). Brisbane, Australia.

Fu, B., Brennan, R., & O'Sullivan, D. (2009). Cross-lingual ontology mapping–an investigation of the impact of machine translation. In A. Gómez-Pérez, Y. Ding, Y. Yong (Eds.), *Asian Semantic Web Conference* (pp. 1–15). Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-10871-6_1

Kauffmann, A., & Azar, J. (2013). *Structural Asymmetries in Machine Translation: The case of English-Japanese*. (PhD Thesis, University of Geneva).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 177–180). Prague, Czech Republic: ACL.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*-Volume 1 (pp. 48–54). Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit* (pp. 79–86). Phuket, Thailand: AAMT.

Lambert, P., & Banchs, R. (2005). Data inferred multi-word expressions for statistical machine translation. In *Proceedings of Machine Translation Summit* X (pp. 396–403).

Lin, S.-C., Wang, J.-C., & Wang, J.-F. (2005). Translation Divergence Analysis and Processing for Mandarin-English Parallel Text Exploitation. In *Proceedings of 17th Conference on Computational Linguistics and speech Processing (ROCLING 2005)*. Tainan, Taiwan.

Losnegaard, G. S., Sangati, F., Parra Escartín, C., Savary, A., Bargmann, S., & Monti, J. (2016). PARSEME Survey on MWE Resources. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia.

Melamed, I. D. (1997). Automatic discovery of noncompositional compounds in parallel data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 97–108).

Moirón, B. V., & Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multi-word expressions in a multilingual context* (pp. 33–40).

Monti, J., & Todirascu, A. (2016). Multiword units translation evaluation in machine translation: another pain in the neck? In G. Corpas Pastor, J. Monti, R. Mitkov, & V. Seretan (Eds.), *Workshop proceedings for Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2015)* (pp. 25–30). Geneva: Editions Tradulex.

Monti, J., Sangati, F., & Arcan, M. (2015). TED-MWE: a bilingual parallel corpus with MWE annotation. In C. Bosco, S. Tonelli, & F. M. Zanzotto (Eds.), *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)* (pp. 193–197). Torino: Accademia University Press srl/Centro Altreitalie. https://doi.org/10.4000/books.aaccademia.1514

Monti, J. (2013). *Processing in Machine Translation. Developing and using language resources for multi-word unit processing in Machine Translation* (PhD Thesis in Linguistica Computazionale, Università degli Studi di Salerno, a.a. 2011–2012).

Monti, J. (2014). An English-Italian MWE dictionary. In *Clic-it Proceedings* 2014. Pisa University Press srl.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29 (1), 19–51.

Okita, T., & Way, A. (2010). Statistical Machine Translation with Terminology. In *Proceedings of the First Symposium on Patent Information Processing (SPIP)* (pp. 1–8).

Pause, P. E. (1997). Interlingual strategies in translation. In C. Hauenschil, & S. Heizmann (Eds.), *Machine Translation and Translation Theory* (pp. 175-190).

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. CICLing 2002. Lecture Notes in Computer Science*, vol 2276. Berlin/ Heidelberg: Springer. https://doi.org/10.1007/3-540-45715-1_1

Sangati, F., & Cranenburgh A. V. (2015). Multiword Expression Identification with Recurring Tree Fragments and Association Measures. In *Proceedings of Annual conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 10–18). Denver, CO: Association for Computational Linguistics.

Savary, A., Sailer, M., Parmentier, Y., Rosnes, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B., Losnegaard, G. S., Parra Escartín, C., Waszczuk, J., Constant, M., Osenova, P., & Sangati, F. (2015). PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*. Poznań, Poland.

Seretan, V., & Wehrli, E. (2007). Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)* (pp. 401–410). Toulouse, France.

Sinha, R. M. K., & Thakur, A. (2005). Translation divergence in English-Hindi MT. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 2005)* (pp. 245–254). Budapest, Hungary.

Thurmair, G., & Aleksić, V. (2012). Creating term and lexicon entries from phrase tables. In *Proceedings of the 16th Conference of the European Association for Machine Translation* (*EAMT 2012*).

Tsvetkov, Y., & Wintner, S. (2010). Extraction of multiword expressions from small parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 1256–1264). Association for Computational Linguistics.

Vintar, S., & Fišer, D. (2008). Harvesting multi-word expressions from parallel corpora. In *LREC. European Language Resources Association*.

Wu, C.C. & Chang, J. S. (2004). Bilingual Collocation Extraction Based on Syntactic and Statistical Analyses. *Computational Linguistics and Chinese Language Processing*, 9(1), 1–20.