DE GRUYTER MOUTON DOI 10.1515/phras-2020-0005 — YoP 2020; 11: 61–80

Carlos Manuel Hidalgo-Ternero and Gloria Corpas Pastor

Bridging the *"gApp"*: improving neural machine translation systems for multiword expression detection

Abstract: The present research introduces the tool *gApp*, a Python-based text preprocessing system for the automatic identification and conversion of discontinuous multiword expressions (MWEs) into their continuous form in order to enhance neural machine translation (NMT). To this end, an experiment with semi-fixed Verb–Noun Constructions (VNC) will be carried out in order to evaluate to what extent *gApp* can optimise the performance of the two main free open-source NMT systems —Google Translate and DeepL— under the challenge of MWE discontinuity in the Spanish into English directionality. In the light of our promising results, the study concludes with suggestions on how to further optimise MWE-aware NMT systems.

Keywords: text preprocessing system, Neural Machine Translation (NMT), Multiword Expression (MWE), Verb–Noun Constructions (VNC), discontinuity

1 Introduction

Twenty years after Sag et al.'s (2002) seminal publication, multiword expressions (MWEs) are still a *pain in the neck* for natural language processing systems, and machine translation is not an exception (see the latest survey by Monti et al. [2018] and the papers in Corpas Pastor and Colson [2020]). Besides their notoriously problematic features such as syntactic anomaly, non-compositionality, and ambiguity, inter alia, a further challenge arises for Neural Machine Translation (NMT): MWEs do not always consist of contiguous tokens (e.g. *take his piece of advice into consideration*), which seriously hinders their automatic identification and translation (Constant et al. 2017; Foufi et al. 2019; Ramisch and Villavicencio 2018; Rohanian et al. 2019).

Carlos Manuel Hidalgo-Ternero and Gloria Corpas Pastor, University of Málaga, Department of Translation and Interpreting, Campus de Teatinos s/n, 29071-Malaga, Spain, cmhidalgo@uma.es, gcorpas@uma.es

In order to overcome the challenges that discontinuous MWEs still pose for even the most robust NMT systems (Zaninello and Birch 2020), we have developed *gApp*, a text preprocessing system for the automatic identification and conversion of discontinuous MWEs into their continuous form, in order to enhance the performance of NMT systems. Against such a background, the overall structure of this paper is as follows. Section 2 offers a brief literature review to provide the context of the current research. Then, the system *gApp* is described in Section 3. Section 4 covers the research methodology. In Section 5, the text preprocessing system's precision and recall will be tested in order to then evaluate to what extent *gApp* can optimise the performance of the two main free open-source NMT systems —Google Translate and DeepL— under the challenge of MWE discontinuity in the Spanish into English directionality. A discussion of the results will then be presented in Section 6. Finally, Section 7 offers concluding remarks on how this tool can enhance the identification and translation of MWEs by NMT systems.

2 Related work

The quest for optimising the treatment of multiword expressions by current NMT systems has laid a fertile ground for research and debate. In this regard, there is a growing body of literature that has already yielded significant advances in the field of multiword expression detection (Finlayson and Kulkarni 2011; Klyueva et al. 2017; Maldonado et al. 2017; Riedl and Bieman 2016; Zampieri et al. 2019; inter alia), and, more precisely, in the automatic identification of discontinuous MWEs (Al Saied et al. 2017, Al Saied et al. 2019; Alegria et al. 2004; Bejček et al. 2011, Bejček et al. 2013; Constant et al. 2017; Foufi et al. 2019; Moreau et al. 2018; Rohanian et al. 2019; Schneider et al. 2014), as well as in the optimisation of MWEs' treatment by current NMT systems (Huang et al. 2018; Rikters and Bojar 2017; Wang et al. 2017; Zaninello and Birch 2020).

Against such a background, there currently exist some MWE processing systems which can perform a token-based MWE identification, as shown, for instance, in Alegria et al. (2004), Bejček et al. (2013), Finlayson and Kulkarni (2011), Nagy and Vincze (2014) and Ramisch (2015), among other systems implemented within the framework of the COST action PARSEME (cf. Ramisch et al. [2018]). All of them employ a lexicon lookup method (Ramisch and Villavicencio 2018), i.e. they resort to a predefined lexicon of patterns for the automatic detection of MWEs in running text. Among these tools, some are able to specifically identify discontinuous expressions by employing a gap-length parameter in

order to delimit the maximum number of tokens permitted within MWE constituents (Ramisch 2015), using *surface realisation schemas* (SRS) with the necessary information on the order of the MWE components, their mandatory or optional continuity and their inflectional restrictions (Alegria et al. 2004), and setting the system so that both lexical items in a discontinuous MWE (the preposed and the postposed) concomitantly trigger the identification process (Foufi et al. 2019), among other techniques.

The text preprocessing tool *gApp* can also perform an analogous token-based identification of discontinuous MWEs. However, in comparison with previous state-of-the-art systems, *gApp* represents a considerable advance since it is not only capable of detecting discontinuous MWEs in running text, but can also automatically convert them into their continuous form, which significantly enhances the performance of, to date, the most robust NMT systems —Google Translate and DeepL—, as it will be demonstrated in the following sections.

3 Overview of the tool

The significant challenges posed by MWE discontinuity have made it necessary to develop *gApp*, a Python-based text preprocessing system for detecting and converting discontinuous MWEs into their continuous form in running text, with a view to enhancing NMT performance. In this regard, *gApp* comprises a lexicon of somatisms that enter as constituents into semi-fixed Verb–Noun Constructions (VNC) in Spanish. Somatisms are terms referring to human or animal body parts. Some examples of somatic VNCs are *sentar la cabeza* (lit., 'to sit the head'; figuratively, 'to settle down'), *tomar el pelo* (lit., 'to take someone's hair'; fig., 'to fool someone'), *meter la pata* (lit., 'to put the leg in'; fig., 'to screw it up'), *verse las caras* (lit., 'to see each other's faces'; fig., 'to confront someone'), etc.

For the automatic identification and conversion of MWEs, *gApp* employs the free open-source library *Spacy*, specialised in performing a wide array of advanced NLP tasks, including non-destructive tokenisation, POS tagging, dependency parsing, lemmatisation, and rule-based matching, among others. Regarding rule-based matching, with Spacy it is possible to automatically detect a set of tokens with a specific pattern together with other optional elements that may occur within (and hence split) that MWE (Honnibal and Montani 2017).

Prior to developing identification patterns for *gApp*, a first necessary step was to establish the kind of n-grams that may appear within the discontinuous form of the somatisms under study. To this end, we followed a corpus-based methodology to query two giga-token web-crawled corpora of Spanish (esTenTen and

JSI Spanish Timestamped Corpus), both available through Sketch Engine. The esTenTen corpus contains over 17.5 billion words of general Spanish (European and American varieties), while the JSI Spanish Timestamped Corpus comprises over 11.6 billion words of news articles obtained from their RSS feeds (Kilgarriff et al. 2003).

Analogously to Hidalgo-Ternero's (2020) corpus-based research methodology, Sketch Engine's Corpus Query Language (CQL) schemas have been applied in order to retrieve both the discontinuous forms of the somatisms under study (henceforth named *relevant results*) as well as other concordances containing analogous patterns but unrelated to the idiomatic sequences (*irrelevant results*), which will eventually determine the necessary restriction for *gApp*'s detection system so as to optimise its precision and recall. In this regard, the implemented CQL schemas are presented in Tab. 1.

Tab. 1: CQL schemas for the discontinuous form of the somatisms

Sequence	CQL schemas
Sentar [1–3 tokens] la cabeza	[lemma="sentar"][]{1,3}[word="la"][word="cabeza"]
Meter [1–3 tokens] la pata	[lemma="meter"][]{1,3}[word="la"][word="pata"]
Tomar [1–3 tokens] el pelo	[lemma="tomar"][]{1,3}[word="el"][word="pelo"]
Verse [1–3 tokens] las caras	[lemma="ver"][]{1,3}[word="las"][word="caras"]

Once the different concordances with both relevant and irrelevant results were analysed, different rule-based matching patterns were developed in order to identify as many relevant results while filtering out as many irrelevant results as possible. By way of illustration, this has been the base pattern employed to detect discontinuous instances of the somatism *tomar el pelo* with *gApp*:

(1) matcher.add('tomar X el pelo', on_match, [{'LEMMA': 'tomar', 'POS': 'VERB'}, {'POS': {'IN': ['DET', 'ADV', 'ADP', 'PRON', 'PROPN']}}, {'OP': '?'}, {'POS': {'NOT_IN': ['VERB', 'CONJ', 'SCONJ', 'ADP']}, 'OP': '?'}, {'ORTH': 'el'}, {'ORTH': 'pelo'}])

Within the method *matcher.add()*, it is possible to include a rule to the matcher, comprising an ID key, a number of patterns and a callback function (in our case, *on_match*) with the arguments *matcher*, *doc*, *id* and *matches,* to act on the matches (Honnibal and Montani 2017). The patterns consist of a list of dictionaries, each of which contains the necessary description of both the exact tokens of the MWE and of those elements that may occur within the sequence. In the case of *tomar el pelo*, the first dictionary of the pattern includes the attribute *'LEMMA*'

for *tomar* with *verb* as its POS tag. For each of the remaining tokens of the MWE (*el pelo*), the attribute *'ORTH'* had to be added since none of them allow inflection in this idiomatic sequence. Regarding the gap, the concordances retrieved from the analysed corpora reveal that this MWE can be split by various adverbial, nominal and prepositional phrases such as *más* ('more'), *tanto* ('so much'), *un poco* ('a bit'), *una vez más* ('once again') or any subject, among others. Therefore, the first constituent of the gap was restricted to tokens with the following POS tags in the form of a list: *'DET'* ('determiner'), *'ADV'* ('adverb'), *'ADP'* ('adposition'), *'PRON'* ('pronoun'), and *'PROPN'* ('proper noun'). For those tokens within the gap whose occurrence is optional, the attribute *'OP': '?'* was included. Finally, in order to also filter out those elements at the final position of the gap belonging to the irrelevant results, the code *{'POS': {"NOT_IN": ["VERB", "CONJ", "SCONJ", "ADP"]}, 'OP': '?'}* had to be incorporated since, according to the concordances retrieved from the analysed corpora, those results containing either a verb, a conjunction, or an adposition directly preceding the bigram *el pelo* were unrelated to discontinuous instances of the somatism *tomar el pelo* and had hence to be excluded so as to increase *gApp*'s precision, as it can be observed in the following instances including irrelevant results:

(2) Sería la última vez que lo verían como Edwin, pero con marcadas muestras de que Pamela ganaba terreno. "Ya **tomaba hormonas, tenía el pelo** largo y me maquillaba". Luego vendría la operación y el cambio de nombre definitivo, que logró hace unos doce años.

(3) "Hoy la mujer está súper predispuesta a ver su pelo diferente. Y, sobre todo, a tener un estilo propio", dice Bebe Sanders, estilista de muchas actrices y modelos. "Se está **tomando conciencia de que el pelo** tiene tanto protagonismo en el puntaje estético de una mujer que es casi imperdonable no disfrutarlo", agrega.

Instances 2 and 3 contain concordances with the pattern *tomar* [3 tokens] *el pelo*, which can be considered as irrelevant results because they are literal sequences that are not related to a discontinuous form of the somatism *tomar el pelo*. In instance 2 it is possible to observe the juxtaposition of two main clauses, separated by a comma: *"Ya tomaba hormonas, tenía el pelo largo…."* In this example, both the verb *tomar* and the noun phrase *el pelo* are used in their literal meanings in the two independent clauses, which could hence be translated as *"I was already taking hormones, had long hair and wore make-up…"*. In instance 3, both *tomar* and *el pelo* belong to different clauses joined by the subordinate conjunction *que*: *"Se está tomando conciencia de que el pelo tiene tanto protagonismo…"*. In this case, *el pelo* is also employed with its literal meaning ('hair'), while *tomar*

is the verbal collocate of *tomar conciencia* ('to become aware'). This sequence from instance 3 could hence be translated as *"People are becoming aware that hair plays such an important role…"*. Despite presenting a similar pattern to the discontinuous form of *tomar el pelo*, these irrelevant results were properly filtered out by *gApp* because in instance 2 the first element of the gap is a noun and the final one is a verb, and in instance 3 the first element of the gap is again a noun and the final one consists of a subordinate conjunction. These elements at the initial and final positions of the gap had been set not to be detected and converted by *gApp* following the predefined patterns for this idiom. Fig. 1 provides a partial overview of the lexicon of somatisms and the rule-based matching patterns.

Despite its refined POS tagger, Spacy still proved to show some limitations. For example, when searching for the lemma *tomar* with the POS tag *verb*, it was possible to observe that Spacy was not capable of properly detecting instances of *tomar* in the infinitive, gerund and imperative form with an enclitic pronoun (*tomarte*, *tomándonos*, *tómales*…) as it detected this token with several POS tags different to its original one, i.e. *verb*. In order to overcome this limitation, an additional alternative first dictionary in that sequence had to be included with the following pattern *{"TEXT": {"REGEX": "^[Tt][oó]m[aáeé][d-t][d-t]*"}}*. With this regular expression, it was possible to also include those instances that had first been mistakenly considered as unrelated to the lemma *tomar* as a verb.

Once the first stage of identification of discontinuous somatisms was completed, it was necessary to develop an additional code for their automatic conversion into the continuous form. To this purpose, with a *for loop* it was possible to create a first condition if the tool matched any of the predefined patterns within the lexicon of somatisms. In this case, the system was set to detect the first dictionary of the match (called *match[1]*) and the last one (called *match[2]*) with the gap as those optional elements within the sequence in reference to the first and last match, i.e. the first token within the gap would be *match[1]+1* (called *gap1*) and the final one would be *match[2]-2* (named *gap3*). Once the gap had been defined, the system was set to automatically print the text from the beginning of the document up to *match[1]*, then *match[2]*, subsequently from *gap1* up to *gap3*, and finally from *match[2]* up to the end of the document, which resulted in the whole original text with the somatism in the continuous form as the output. If the first condition was not met, i.e. if none of the predefined patterns within the lexicon of somatisms was matched, the system was set not to perform any change in the document.

gApp.py*

```
import spacy
import re
from spacy.matcher import Matcher
nlp = spacy.load("es_core_news_sm")
statement = input("Please insert your text here: ")
def on_match(matcher, doc, id, matches):
    print('Matched!')
matcher = Matcher(nlp.vocab)
#BRAZO
matcher.add("bajar X los brazos", on_match, [{"LEMMA": "bajar"}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'OP': '?'}, {'POS':
#CABEZA
matcher.add("esconder X la cabeza", on_match, [{'LEMMA': "esconder", 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {
matcher.add("jugarse X la cabeza", on_match, [{'LEMMA': 'jugar', 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'OP'
matcher.add("levantar X la cabeza", on_match, [{'LEMMA': 'levantar', 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {
matcher.add("meter X la cabeza", on_match, [{'LEMMA': 'meter', 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'OP':
matcher.add("perder X la cabeza", on_match, [{'LEMMA': 'perder', 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'OP'
matcher.add("romperse X la cabeza", on_match, [{'LEMMA': 'romper', 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'O
matcher.add("sentar X la cabeza", on_match, [{'LEMMA': 'sentar'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'OP': '?'}, {'POS':
#CARA
matcher.add("dar X la cara", on_match, [{'LEMMA': 'dar', 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'OP': '?'},
matcher.add("verse X las caras", on_match, [{'LEMMA': 'ver', 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'OP': '?
matcher.add("verse X las caras", on_match, [{"TEXT": {"REGEX": "^[Vv](iéndo|er|eámo|éan[nos,][eso,]*)|eos|isto"}}, {'POS': {"IN": ["DET", "ADV"
#CODOS
matcher.add("hincar X los codos", on_match, [{'LEMMA': 'hincar', 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'OP'
#CUERNOS
matcher.add("poner X los cuernos", on_match, [{"LEMMA": "poner"}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN", "CONJ", "NUM"]}}, {'OP'
matcher.add("poner X los cuernos", on_match, [{"TEXT": {"REGEX": "^[Pp][oó]n[e-t][a-r]*"}}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN
#DEDO
matcher.add("chuparse X el dedo", on_match, [{'LEMMA': 'chupar', 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'OP'
matcher.add("chuparse X los dedos", on_match, [{'LEMMA': 'chupar', 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'O
matcher.add("mover X un dedo", on_match, [{'LEMMA': 'mover', 'POS': 'VERB'}, {'POS': {"IN": ["DET", "ADV", "ADP", "PRON", "PROPN"]}}, {'OP': '?
```

Fig. 1: Overview of *gApp*'s rule-based matching lexicon of somatisms

4 Methodology

After *gApp*'s mechanism for the automatic identification and conversion of discontinuous somatisms has been introduced, this section presents the research conducted in order to evaluate to what extent *gApp* can enhance the performance of the NMT systems DeepL and Google Translate. Following Hidalgo-Ternero (2020), the concordances containing the discontinuous somatisms under study have been retrieved from esTenTen corpus and Spanish JSI Timestamped Corpus, comprising a heterogeneous sample in terms of text sources, types and language varieties. In spite of the challenge that noisy input still poses to even the most robust NMT systems (Sperber et al. 2017; Belinkov and Bisk 2018; Anastasopoulos 2019; Niu et al. 2020), concordances containing user-generated content (UGC) were also included in the test so as to mitigate sample bias, which could otherwise arise from solely analysing NMT canonical training data for the somatisms under study. In this regard, a total of 560 cases was examined, considering the continuous and discontinuous forms of the somatisms *sentar la cabeza*, *meter la pata*, *verse las caras*, and *tomar el pelo*, split by different unigrams, bigrams or trigrams. For each somatism, 70 irrelevant results were compiled, all of which included sequences unrelated to the idiom but whose form may pose some challenges for the automatic identification through *gApp*, in order to calculate, at a first stage, both the precision and recall of this text preprocessing system.

After the calculation of both parameters, at a second stage, the results concerning the NMTs' performance for the different concordances were classified within three main categories depending on whether the somatisms under study were presented in their discontinuous form (category 1) or in their continuous form after being automatically converted through *gApp* (category 2) in contrast to their manual conversion (category 3). DeepL and Google Translate's outputs for these different scenarios were then manually assessed following a reference-based MT evaluation (Hidalgo-Ternero 2020) with several possible target-text candidates for each of the somatisms in both their continuous and discontinuous forms. In this regard, morphological, syntactic, and/or orthotypographic divergences or source-text/translation inaccuracies affecting other elements in the sentences were not listed *per se* as errors if they were not related to the phenomenon of MWE discontinuity for the somatisms under study.

5 Results

In this section, the results will be analysed and presented in two stages. Thus, *gApp*'s precision and recall for each of the somatisms under study will be set out

in order to assess to what extent *gApp* optimises the performance of the two NMT systems —Google Translate and DeepL— under the challenge of MWE discontinuity. In the case of *sentar la cabeza*, *gApp* automatically detected and converted 69 forms, 65 of which were true positives and 4 were false positives. In this way, *gApp*'s precision for this idiomatic sequence was 94.2% (65/69 cases) with a recall of 92.9% (65/70 cases). The NMTs' results for the continuous and discontinuous form of this somatism are set out in Fig. 2.

Regarding DeepL's output, it can be observed that the automatic conversion into the idiom's continuous form through *gApp* did not lead to a significant improvement in this NMT's performance compared to the discontinuous scenario (only 1-case difference, i.e. 1.4%), whereas its manual conversion presented a considerable enhancement (a 6-case difference, i.e. 8.6%). In the case of Google Translate, *gApp* could approach the manual's amelioration with a 5-case divergence, i.e. 7.1%, versus a 7-case one, i.e. 10%, respectively. An illustration of the conversion from the discontinuous to the continuous form of *sentar la cabeza* with *gApp* is presented in Tab. 2 and the NMTs' performance before and after this conversion through *gApp* is shown in Tab. 3.

As it can be observed in Tab. 3, in DeepL both the source-text (ST) continuous and discontinuous idiom was identified and an appropriate equivalent was provided in the target text (TT). However, that was not the case with Google Translate. Before the conversion with *gApp*, this NMT could not properly identify the ST discontinuous idiom, which hence led to a literal and inappropriate translation in the target text. It was only after the conversion with *gApp* that Google

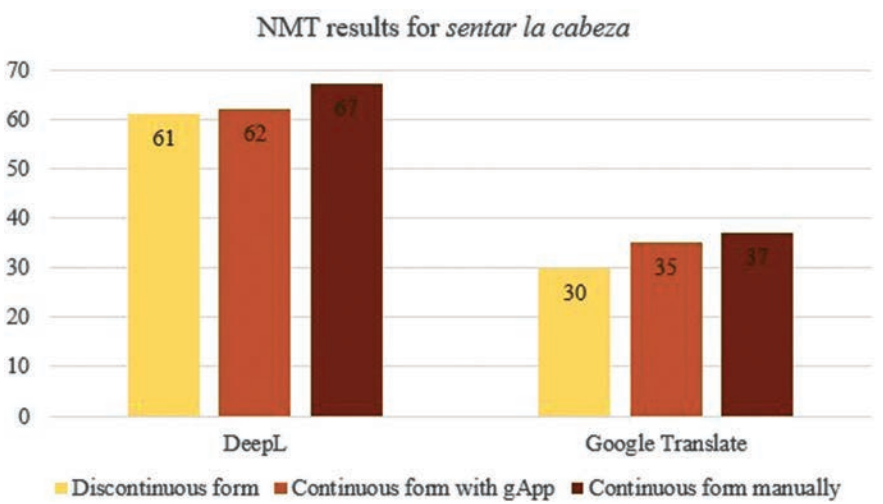

Fig. 2: NMT results for *sentar la cabeza*

Tab. 2: Source-text KWIC extracts with *sentar la cabeza* before and after the conversion with *gApp*

	KWIC extracts
ST [ES] Discontinuous form (original version, before *gApp*)	Exiliado, Piccolo ha dejado de ser una amenaza para la humanidad y Goku, sin perder su pasión por la lucha, parece **haber sentado ligeramente la cabeza** tras el nacimiento de su hijo Son Gohan.
ST [ES] Continuous form (after *gApp*)	Exiliado, Piccolo ha dejado de ser una amenaza para la humanidad y Goku, sin perder su pasión por la lucha, parece **haber sentado la cabeza ligeramente** tras el nacimiento de su hijo Son Gohan.

Tab. 3: DeepL and Google Translate's outcomes before and after the conversion of the ST idiom *sentar la cabeza* with *gApp*

	NMT outcomes
TT [EN] Discontinuous form in DeepL (before *gApp*)	In exile, Piccolo is no longer a threat to humanity, and Goku, without losing his passion for fighting, seems to **have settled down slightly** after the birth of his son Son Gohan.
TT [EN] Continuous form in DeepL (after *gApp*)	In exile, Piccolo is no longer a threat to humanity, and Goku, without losing his passion for fighting, seems to **have settled down slightly** after the birth of his son Son Gohan.
TT [EN] Discontinuous form in Google Translate (before *gApp*)	Exiled, Piccolo is no longer a threat to humanity and Goku, without losing his passion for fighting, seems to **have slightly settled his head** after the birth of his son Son Gohan.
TT [EN] Continuous form in Google Translate (after *gApp*)	Exiled, Piccolo is no longer a threat to humanity and Goku, without losing his passion for fighting, seems to **have settled down slightly** after the birth of his son Son Gohan.

Translate could properly detect and offer an adequate equivalent in English for the ST idiom.

As for the somatism *meter la pata*, *gApp* automatically converted 68 forms, all of them corresponding to true positives. Therefore, this tool's precision amounted to 100% (68/68 cases) and its recall equated to 97.1% (68/70 cases). An overview of the NMTs' performance for the continuous and discontinuous form of this somatism is presented in Fig. 3. When contrasting the data, it is possible to notice that the continuous form of the somatism attained a quasi-analogous

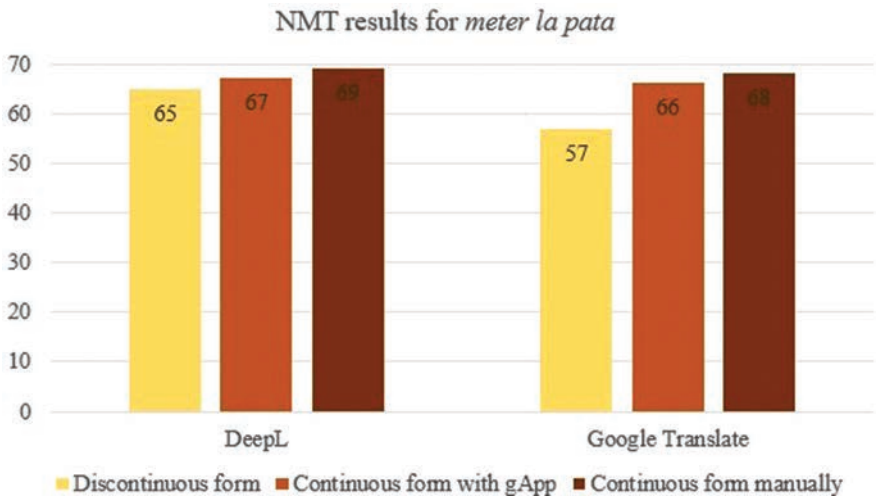

Fig. 3: NMT results for *meter la pata*

outcome for DeepL and Google Translate both through *gApp* (67 and 66 cases, respectively) and the manual conversion (69 and 68). Nevertheless, when comparing it with the discontinuous scenario, major dissimilarities are to be observed: whereas *gApp* and the manual conversion could only result in a slight enhancement for DeepL (2.9% and 5.7%, respectively), both did considerably ameliorate Google Translate's performance for this idiomatic sequence (12.9% and 15.7%, respectively).

With regard to *tomar el pelo* (Fig. 4), *gApp* transformed 73 cases, including 68 true positives and 5 false positives. In this way, the tool's precision came to 93.2% (68/73) and its recall amounted to 97.1% (68/70 cases) for this somatism. Analogously to the case of *meter la pata*, the continuous form of *tomar el pelo* presented a similar performance for the two NMTs both through the automatic (65 and 62 cases) and the manual (67 and 64) conversion. In contrast to the previous somatism, when comparing the NMTs' performance for the discontinuous and continuous form of *tomar el pelo*, only minor differences have been observed. DeepL's performance improved by 18.6% with *gApp* and by 21.4% through manual conversion, while Google Translate's outcome improved even less (*gApp*: 15.7%; manual conversion: 18.6%). In Tab. 4 it is possible to observe an instance of the conversion from the discontinuous to the continuous form of *tomar el pelo* with *gApp*. Tab. 5 illustrates the NMTs' performance before and after *gApp*.

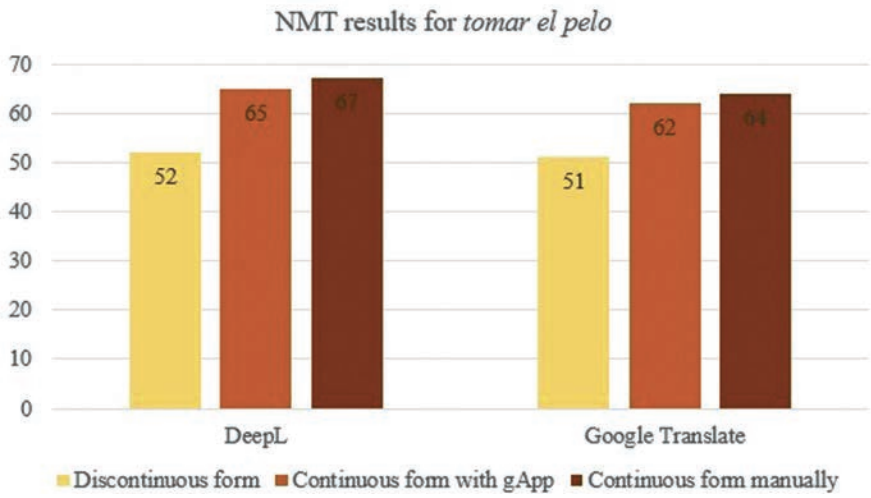

Fig. 4: NMT results for *tomar el pelo*

Tab. 4: Source-text KWIC extracts with *tomar el pelo* before and after the conversion with *gApp*

	KWIC extracts
ST [ES] Discontinuous form (original version, before *gApp*)	Así consta, ha dicho Monago, en el registro mercantil del Reino Unido, que refleja un reporte anual y el estado financiero de la empresa, según informa el PP extremeño en nota de prensa. </s><s> **"¿Se le puede tomar a la gente el pelo en campaña electoral?"**, ha planteado Monago, quien ha señalado que se están riendo de los extremeños diciéndoles que una empresa de 14.000 euros de capital social va a generar tres_millones de turistas al año y va a hacer rascacielos en Castilblanco.
ST [ES] Continuous form (after *gApp*)	Así consta, ha dicho Monago, en el registro mercantil del Reino Unido, que refleja un reporte anual y el estado financiero de la empresa, según informa el PP extremeño en nota de prensa. </s><s> **"¿Se le puede tomar el pelo a la gente en campaña electoral?"**, ha planteado Monago, quien ha señalado que se están riendo de los extremeños diciéndoles que una empresa de 14.000 euros de capital social va a generar tres_millones de turistas al año y va a hacer rascacielos en Castilblanco.

The instances in Tab. 5 show distinctly different results before and after the automatic conversion of the source-text somatism for both NMT systems. In the case of DeepL, the ST discontinuous idiom could not properly be detected, leading to an inappropriate omission of the whole sequence, but after the conversion through *gApp* the continuous idiom could then be identified and an adequate equivalent was provided for the target text. In Google Translate, the ST idiom

Tab. 5: DeepL and Google Translate's outcomes before and after the conversion of the ST idiom *tomar el pelo* with *gApp*

	NMT outcomes
TT [EN] Discontinuous form in DeepL (before *gApp*)	This is what is stated, said Monago, in the United Kingdom's commercial register, which reflects an annual report and the financial status of the company, according to the Extremadura PP in a press release. </Monago said that they are making fun of the people of Extremadura by telling them that a company with 14,000 euros of share capital will generate three million tourists a year and will build skyscrapers in Castilblanco.
TT [EN] Continuous form in DeepL (after *gApp*)	This is what is stated, said Monago, in the United Kingdom's commercial register, which reflects an annual report and the financial status of the company, according to the Extremadura PP in a press release. </**Can you make fun of people in an election campaign,"** said Monago, who pointed out that they are laughing at the people of Extremadura by telling them that a company with 14,000 euros of share capital will generate three million tourists a year and will build skyscrapers in Castilblanco?
TT [EN] Discontinuous form in Google Translate (before *gApp*)	This is stated, Monago has said, in the commercial register of the United Kingdom, which reflects an annual report and the financial status of the company, as reported by the Extremadura PP in a press release. </s> <s> **"Can people tease people in an electoral campaign?"** asked Monago, who pointed out that they are laughing at Extremadurans telling them that a company with 14,000 euros of share capital is going to generate three_millions of tourists a year and will make skyscrapers in Castilblanco.
TT [EN] Continuous form in Google Translate (after *gApp*)	This is stated, Monago has said, in the commercial register of the United Kingdom, which reflects an annual report and the financial status of the company, as reported by the Extremadura PP in a press release. </s> <s> **"Can you fool people in the electoral campaign?"** asked Monago, who pointed out that they are laughing at Extremadurans telling them that a company with 14,000 euros of share capital is going to generate three_millions of tourists a year and will make skyscrapers in Castilblanco.

was detected both in its discontinuous and continuous form. However, before *gApp*, the prepositional phrase *a la gente* ('to the people') following the head verb *tomar* led to its erroneous interpretation both as the subject and the object of the source-text VNC *tomar el pelo*. This prompted an inappropriate translation outcome with the agent and the recipient concomitantly being the same person

"Can people tease people in an electoral campaign?". After the conversion through *gApp*, a straightforward analysis of the source text could be performed by Google Translate, which resulted in a more adequate translation outcome *"Can you fool people in the electoral campaign?"*, with an analogous effect to the impersonal ST sentence *"¿Se le puede tomar a la gente el pelo en campaña electoral?"*.

Finally, in the case of *verse las caras* (Fig. 5), *gApp* converted 76 forms, 70 of which were true positives and 6 were false positives. Consequently, this tool achieved a precision of 92.1% (70/76 cases) and a recall of 100% (70/70 cases). This recall of all the true-positive cases hence resulted in an equal performance of both the automatic and the manual conversion for DeepL (60 cases each) and Google Translate (48 cases each). The comparison of both the continuous and discontinuous scenarios shows that *gApp* led to a considerable improvement for DeepL (17.1%) and, even more significantly, for Google Translate (22.9%).

After all 560 cases have been analysed and classified, the final results are displayed in Fig. 6. For DeepL, the conversion into the continuous form through *gApp* led to an amelioration of 10% (28-case difference), whereas its manual conversion increased its accuracy by 13.2% (37-case divergence). Regarding Google Translate's performance, a further differential improvement can be perceived: 14.6% (41 cases) with *gApp* and 16.8% (47 cases) with the manual conversion.

Finally, the global results (Fig. 7), combining both NMTs' outcomes, show an ultimate contrastive enhancement of 12.3% (69 cases) with *gApp* and of 15% (84 cases) with the manual conversion.

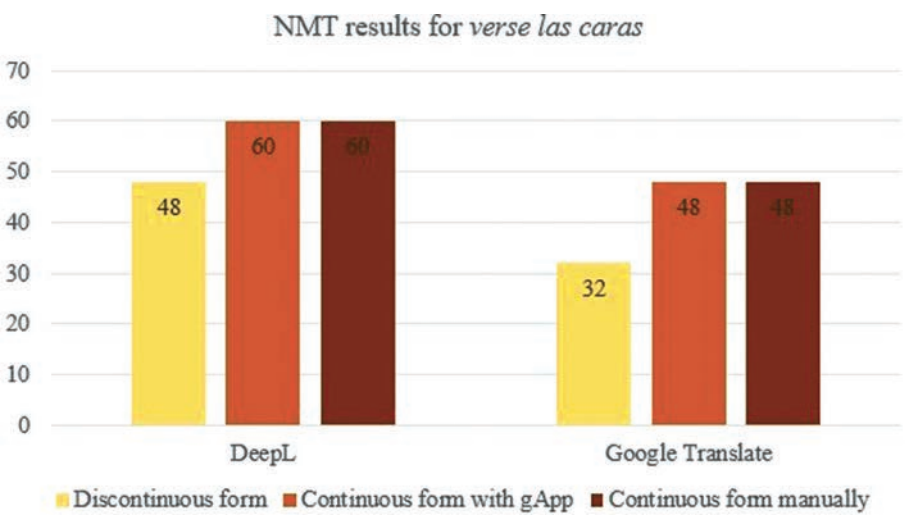

Fig. 5: NMT results for *verse las caras*

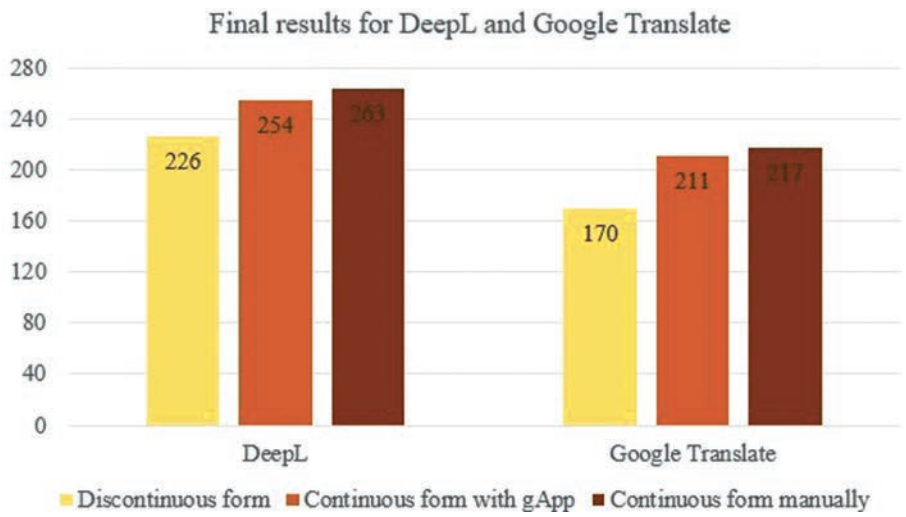

Fig. 6: Final results for DeepL and Google Translate

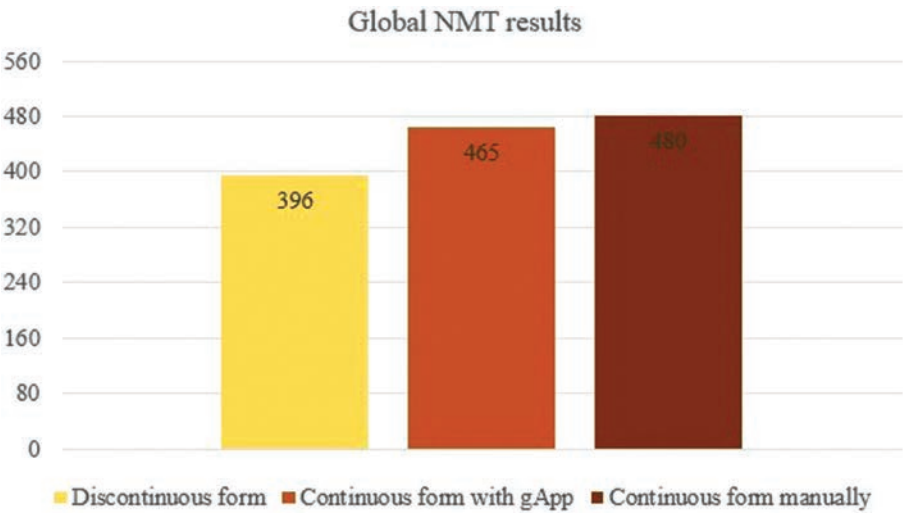

Fig. 7: Global NMT results

6 Discussion

The overall NMT results show that *gApp*'s preprocessing system nearly attains the manual conversion for the discontinuous somatisms under study with only a 2.7% difference. This quasi-analogous performance is chiefly due to *gApp*'s

refined detection system both in terms of final average precision (94.8%) and recall (96.8%), which means that only 5.2% of the global irrelevant results could penetrate the system and solely 3.2% of all the total relevant results could not be automatically identified.

In order to attain these high accuracy rates in terms of precision and recall some refinements needed to be implemented in *gApp*'s matching system. In spite of its advanced NLP pipeline, Spacy still encountered serious difficulties to automatically detect those cases in which the head verb occurred with an enclitic pronoun, which led to an erroneous POS tagging of these somatisms and, thus, resulted in non-identification. This obstacle hence necessitated the inclusion of regular expressions to also encompass all those relevant results. Despite the insertion of regex, Spacy was still unable to properly detect some instances containing orthotypographic mistakes (such as the absence of written accents when they were due or the non-capitalisation of proper names, inter alia) or rare words (such as *Vd.*, abbreviation of *usted*, i.e. Spanish second person singular pronoun employed in formal forms of address), which emphasises the need for optimising state-of-the-art POS taggers both regarding non-canonical UGC and rare and out-of-vocabulary words (Derczynski et al. 2013; Neunerdt et al. 2013; Gui et al. 2017).

In spite of these limitations in *gApp*'s detection system, the overall NMT results permit us to observe that the tool still offered a considerable enhancement in the translation of somatisms in relative terms, i.e. when comparing the discontinuous versus the continuous scenario after the idiom's automatic conversion, with an average amelioration of 10% (28-case difference) for DeepL and 14.6% (41-case divergence) for Google Translate. Nevertheless, this amelioration was still far from an optimal performance in absolute terms: whereas the continuous form through *gApp* attained a final average accuracy rate of 90.7% (254/280 cases) for DeepL, in the case of Google Translate its overall performance could only amount to 75.4% (211/280 cases). Two somatisms were particularly challenging for this latter NMT system: *verse las caras*, with an accuracy rate of 68.6% (48/70 cases) for the continuous form through *gApp* and, specially, *sentar la cabeza*, with a final score of 50% (35/70 cases). These results hence accentuate that, despite eliminating the obstacle of discontinuity, a thoroughly accurate performance still necessitates a further optimisation of the NMT systems regarding MWE identification in all its facets.

7 Conclusion

The text preprocessing system *gApp* has proved to optimise NMT performance for the somatisms under study, with an ultimate average enhancement of 12.3% for the analysed NMT systems. It has also proved to attain quasi-analogous results

to the manual conversion, which in contrast achieved a global 15% improvement. These promising results with semi-fixed VNC somatisms in the Spanish into English directionality invite to further expand *gApp*'s detection lexicon and conversion mechanism in order to examine to what extent it can also lead to NMT enhancement for other MWE categories affected by discontinuity as well as for other NMT systems.

Additionally, the present study can also lay the groundwork for further research to determine the scalation of this model to other language-dependent text preprocessing systems for the automatic conversion of discontinuous MWEs in syntactically flexible languages, in order to enhance MWE-aware NMT systems.

Acknowledgements

This paper has been carried out in the framework of various research projects on language technologies applied to translation and interpretation (ref. FFI2016-75831-P, UMA18-FEDERJA-067, CEI-RIS3 and EUIN2017-87746). It has also been funded by the Spanish Ministry of Education (FPU16/02032).

References

Alegria, Iñaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola & Ruben Urizar. 2004. Representation and treatment of multiword expressions in Basque. In *Proceedings of the second ACL workshop on multiword expressions: Integrating processing*, 48–55. https://www.aclweb.org/anthology/W04-0407.pdf

Al Saied, Hazem, Mathieu Constant & Marie Candito. 2017. The ATILF-LLF system for Parseme shared task: a transition-based verbal multiword expression tagger. In *Proceedings of the 13th workshop on multiword expressions (MWE 2017)*, 127–132. https://www.aclweb.org/anthology/W17-1717.pdf

Al Saied, Hazem, Marie Candito & Mathieu Constant. 2019. Comparing linear and neural models for competitive MWE identification. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 86–96. https://www.aclweb.org/anthology/W19-6109.pdf

Anastasopoulos, Antonios. 2019. An analysis of source-side grammatical errors in NMT. In *Proceedings of the 2019 ACL workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, 213–223. https://www.aclweb.org/anthology/W19-4822

Bejček, Eduard, Pavel Straňák & Daniel Zeman. 2011. Influence of treebank design on representation of multiword expressions. In Alexander F. Gelbukh (ed.), *Computational Linguistics and intelligent text processing – 12th international conference, CICLing 2011, vol. 6608 (Lecture notes in Computer Science)*, 1–14. Berlin & Heidelberg: Springer.

Bejček, Eduard, Pavel Straňák & Pavel Pecina. 2013. Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In *Proceedings*

of the 9th workshop on multiword expressions, 106–115. https://www.aclweb.org/anthology/W13-1016.pdf

Belinkov, Yonatan & Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. *ArXiv*. https://arxiv.org/abs/1711.02173

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner & Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics* 43(4). 1–92.

Corpas Pastor, Gloria & Jean-Pierre Colson (eds.). 2020. *Computational phraseology* (IVITRA Research in Linguistics and Literature, 24). Amsterdam & Philadelphia: John Benjamins.

Derczynski, Leon, Alan Ritter, Sam Clark & Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: overcoming sparse and noisy data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, 198–206. http://www.aclweb.org/anthology/R13-1026

Finlayson, Mark & Nidhi Kulkarni. 2011. Detecting multiword expressions improves word sense disambiguation. In *Proceedings of the eighth ALC workshop on multiword expressions (MWE 2011)*, 20–24. https://www.aclweb.org/anthology/W11-0805.pdf

Foufi, Vasiliki, Luca Nerima & Eric Wehrli. 2019. Multilingual parsing and MWE detection. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 217–237. Berlin: Language Science Press.

Gui, Tao, Qi Zhang, Haoran Huang, Minlong Peng & Xuanjing Huang. 2017. Part-of-speech tagging for twitter with adversarial neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2411–2420. https://www.aclweb.org/anthology/D17-1256.pdf

Hidalgo-Ternero, Carlos Manuel. 2020 (forthcoming). Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation. *MonTI* 6 (Special issue, "Análisis multidisciplinar del fenómeno de la variación en traducción e interpretación / Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting").

Honnibal, Matthew & Inés Montani. 2017 (to appear). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Huang, Po-Sen, Chong Wang, Sitao Huang, Denny Zhou & Li Deng. 2018. Towards neural phrase-based machine translation. Paper presented at the sixth International Conference on Learning Representations (ICLR), Vancouver Convention Center, 30 April–3 May 2018. https://arxiv.org/pdf/1706.05565.pdf

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovvář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2003. *The Sketch Engine*. https://www.sketchengine.eu (accessed 4 March 2020)

Klyueva, Natalia, Antoine Doucet & Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proceedings of the 13th workshop on multiword expressions (MWE 2017)*, 60–65. https://www.aclweb.org/anthology/W17-1707.pdf

Maldonado, Alfredo, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel & Qun Liu. 2017. Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on multiword expressions (MWE 2017)*, 114–120. https://www.aclweb.org/anthology/W17-1715.pdf

Monti, Johanna, Violeta Seretan, Gloria Corpas Pastor & Ruslan Mitkov. 2018. Multiword units in machine translation and technology. In Ruslan Mitkov, Johanna Monti, Gloria Corpas Pastor & Violeta Seretan (eds.), *Multiword units in translation and translation technology*, 1–37. Amsterdam: John Benjamins.

Moreau, Erwan, Ashjan Alsulaimani, Alfredo Maldonado & Carl Vogel. 2018. CRF-Seq and CRFDepTree at PARSEME Shared Task 2018: Detecting verbal MWEs using sequential and dependency-based approaches. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 241–247. https://www.aclweb.org/anthology/W18-4926.pdf

Nagy, István T. & Veronika Vincze. 2014. VPCTagger: Detecting verb-particle constructions with syntax-based methods. In *Proceedings of the 10th workshop on multiword expressions (MWE 2014)*, 17–25. https://www.aclweb.org/anthology/W14-0803.pdf

Neunerdt, Melanie, Bianka Trevisan, Michael Reyer & Rudolf Mathar. 2013. Part-of-speech tagging for social media texts. In Iryna Gurevych, Chris Biemann & Torsten Zesch (eds.), *Language processing and knowledge in the web. Lecture notes in computer science 8105*, 139–150. Berlin & Heidelberg: Springer.

Niu, Xing, Prashant Mathur, Georgiana Dinu & Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. *ArXiv*. https://arxiv.org/pdf/2005.00580.pdf

Ramisch, Carlos. 2015. *Multiword expressions acquisition: A generic and open framework* (Theory and applications of natural language processing series XIV). Cham: Springer.

Ramisch, Carlos, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya & Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 222–240. https://www.aclweb.org/anthology/W18-4925.pdf

Ramisch, Carlos & Aline Villavicencio. 2018. Computational treatment of multiword expressions. In Ruslan Mitkov (ed.), *Oxford handbook of Computational Linguistics* (2nd edn). N. p.: Oxford University Press.

Riedl, Martin & Chris Biemann. 2016. Impact of MWE resources on multiword recognition. In *Proceedings of the twelfth workshop on multiword expressions (MWE 2016)*, 107–111. https://www.aclweb.org/anthology/W16-1816.pdf

Rikters, Matīss & Ondřej Bojar. 2017. Paying attention to multi-word expressions in neural machine translation. *ArXiv*. https://arxiv.org/abs/1710.06313

Rohanian, Omid, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha & Ruslan Mitkov. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. I*, 2692–2698. https://www.aclweb.org/anthology/N19-1275.pdf

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbukh (ed.), *Computational Linguistics and intelligent text processing. CICLing 2002. Lecture Notes in Computer Science*, 1–15. Berlin & Heidelberg: Springer.

Schneider, Nathan, Emily Danchik, Chris Dyer & Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *TACL* 2. 193–206.

Sperber, Matthias, Jan Niehues & Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *International Workshop on Spoken Language Translation (IWSLT)*. https://pdfs.semanticscholar.org/88ed/f12127a628bed608cae0bdf3700d00824df4.pdf

Wang, Xing, Zhaopeng Tu, Deyi Xiong & Min Zhang. 2017. Translating phrases in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2017*), 1421–1431. https://www.aclweb.org/anthology/D17-1149.pdf

Zampieri, Nicolas, Carlos Ramisch & Geraldine Damnati. 2019. The impact of word representations on sequential neural MWE identification. *Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 169–175. https://www.aclweb.org/anthology/W19-5121.pdf

Zaninello, Andrea & Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 3816–3825. https://www.aclweb.org/anthology/2020.lrec-1.471.pdf