GOOGLE TRANSLATE VS. DEEPL: ANALYSING NEURAL MACHINE TRANSLATION PERFORMANCE UNDER THE CHALLENGE OF PHRASEOLOGICAL VARIATION[1]

CARLOS MANUEL HIDALGO-TERNERO
cmhidalgo@uma.es
University of Malaga (Spain)

Abstract

The present research analyses the performance of two free open-source neural machine translation (NMT) systems —Google Translate and DeepL— in the (ES>EN) translation of somatisms such as *tomar el pelo* and *meter la pata,* their nominal variants (*tomadura/tomada de pelo* and *metedura/metida de pata*), and other lower-frequency variants such as *meter la pata hasta el corvejón*, *meter la gamba* and *metedura/metida de gamba*. The machine translation outcomes will be contrasted and classified depending on whether these idioms are presented in their continuous or discontinuous form (Anastasiou 2010), i.e., whether different n-grams split the idiomatic sequence (or not), which may pose some difficulties for their automatic detection and translation.

Overall, the insights gained from this study will prove useful in determining for which of the different scenarios either Google Translate or DeepL delivers a better performance under the challenge of phraseological variation and discontinuity.

Keywords: Neural machine translation; Phraseological variation; Discontinuity; Somatisms.

Resumen

La presente investigación tiene por objetivo analizar el rendimiento de dos sistemas de traducción automática neuronal (TAN) —Google Translate y DeepL— en la traducción (ES>EN) de somatismos tales como *tomar el pelo* y *meter la pata,* sus variantes nominales (*tomadura/tomada de pelo* y *metedura/metida de pata*) y otras variantes con menor frecuencia de aparición tales como *meter la pata hasta el corvejón*, *meter la gamba* y *metedura/metida de gamba*. Los resultados de ambos sistemas se contrastarán y clasificarán según los diferentes somatismos se presenten en su forma continua o discontinua (Anastasiou 2010), es decir, dependiendo de si distintos n-gramas dividen la secuencia idiomática (o no), lo que puede presentar serias dificultades para su detección y traducción automáticas. En general, los resultados obtenidos permitirán dilucidar en cuál de los diferentes escenarios Google Translate o DeepL ofrece un mejor rendimiento ante el desafío de la variación y discontinuidad fraseológica.

Palabras clave: Traducción automática neuronal; Variación fraseológica; Discontinuidad; Somatismos.

1. Introduction

The not so paradoxical instability of fixed expressions poses serious challenges for Natural Language Processing in general, and Machine Translation in particular. In this regard, one of the fertile grounds for debate in computational linguistics is how to approach these fixed expressions, admitting complete fixedness is the exception.

In this context, Sag, Baldwin et al. (2002) have thoroughly elaborated on the two opposing (and yet complementary) approaches in the computational treatment of multiword expressions (MWE). On the one hand, a *fully compositional account*, which may experience no problem with lexical variation, but it does with idiomaticity (for example, interpreting the "joke" semantics in *pull someone's leg*) and overgeneration (for example, in creating

unregistered instances such as *the leg was pulled*). On the other hand, the *word-with-spaces* approach suffers from a flexibility problem as it does not properly capture phraseological and inflectional variation, and it also impedes lexical proliferation (e.g. light-verb constructions such as *make a payment, make a deposit*, etc. would have to be listed individually, which entails an important lack of generality and prediction) (Sag, Baldwin et al. 2002: 2-3). This approach does not consider either that MWEs do not always consist of adjacent tokens (e.g. _take my piece of advice into account_), which hence poses further serious difficulties for the automatic detection of these expressions (Rohanian, Taslimipoor et al. 2019).

Against this background, it will first be necessary to analyse the somatisms (i.e., idioms containing terms that refer to human or animal body parts) under study in order to determine to which category it is possible to relate them. These somatisms include *tomar el pelo* (literally, 'to take someone's hair'; figuratively, 'to fool someone'), and *meter la pata* (lit., 'to put the leg in'; fig., 'to mess things up'), and their variants. In the case of *tomar el pelo*, only the nominal variants, *tomadura de pelo* and *tomada de pelo*, retrieve a significant number of concordances in the analysed corpora. Regarding other possible variants such as *tomar el cabello* or *tomar los pelos*, a very scarce number of concordances of them as idioms has been recalled: 5 in the case of *tomar el cabello* and 3 for *tomar los pelos*. The latter ones can therefore be considered as *anti-collocations*, i.e., "potential lexical variants of a given institutionalized phrase which are observed with zero or markedly low frequency" (Pearce, in Sag, Baldwin et al. 2002: 7).

Concerning the idiom *meter la pata*, it was possible to detect a wider range of variants: not only *variants through syntagmatic relation* (Koike 2007) such as *metedura de pata* and *metida de pata*, but also through insertion, *meter la pata hasta el corvejón*, and *through paradigmatic relation* (Koike 2007), in a relation of co-hyponymy, such as *meter la gamba*, which also holds two nominal variants, *metedura de gamba* and *metida de gamba*. These latter variants are diaphasically marked as informal, and their appearance usually connotes irony or humour, as it was possible to observe in the concordances retrieved from the analysed corpora.

Regarding the idioms *tomar el pelo* and *meter la pata*, and their variants, the different retrieved concordances reveal that both can undergo verb inflection (*tomo, tomas, tomó… el pelo*) but they are not subject to noun inflection (**tomar los pelos*, **meter las patas*). In these sequences, *pelo* and *pata* are only modified by the definite article (**tomar un pelo, *tomar otro pelo, *meter una pata...*). Therefore, both sequences can be considered as *semi-fixed* MWEs as they can undergo both morphosyntactic and internal alterations from their canonical form, unlike other MWEs such as *by and large, in vitro* or *in a nutshell,* which are hence designated as *fixed expressions* (Sag, Baldwin et al. 2002: 4).

As *semi-fixed* MWEs, both somatisms (*tomar el pelo* and *meter la pata*) and their variants (*tomadura de pelo, metedura de pata, meter la gamba…*) present a certain degree of syntactic flexibility. The different concordances reveal that adverb phrases such as *más* ('more'), *tanto* ('so much') or *un poco* ('a bit'), noun phrases such as *otra vez* ('again') or the subject, among others, consistently occur within (and hence split) the sequences in the verb-phrase variants of the idioms (*tomar el pelo*, *meter la pata* and *meter la gamba*). For the noun-phrase variants, the sequence is regularly split by different adjective modifiers such as *total* ('total'), *descomunal* ('enormous') or *constante* ('constant') in the case of *tomadura/tomada de pelo*, and *total*, *grave* ('serious'), or *histórica* ('historic') in the case of *metedura/metida de pata*. Therefore, whereas it would not be possible to capture the full range of lexical variants of these idioms if being treated as *words-with-spaces*, a *fully compositional account* would overgenerate anti-collocations such as *tomar los pelos* or *tomar los cabellos*, all of which thus necessitates employing a hybrid approach (Sag, Baldwin et al. 2002).

Throughout this research we will thus distinguish between continuous or discontinuous forms of the idiom (Anastasiou 2010) depending on whether or not the sequence is split by those different n-grams. One of the main objectives of our research will hence be to determine in which cases and to what extent discontinuity poses some challenges for the automatic detection and translation of the idioms. The ultimate objective will be to analyse and contrast the performance of two neural machine translation (NMT) systems, Google Translate vs. DeepL, when accurately detecting and translating the continuous and discontinuous forms of the somatisms *tomar el pelo* and *meter la pata,* their nominal variants (*tomadura de pelo, tomada de pelo, metedura de*

pata, and *metida de pata*), as well as other variants such as *meter la pata hasta el corvejón*, *meter la gamba, metedura de gamba*, and *metida de gamba*, in order to determine which of these NMT systems delivers a better performance under the challenge of phraseological variation and discontinuity.

The overall structure of this paper takes the form of five sections. Once the somatisms under study have been introduced in section 1, they will be further analysed in section 2, shifting the focus onto their diatopic variation. Section 3 begins by describing the methodology of the present research, and section 4 exposes the results retrieved from both NMT systems when facing the challenge of phraseological variation and discontinuity. In this regard, the results will be contrasted and classified depending on whether these idioms are presented in their continuous or discontinuous form (Anastasiou 2010), i.e., whether different n-grams split the idiomatic sequence (or not). In this line, section 5 analyses the repercussions of the retrieved results for the NMT systems and, finally, the conclusions of the present research will be displayed in section 6.

2. Diatopic variation of the somatisms under study

In order to exhibit the diatopic variation of the somatisms under study, Corpas Pastor's (2015 & 2018) corpus-based research protocol will be implemented. In this regard, the somatisms have been retrieved and analysed from the monolingual corpora esTenTen (general Spanish), eseuTenTen (European Spanish), and esamTenTen (Latin American Spanish). The latter includes subcorpora corresponding to the 18 American national varieties. These corpora, belonging to the TenTen family, are available through Sketch Engine, a language corpus management and query system with 500 corpora in more than ninety different languages. Comprising over 17.5 billion words, esTenTen corpus (including eseuTenTen and esamTenTen) is compiled from texts retrieved from the Internet. These corpora are lemmatised and part-of-speech tagged (Kilgarriff, Baisa et al. 2003).

The different somatisms will be displayed in Table 1 along with their raw/normalised frequencies for each of the 19 Spanish national varieties (identified by their country code top-level domain: .ar, .bo, .cl, .co, .es, etc.). Whereas the first column, representing the *raw frequency*, refers to the total

number of occurrences of a specific somatism, the second one (in bold) stands for the *normalised frequencies*, presented as percentage scores per million tokens. Analysing normalised frequencies to a common base is essential when contrasting results retrieved from corpora of dissimilar sizes (Corpas Pastor 2018).

	Tomar el pelo		Tomadura de pelo		Tomada de pelo		Meter la pata		Metedura de pata		Metida de pata		Meter la gamba		Metedura de gamba		Metida de gamba		Meter la pata hasta el corvejón	
.AR	3775	**1.13**	388	**0.12**	1322	**0.4**	2651	**0.79**	159	**0.05**	710	**0.21**	80	**0.02**	1	**>0.01**	10	**>0.01**	1	**>0.01**
.BO	29	**0.45**	13	**0.2**	3	**0.05**	40	**0.62**	3	**0.05**	41	**0.64**	–	–	–	–	–	–	–	–
.CL	576	**0.49**	234	**0.2**	63	**0.05**	587	**0.5**	55	**0.05**	582	**0.5**	–	–	–	–	1	**>0.01**	–	–
.CO	157	**0.32**	27	**0.05**	7	**0.01**	99	**0.2**	17	**0.03**	134	**0.27**	–	–	–	–	–	–	–	–
.CR	27	**0.43**	14	**0.22**	2	**0.03**	33	**0.52**	4	**0.06**	33	**0.52**	–	–	–	–	–	–	–	–
.CU	125	**0.45**	69	**0.25**	–	–	141	**0.5**	74	**0.26**	24	**0.09**	–	–	–	–	–	–	–	–
.DO	54	**0.92**	13	**0.22**	–	–	141	**2.39**	12	**0.2**	58	**0.99**	1	**0.02**	–	–	–	–	–	–
.EC	36	**0.41**	45	**0.51**	1	**0.01**	43	**0.49**	16	**0.18**	12	**0.14**	–	–	–	–	–	–	–	–
.SV	9	**0.25**	7	**0.19**	–	–	18	**0.49**	6	**0.16**	22	**0.6**	–	–	–	–	–	–	–	–
.GT	22	**0.61**	4	**0.11**	1	**0.03**	63	**1.75**	2	**0.06**	57	**1.58**	–	–	–	–	–	–	–	–
.HN	6	**0.57**	4	**0.38**	–	–	–	–	3	**0.28**	–	–	–	–	–	–	–	–	–	–
.MX	1240	**0.63**	455	**0.23**	488	**0.25**	1054	**0.53**	101	**0.05**	459	**0.23**	3	**>0.01**	–	–	–	–	1	**>0.01**
.NI	32	**0.46**	5	**0.07**	1	**0.01**	40	**0.57**	5	**0.07**	49	**0.7**	–	–	–	–	–	–	–	–
.PA	13	**0.62**	4	**0.19**	–	–	45	**2.14**	3	**0.14**	38	**1.81**	–	–	–	–	–	–	–	–
.PY	28	**0.4**	15	**0.22**	1	**0.01**	41	**0.59**	4	**0.06**	22	**0.32**	–	–	–	–	–	–	–	–
.PE	264	**0.78**	111	**0.33**	7	**0.02**	642	**1.9**	21	**0.06**	482	**1.43**	–	–	–	–	–	–	6	**0.02**
.UY	496	**2.35**	168	**0.8**	180	**0.85**	221	**1.05**	9	**0.04**	87	**0.41**	2	**0.01**	–	–	–	–	–	–
.VE	57	**0.19**	27	**0.09**	2	**0.01**	129	**0.44**	28	**0.1**	77	**0.26**	1	**>0.01**	–	–	3	**0.01**	–	–
.ES	4429	**1.89**	3356	**1.43**	29	**0.01**	4523	**1.93**	1979	**0.84**	220	**0.09**	276	**0.12**	17	**0.01**	9	**>0.01**	41	**0.02**

Table 1: Raw and normalised frequencies of the somatisms under study in the 19 national varieties of *esTenTen*

The results in Table 1 exhibit distinctly different frequency scores for the somatisms across national varieties. A significant number of occurrences for the verbal variants (*tomar el pelo* and *meter la pata*) has been detected in the 19 national varieties, albeit with different frequency scores: whereas *tomar el pelo* displays a higher normalised frequency in the dialectal areas of Uruguay (2.35), Spain (1.89) and Argentina (1.13), *meter la pata* is more frequent in the varieties of Dominican Republic (2.39), Panama (2.14), Spain (1.93) and Peru (1.9).

Regarding the nominal variants, quite revealing results have been retrieved on the frequency scores of *tomadura de pelo* vs. *tomada de pelo*, and *metedura de pata* vs. *metida de pata*. While, in the first dichotomy, *tomadura de pelo* appears as the most frequent variant in most national varieties, except for Argentina (0.12 vs. 0.4), Mexico (0.23 vs. 0.25), and Uruguay (0.8 vs. 0.85);

in the second opposition, *metida de pata* is the variant with the higher nor-malised frequency across national varieties, with the exception of Cuba (0.26 vs. 0.09), Ecuador (0.18 vs. 0.14) and Spain (0.84 vs. 0.09).

With respect to the somatism *meter la gamba*, the results permit to observe that its occurrence is predominant in European Spanish (0.12), distantly followed by the national variety of Argentina (0.02), Dominican Republic (0.02) and Uruguay (0.01). Its nominal variants, *metedura de gamba* and *metida de gamba*, exhibit low frequency scores in the few varieties presenting occurrences for these sequences: 0.01 in European Spanish in the case of *metedura de gamba* and 0.01 in the national variety of Venezuela for *metida de gamba*. Finally, regarding the variant through insertion *meter la pata hasta el corvejón,* only significant occurrences have been detected in the national varieties of Spain and Peru (0.02 each). In the case of Peru, 3 out of the 6 retrieved occurrences presented instances of another variant of this somatism, *meter la pata hasta los corvejones.*

Once the diatopic variation of the somatisms under study has been ana-lysed, one of the objectives of the research will be to examine whether, in all the cases, there is a direct correspondence between a higher frequency score of a given somatism across national varieties and a better performance of the NMT system, or whether other factors enter the scene, such as some specific national varieties being privileged over others in the translation of these idioms. The final results will also shed some light on these NMTs' performance regarding rare and out-of-vocabulary (OOV) words, as is the case of the lower-frequency variants *meter la pata hasta el corvejón*, *meter la gamba*, and their nominal variants *metedura de gamba* and *metida de gamba*.

3. Methodology

Following a corpus-based methodology, the cases under study have also been retrieved from the corpus esTenTen through Sketch Engine. In this context, it is necessary to take into account that this corpus is created from a wide compilation of texts collected from the web and hence it includes linguisti-cally valuable text corpora from heterogeneous sources ranging from NMT canonical parallel training data to non-canonical user-generated content

(UGC), with the acknowledged challenges that UGC's pervasive source-text error, noise and out-of-vocabulary tokens, inter alia, still pose to even the most robust NMT systems (Sperber, Niehues et al. 2017; Belinkov & Bisk 2018; Anastasopoulos 2019; Lohar, Popović et al. 2019). This heterogeneity in terms of text sources, types and diatopic varieties has however been maintained in the retrieval and analysis process in order to avoid sampling bias, which would otherwise originate from exclusively selecting a sole text typology for the somatisms under study. Since the main objective of the present research is to determine these NMT systems' performance under the challenge of phraseological variation and discontinuity, the results have been manually assessed following a reference-based MT evaluation with several possible target-text candidates for each of the somatisms under study in their continuous and discontinuous form for the different contexts. In this way, those results presenting morphological, syntactic, and/or orthotypographic dissimilarities or containing either source-text or translation errors (omissions, additions, etc.) which affected other elements in the sentences were unified within the same category depending on the NMTs' performance in the translation of the analysed source-text somatisms under analogous phenomena (continuous or discontinuous form split by a subject, object, adjective modifier, etc.)

In this context, in order to retrieve the discontinuous forms of the somatisms under study, Sketch Engine's CQL (Corpus Query Language) code has been employed. CQL is a special query code to search for complex lexico-grammatical patterns which would otherwise be laborious to retrieve through the simple query system. In order to exclude as much noise as possible, the following CQL schemas have been employed:

Sequence	CQL schemas
"Tomar [1-2 n-grams] el pelo"	[lemma="tomar"][]{1,2}[word="el"][word="pelo"]
"Tomadura/tomada [1-2 n-grams] de pelo"	[lemma="tomadura|tomada"][]{1,2}[word="de"][word="pelo"]
"Meter [1-2 n-grams] la pata"	[lemma="meter"][]{1,2}[word="la"][word="pata"]
"Metedura/metida [1-2 n-grams] de pata"	[lemma="metedura|metida"][]{1,2}[word="de"][word="pata"]

Table 2: CQL schemas for the discontinuous form of the somatisms

Whereas the CQL code *lemma* finds examples of all inflected forms of the introduced token (e.g. *tomar, tomó, tomaría, tomaste* in the case of *tomar*), the code *word* retrieves the exact form of the token, without inflection. In the somatisms *tomar el pelo* and *meter la pata,* and their nominal variants (*tomadura de pelo*, *tomada de pelo, metedura de pata*, and *metida de pata*) only the verbs *tomar* and *meter*, and the nouns *tomadura/tomada* and *metedura/metida*, respectively, allow inflection. That is the reason why, in order to refine the concordances retrieved in esTenTen, the CQL code *lemma* has only been employed for those elements in the idiom allowing inflection, and the code *word* has been selected for the rest.

In order to search for the different n-grams that may split the sequence, the CQL codes "[]{1,2}" have regularly been used, which would retrieve instances of unigrams or bigrams within the sequence. In two specific cases the number of n-grams has been augmented to "[]{1,4}": when searching for indirect objects within the sequence *tomar el pelo,* and when seeking several different n-grams dividing the verb-phrase variants of the idioms *tomar el pelo* and *meter la pata,* all of which can contain a number of n-grams higher than 2.

The retrieved concordances, containing discontinuous somatisms, have been stored and classified depending on which n-gram split the sequence. Two free open-source NMT systems —Google Translate and DeepL— have then been employed to translate these somatisms in their discontinuous scenarios. These sequences have subsequently been manually altered into their continuous form in order to evaluate and contrast the performance of both

NMT systems against both the continuous and discontinuous display of the somatisms in identical contexts, with the purpose of securing that a dissimilar performance in the NMT systems was chiefly caused by phraseological variation and discontinuity, discarding other possible factors. Against such a background, the different translation outcomes will be presented in the following section.

4. Results

A total of 380 cases has been analysed, considering the continuous and discontinuous forms of the idioms *tomar el pelo, meter la pata* and their nominal variants (*tomadura de pelo, tomada de pelo, metedura de pata*, and *metida de pata*) split by different modifiers (*tanto*, *un poco*, *otra vez*, etc.), as well as other variants such as *meter la pata hasta el corvejón*, *meter la gamba, metedura de gamba* and *metida de gamba*.

In the case of *tomar el pelo*, 140 cases have been examined (70 for the continuous and 70 for the discontinuous form of the sequence). The itemised results for *tomar el pelo* (Chart 1) display notable differences between both NMT systems' performance. Whereas the inclusion of modifiers such as *más* (9 vs. 10)[2] or the indirect object (6 in both) only presented minor dissimilarities between both NMT systems, DeepL delivered a better performance in those cases where the sequence was split by other modifiers such as *tanto* (6 vs. 10), *otra vez* (7 vs. 10), and the subject (7 vs. 9). In two other scenarios, it was Google Translate that offered more suitable equivalents: when the sequence was divided by *un poco* (6 vs. 4) and more significantly when several different n-grams were included within the sequence (10 vs. 3). An example of these NMTs' performance in the translation of the continuous and discontinuous form of *tomar el pelo* will be illustrated in Table 3.
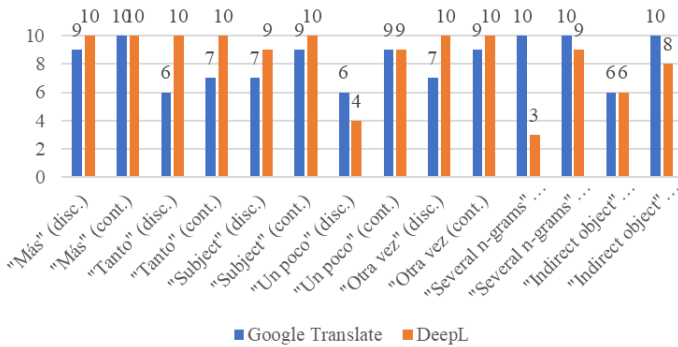
Itemised NMT results for *tomar el pelo*

Chart 1: Itemised NMT results for *tomar el pelo*

	KWIC extracts	Comments
ST [ES] Disc. form	En 1969, cuando una periodista le preguntó si este "¿no?" provenía de su herencia santiagueña (junto con las empanadas), ella respondió: "pare un poco, ya no lo uso más". Al respecto, decía: "<u>me tomaron tanto el pelo</u> en todas partes que lo tuve que dejar" (Mactas 1969: 55).	
TT [EN] Disc. form with DeepL	In 1969, when a journalist asked her if this "no" came from her Santiago heritage (along with the empanadas), she replied: "stop a little, I don't use it anymore". In this respect, she said: "<u>I was teased so much</u> everywhere that I had to give it up" (Mactas 1969: 55).	The ST idiom has been detected and an appropriate TT equivalent has been provided.
TT [EN] Disc. form with Google Translate	In 1969, when a journalist asked if this "right?" It came from her Santiago heritage (along with the empanadas), she replied: "Stop a little, I don't use it anymore". In this regard, he said: "<u>They took my hair so much</u> everywhere I had to leave it" (Mactas 1969: 55).	A literal and inadequate translation of the ST idiom has been provided: "they took my hair so much…" instead of "I was teased so much…"
ST [ES] Cont. form	En 1969, cuando una periodista le preguntó si este "¿no?" provenía de su herencia santiagueña (junto con las empanadas), ella respondió: "pare un poco, ya no lo uso más". Al respecto, decía: "<u>me tomaron el pelo tanto</u> en todas partes que lo tuve que dejar" (Mactas 1969: 55).	
TT [EN] Cont. form with DeepL	In 1969, when a journalist asked her if this "no" came from her Santiago heritage (along with the empanadas), she replied: "stop a little, I don't use it anymore". She said: "<u>I was teased so much</u> everywhere that I had to give it up" (Mactas 1969: 55).	The ST idiom has been detected and an appropriate TT equivalent has been provided (analogously to the discontinuous scenario).
TT [EN] Cont. form with Google Translate	In 1969, when a journalist asked if this "right?" It came from her Santiago heritage (along with the empanadas), she replied: "Stop a little, I don't use it anymore". In this regard, he said: "<u>They teased me so much</u> everywhere that I had to leave it" (Mactas 1969: 55).	In contrast to the discontinuous scenario with Google Translate, the ST idiom has now been detected and an appropriate TT equivalent has been provided.

Table 3: KWIC extracts with the translation of *tomar el pelo*

However, when analysing the total NMT results (Chart 2), only slight differences have been observed in the performance of Google Translate versus DeepL when adequately detecting and establishing equivalences for the idiom: 51 (73%) vs. 52 (74%) appropriate equivalents, respectively, for the discontinuous form, and 64 (92%) vs. 67 (96%) adequate equivalents for the continuous form of the sequence.
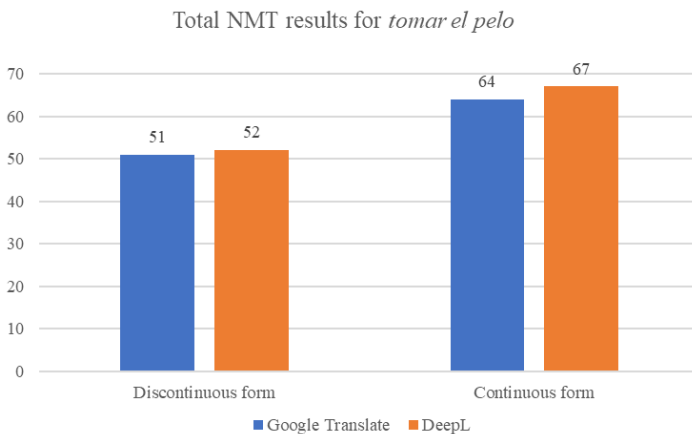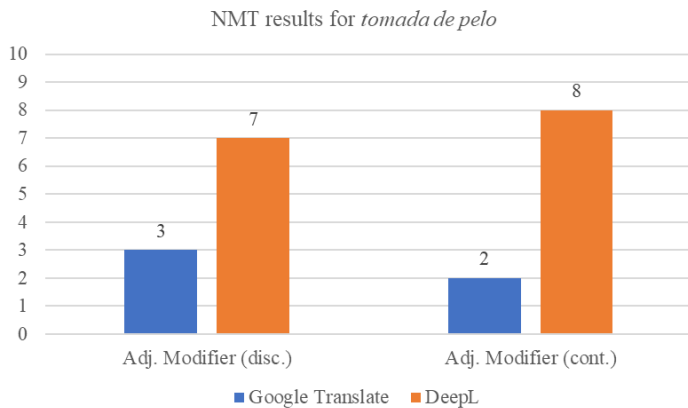

Chart 2: Total NMT results for *tomar el pelo*

When examining the results of the nominal variant *tomadura de pelo* (Chart 3), it can be observed that Google Translate provided more appropriate equivalents in the discontinuous form, when the sequence was split by an adjective modifier (9 vs. 7). Both NMT systems offered appropriate equivalents in all those concordances with the continuous form of *tomadura de pelo* (10 out of 10). Regarding the somatism *tomada de pelo* (Chart 4), while DeepL exhibited similar results, the performance of Google Translate dramatically diminished both in the discontinuous (3 vs. 7) and the continuous (2 vs. 8) form of the sequence.

NMT results for *tomadura de pelo*
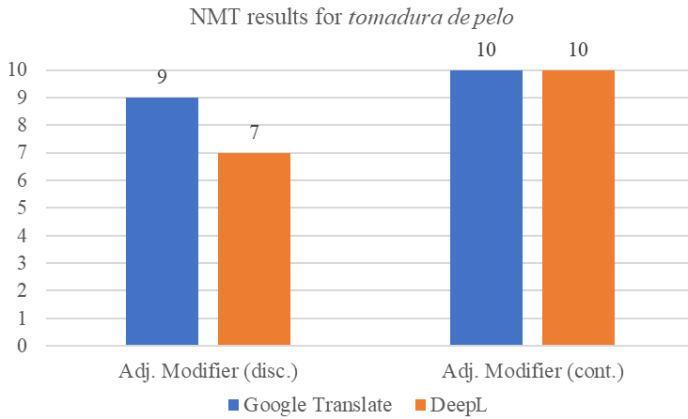
Chart 3: NMT results for *tomadura de pelo*

NMT results for *tomada de pelo*

Chart 4: NMT results for *tomada de pelo*

In the case of the idiom *meter la pata*, 120 cases have been analysed (60 for the continuous and 60 for the discontinuous form of the sequence). Concordances with the same modifiers as in the case of *tomar el pelo* were translated and examined, except for the absence of indirect object in the scenario of *meter la pata*, since it does not take the dative. The itemised

NMT results (Chart 5) exhibit, overall, a quasi-analogous outcome for both NMT systems. Google Translate and DeepL delivered a similar performance in those instances in which the sequence was split by modifiers such as *más* (9 in both), *tanto* (10), the subject or *un poco* (9 vs. 10), and *otra vez* (8 vs. 9). However, there was a specific case in which more dramatic differences could be observed: when several different n-grams divided the sequence (as those in the examples of Table 4), Google Translate achieved an accuracy rate of 5 out of 10, while DeepL's performance obtained a score of 9 out of 10.
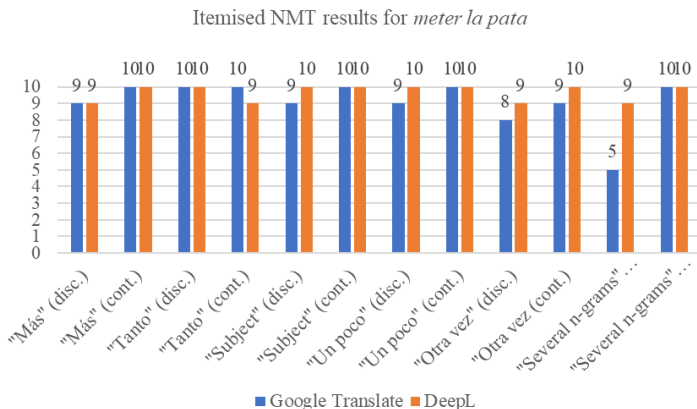

Chart 5: Itemised NMT results for *meter la pata*

	KWIC extracts	Comments
ST [ES] Disc. form	Una vez que hayamos hecho bien el trimming lo que toca ahora es el secado y curado de nuestras plantas de marihuana para obtener el aroma y fragancia adecuada, no podemos <u>meter aquí tampoco la pata</u>.	
TT [EN] Disc. form in DeepL	Once we have done the trimming well, it is now time to dry and cure our marijuana plants to obtain the right aroma and fragrance.	The ST sentence containing the discontinuous idiom has been omitted in the TT.
TT [EN] Disc. form in Google Translate	Once we have done the trimming well, what it touches now is the drying and curing of our marijuana plants to obtain the right aroma and fragrance, we can not put the leg here either.	A literal and inadequate translation of the ST discontinuous somatism has been provided in the TT.
ST [ES] Cont. Form	Una vez que hayamos hecho bien el trimming lo que toca ahora es el secado y curado de nuestras plantas de marihuana para obtener el aroma y fragancia adecuada, no podemos <u>meter la pata</u> aquí tampoco.	
TT [EN] Cont. form in DeepL	Once we've done the trimming right, it's time to dry and cure our marijuana plants to get the right aroma and fragrance, we can't screw up here either.	The ST continuous idiom has been detected and an appropriate equivalent has been provided.
TT [EN] Cont. form in Google Translate	Once we have done the trimming well, what it touches now is the drying and curing of our marijuana plants to obtain the right aroma and fragrance, we cannot screw up here either.	The ST continuous idiom has been detected and an appropriate equivalent has been provided.

Table 4: KWIC extracts with the translation of *meter la pata*

The global results for the sequence (Chart 6) also show that, analogously to *tomar el pelo*, no notable differences in the overall performance of both NMT systems have been detected: whereas Google Translate offered an accuracy rate of 83% (50 out of 60 cases) for the discontinuous form of the sequence, and of 98% (59/60 cases) for the continuous one, DeepL slightly outperformed Google Translate in the discontinuous case up to 95% (57/60 cases) but it showed a similar accuracy rate to Google Translate for the continuous form of the sequence (98%, i.e., 59/60 cases).
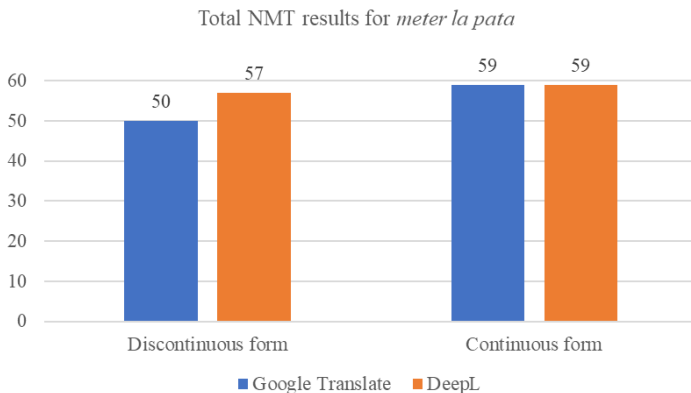
Total NMT results for *meter la pata*

Chart 6: Total NMT results for *meter la pata*

Furthermore, as can be observed in Chart 7, Google Translate and DeepL also delivered a similar performance for the nominal variant *metedura de pata*: 7 vs. 8 for the discontinuous form and 8 in both for the continuous one. That was not the case of the variant *metida de pata* (Chart 8), for which Google Translate's performance drastically declined both in the discontinuous (1 vs. 7) and the continuous form (3 vs. 8).
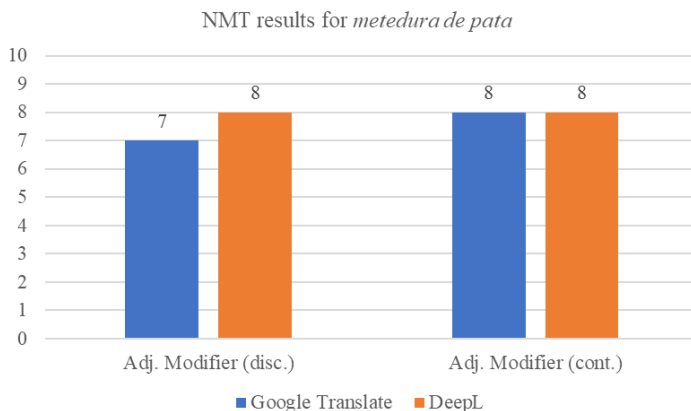
NMT results for *metedura de pata*
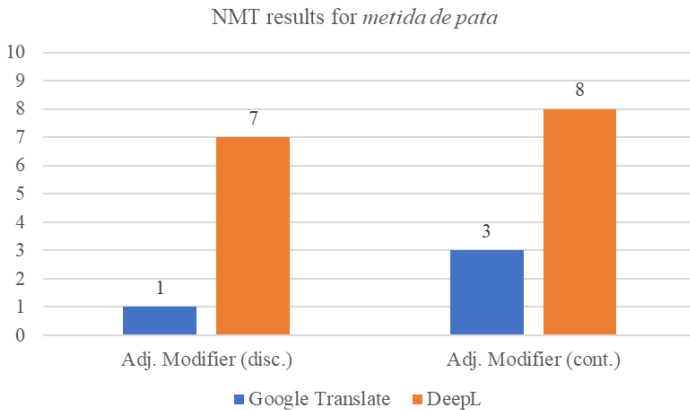
Chart 7: NMT results for *metedura de pata*

Chart 8: NMT results for *metida de pata*

Finally, concerning the lower-frequency variants *meter la pata hasta el corvejón, meter la gamba*, *metedura de gamba* and *metida de gamba*, neither Google Translate nor DeepL could adequately detect and translate these idioms, which led to a literal (and, hence, inappropriate) interpretation and translation of them in a total of 40 analysed cases.

Once the total of 380 cases has been classified and exposed, the global results will be displayed. As can be observed in Chart 9 and 10, overall, DeepL slightly outperformed Google Translate in the task of finding appropriate equivalents for the somatisms *tomar el pelo* and *meter la pata*, and their nominal variants *tomadura de pelo*, *tomada de pelo, metedura de pata,* and *metida de pata*. For the continuous form of these idioms, while Google Translate offered an accuracy rate of 86% (146 out of 170), DeepL achieved a score of 94% (160/170). In the discontinuous scenario, Google Translate was newly surpassed by DeepL: 71% (121/170) vs. 81% (138/170). The addition of both continuous and discontinuous results hence shows a minor contrast between both NMTs' performance: 79% (267/340) with Google Translate and 88% (298/340) with DeepL. Finally, when including all the remaining variants (*meter la pata hasta el corvejón*, *meter la gamba*, *metedura de gamba*, and *metida de gamba*) the global results exhibit an accuracy rate of 70% (267/380) for Google Translate and 78% (298/380) for DeepL.
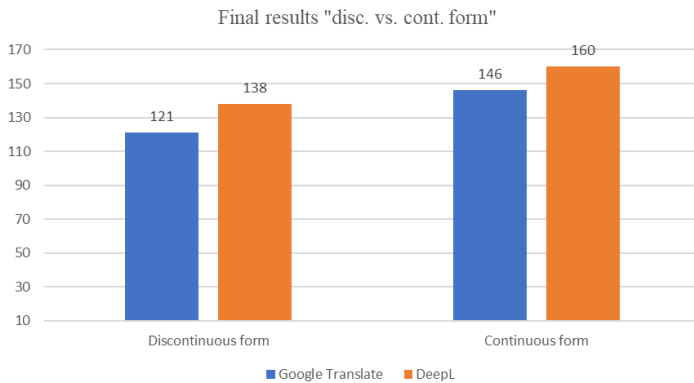
Final results "disc. vs. cont. form"

Chart 9: Final results "disc. vs. cont. form"
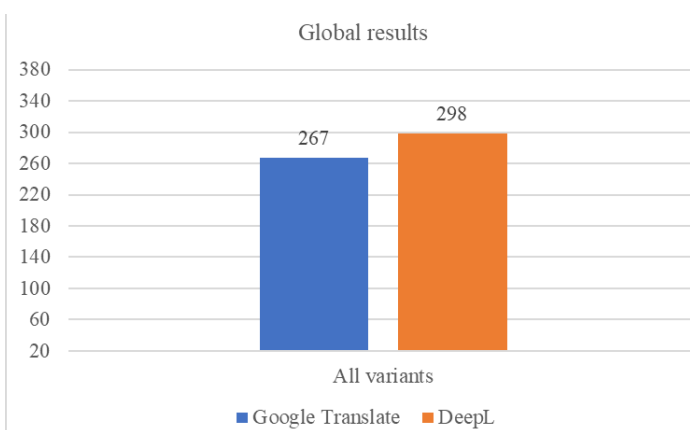
Global results

Chart 10: Global results

5. Analysis of the results

The overall results show that both Google Translate and DeepL delivered a high-quality performance in the detection and establishment of equivalences for the idioms *tomar el pelo* and *meter la* pata, and their respective nominal variants *tomadura de pelo,* and *metedura de pata*, both in their continuous and discontinuous forms, with a final global accuracy rate of 86% with Google Translate and 89% with Deepl for these somatisms.

Nevertheless, that was not the case of the nominal variants *tomada de pelo* and *metida de pata*, for which Google Translate's performance dramatically declined to 20% in the continuous and 30% in the discontinuous form of the sequence *tomada de pelo*, and to 30% in the continuous and 10% in the discontinuous form of *metida de pata*. In the case of *tomada de pelo*, despite presenting higher normalised frequencies in widely spoken varieties such as those of Argentina (0.12 vs. 0.4), Mexico (0.23 vs. 0.25), and Uruguay (0.8 vs. 0.85) and occurrences in most of the remaining national varieties, Google Translate's accuracy rate decreased 80% in the continuous and 60% in the discontinuous form when compared to the variant *tomadura de pelo*, for which European Spanish presents the highest comparative normalised frequency (1.43 vs. 0.01). Even more revealing is the case of *metida de pata* since, despite presenting a higher normalised frequency in most of the national varieties, it posed considerably more serious difficulties for Google Translate than *metedura de pata*, with a comparative decline of 50% in the continuous and 60% in the discontinuous form; even if *metedura de pata* is exclusively predominant in the dialectal areas of Cuba (0.26 vs. 0.09), Ecuador (0.18 vs. 0.14) and, mainly, Spain (0.84 vs. 0.09).

Therefore, the inclusion of the diatopically-marked variants *tomada de pelo* and *metida de pata* in the analysis resulted in a considerable decrease in the global NMTs' accuracy rate from 89% to 88% with Deepl and, more dramatically, from 86% to 79% with Google Translate. These results are in line with previous studies on phraseological variation (Corpas Pastor 2015 & 2018; Valencia Giraldo & Corpas Pastor 2019) which corroborate the overall tendency towards European Spanish to the detriment of American national varieties in the translation of phraseology. Furthermore, the results also highlight the pressing need for ameliorating NMT training data to encompass the diatopic variation of a given language as well as the necessity of searching alternative approaches for low-resourced language varieties (Honnet, Popescu-Belis et al. 2017; Costa-Jussà, Zampieri et al. 2018; Farhan, Talafha et al. 2020).

In addition to these obstacles, neither Google Translate nor DeepL could adequately detect and translate other lower-frequency variants such as *meter la pata hasta el corvejón* and *meter la gamba* (or the nominal variants *metedura de gamba* and *metida de gamba*). That is the reason why the final global results

(those containing all the variants) showed a further dramatic decline in the overall NMTs' performance from 79% to 70% for Google Translate, and from 88% to 78% for DeepL, all of which emphasises the additional necessity to optimise the NMTs' performance regarding rare and out-of-vocabulary words (Sennrich, Haddow et al. 2015; Gülçehre, Ahn et al. 2016; Liu, May et al. 2018)

It is also possible to detect in the results a decline in the NMTs' performance from the continuous to the discontinuous form of the idioms (from 86% to 71% in the case of Google Translate, and from 94% to 81% with DeepL). As can be observed in the itemised results, in the case of the verbal variants (*tomar el pelo* and *meter la pata*) only the insertion of common 1-gram modifiers such as *más* or the subject, and the 2-gram modifier *otra vez*, did not pose serious problems for any of the NMT systems. However, in the case of *tomar el pelo* the 1-gram modifier *tanto* presented difficulties for Google Translate, and so did the inclusion of several different n-grams for DeepL. Other modifiers such as *un poco* or the indirect object caused problems for both NMT systems. In the scenario of *meter la pata*, none of the modifiers posed an obstacle, except for the inclusion of several different n-grams, in this case only for Google Translate. As for the nominal variants (*tomadura de pelo* and *metedura de pata*), none of the adjective modifiers presented, overall, any serious difficulties for the NMT systems. Consequently, the itemised results indicate that the machine translation outcomes largely depend on the interplay of the idiom in question, the modifiers within the discontinuous form of the sequence, and the NMT system itself.

6. Conclusion

The present research has aimed to contribute to the growing body of literature (Rikters & Bojar 2017; Wang, Tu et al. 2017; Huang, Wang et al. 2018, inter alia) that is bringing into focus the treatment of MWEs by current neural machine translation systems as well as new avenues for its optimisation. In this context, the NMT systems Google Translate and DeepL have proved to deliver an overall high-quality performance in the detection and establishment of equivalences for idioms with high normalised frequencies such as *tomar el pelo* and *meter la pata*, and their nominal variants *tomadura de pelo*

and *metedura de pata*, both in their continuous and discontinuous forms, with a final average of 86% vs. 89%, respectively.

In spite of their high accuracy rate in the translation of these idiomatic sequences, both NMTs still encountered serious difficulties to adequately detect and translate other lower-frequency phraseological variants of the idioms (such as *meter la pata hasta el corvejón*, *meter la gamba*, and its nominal variants *metedura de gamba* and *metida de gamba*) or, in the case of Google Translate, other variants with higher frequency scores in widely spoken Latin-American national varieties, as was the case of the diatopically-marked variants *tomada de pelo* and *metida de pata*. These results hence concur with the need for further research to improve on NMTs' performance not only for rare and out-of-vocabulary words but also when facing the challenge of diatopic variation within a given language, which could not be properly seized by standard one-size-fits-all models.

References

Anastasiou, Dimitra. (2010) *Idiom Treatment Experiments in Machine Translation*. Newcastle upon Tyne: Cambridge Scholars Publishing. ISBN-13: 978-1-4438-2515-3 ISBN-10: 1-4438-2515-8

Anastasopoulos, Antonios. (2019) "An analysis of source-side grammatical errors in NMT." *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy: Association for Computational Linguistics, pp. 213–223.

Belinkov, Yonatan & Yonatan Bisk. (2018) "Synthetic and natural noise both break neural machine translation." *CoRR* abs/1711.02173

Corpas Pastor, Gloria. (2015) "Translating English Verbal Collocations into Spanish: on Distribution and other Relevant Differences related to Diatopic Variation." *Lingvisticae Investigationes Special Issue 'Spanish Phraseology. Varieties and variations'*, 38:2, pp. 229–262. ISSN: 03784169 / E-ISSN: 15699927. DOI 10.1075/li.38.2.03cor.

Corpas Pastor, Gloria. (2018) "Laughing one's head off in Spanish subtitles: a corpus-based study on diatopic variation and its consequences for translation." In: Mogorrón Huerta, Pedro & Albadalejo-Martínez, Antonio. (eds.) 2018. *Fraseología, Diatopía y Traducción, Colección "IVITRA Research in*

Linguistics and Literature". Ámsterdam: John Benjamins, pp. 54-106. ISBN: 9789027202253.

Costa-jussà, Marta R.; Marcos Zampieri & Santanu Pal. (2018) "A neural approach to language variety translation." *Proceedings of the fifth workshop on NLP for similar languages, varieties and dialects* (VarDial 2018), pp. 275–282.

Farhan, W.; Bashar Talafha; Analle Abuammar; Ruba Jaikat; Mahmoud Al-Ayyoub; Ahmad Bisher Tarakji & Anas Toma. (2020) "Unsupervised dialectal neural machine translation." *Information Processing & Management*, 57:3. ISSN 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2019.102181

Gülçehre, Çaglar; Sungjin Ahn; Ramesh Nallapati; Bowen Zhou & Yoshua Bengio. (2016) "Pointing the unknown words." *CoRR*. abs/1603.08148

Honnet, Pierre-Edouard; Andrei Popescu-Belis; Claudiu Musat & Michael Baeriswyl. (2017) "Machine translation of low-resource spoken dialects: Strategies for normalizing swiss german." *CoRR*. abs/1710.11035.

Huang, Po-Sen; Chong Wang; Sitao Huang; Denny Zhou, & Li Deng. (2018) "Towards neural phrase-based machine translation." *ICLR*.

Kilgarriff, Adam; Vít Baisa; Jan Bušta; Miloš Jakubíček; Vojtěch Kovvář; Jan Michelfeit; Pavel Rychlý & Vít Suchomel. (2003) "The Sketch Engine." <https://www.sketchengine.eu>

Koike, Kazumi. (2007) "Relaciones paradigmáticas y sintagmáticas de las locuciones verbales en español." In: Cuartero Otal, Juan & Emsel, Martina (eds.), *Vernetzungen Bedeutung in Wort, Satz und Text. Festschrift für Gerd Wotjak zum 65. Geburtstag*. Frankfurt am Main: Peter Lang, pp. 263-275.

Liu, Nelson F.; Jonathan May; Michael Pust & Kevin Knight. (2018) "Augmenting statistical machine translation with subword translation of out-of-vocabulary words." *arXiv* preprint arXiv:1808.05700.

Lohar, Pintu; Maja Popović; Haithem Alfi & Andy Way. (2019) "A systematic comparison between SMT and NMT on translating user-generated content." *20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*. La Rochelle, France.

Rikters, Matīss & Ondřej Bojar. (2017) "Paying attention to multi-word expressions in neural machine translation." *arXiv* preprint arXiv: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." *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pp. 2692 - 2698.

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.) 2002. *Computational Linguistics and Intelligent Text Processing. CICLing 2002. Lecture Notes in Computer Science.* Berlin, Heidelberg: Springer, pp. 1–15.

SENNRICH, Rico; Barry Haddow & Alexandra Birch. (2015) "Neural machine translation of rare words with subword units." *arXiv* preprint arXiv:1508.07909.

SPERBER, Matthias; Jan Niehues & Alex Waibel. (2017) "Toward robust neural machine translation for noisy input sequences." *International Workshop on Spoken Language Translation (IWSLT)*. Tokyo, Japan.

VALENCIA GIRALDO, M. Victoria, & Gloria Corpas Pastor. (2019) "The Portrait of Dorian Gray: A Corpus-Based Analysis of Translated Verb + Noun (Object) Collocations in Peninsular and Colombian Spanish." In: Gloria Corpas Pastor & Ruslan Mitkov (eds.) 2019. *Computational and Corpus-Based Phraseology. Europhras 2019, LNAI 11755.* Cham: Springer Nature Switzerland AG. DOI: https://doi.org/10.1007/978-3-030-30135-4_13

WANG, Xing; Zhaopeng Tu; Deyi Xiong & Min Zhang. (2017) "Translating phrases in neural machine translation." *EMNLP 2017*, pp. 1421–1431.

NOTA BIOGRÁFICA / BIONOTE

CARLOS MANUEL HIDALGO TERNERO es graduado en Traducción e Interpretación (ES<>EN/DE) por la Universidad de Málaga (2015), con un Máster Universitario en Enseñanza del Español como Lengua Extranjera y de Otras Lenguas Modernas, por la Universidad de Sevilla (2016) y un Máster en Profesorado de Educación Secundaria Obligatoria y Bachillerato, Formación Profesional y Enseñanza de Idiomas (especialidad lengua inglesa) por la UMA (2017). En la actualidad es Personal Investigador en Formación (PIF) en la UMA con un contrato FPU, ref. FPU16/02032, concedido por el Ministerio de Educación, Ciencia y Deporte. En este contexto, lleva a cabo su labor investigadora en varios proyectos tales como Interpreta 2.0 (PIE 17-015, type B), VIP (FFI2016-75831-P) y Termitur (HUM2754), así como ha sido miembro del Comité Organizador del Congreso Europhras 2019.

CARLOS MANUEL HIDALGO TERNERO holds a Bachelor's Degree in Translation and Interpreting (ES<>EN/DE), from the University of Malaga (2015), as well as a Master's Degree in Teaching Spanish as a Foreign Language and Other Modern Languages, from the University of Seville (2016), and a Master's Degree in Secondary Education and Language Teaching (English), from the University of Malaga (2017). He's currently a trainee researcher at the University of Malaga thanks to a FPU (Formación de Profesorado Universitario) contract, ref. FPU16/02032, awarded by the Ministerio de Educación, Ciencia y Deporte (Spanish Ministry of Education, Science and Sport). In this context, he carries out his research work in different projects such as Interpreta 2.0 (PIE 17-015, type B), VIP (FFI2016-75831-P) and TERMITUR (HUM2754), and he has also been a member of the Organising Committee for the international conference EUROPHRAS 2019.