

UNIVERSIDAD DE MÁLAGA
FACULTAD DE FILOSOFÍA Y LETRAS



TESIS DOCTORAL

**RECURSOS DOCUMENTALES Y TECNOLÓGICOS
PARA LA TRADUCCIÓN DE SOMATISMOS**

**TECHNOLOGICAL AND INFORMATION RESOURCES
FOR THE TRANSLATION OF SOMATISMS**

D. CARLOS MANUEL HIDALGO TERNERO

DIRIGIDA POR

DRA. D.^a GLORIA CORPAS PASTOR

TUTORIZADA POR

DR. D. DAVID MARÍN HERNÁNDEZ

Programa de Doctorado en Lingüística, Literatura y Traducción
Facultad de Filosofía y Letras
Universidad de Málaga

2022



UNIVERSIDAD
DE MÁLAGA

AUTOR: Carlos Manuel Hidalgo Ternero

 <https://orcid.org/0000-0002-8338-2627>

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga



Esta obra está bajo una licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional:

<http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

Cualquier parte de esta obra se puede reproducir sin autorización
pero con el reconocimiento y atribución de los autores.

No se puede hacer uso comercial de la obra y no se puede alterar, transformar o hacer obras derivadas.

Esta Tesis Doctoral está depositada en el Repositorio Institucional de la Universidad de Málaga (RIUMA): riuma.uma.es



ÍNDICE

CAPÍTULO I. SUMMARY	4
1.1 Practical motivation.....	9
1.2 Hypothesis and objectives.....	7
1.3 Methodology	8
1.4 Results	10
1.5 List of associated publications	12
1.6 Thesis structure	20
CAPÍTULO II. INTRODUCCIÓN	21
2.1 Motivación	21
2.2 Hipótesis y objetivos	26
2.3 Metodología	29
CAPÍTULO III. MARCO TEÓRICO: LOS SOMATISMOS Y SU TRADUCCIÓN	37
3.1 Sobre el concepto de <i>fraseología</i> y <i>unidad fraseológica</i>	39
3.2 Sobre el concepto de <i>somatismo</i> y su lugar en el universo fraseológico	43
3.3 Sobre la traducción de las unidades fraseológicas	47
3.3.1 La equivalencia translémica	49
3.3.2 La equivalencia textual	52
3.4 Características principales de las unidades fraseológicas	55
3.4.3 Fijación	55
3.4.4 Idiomatidad	58
3.4.5 La variabilidad en la fraseología	59
CAPÍTULO IV. RECURSOS Y HERRAMIENTAS ELECTRÓNICOS APLICADOS A LA TRADUCCIÓN DE SOMATISMOS SEGÚN LOS DISTINTOS RETOS	67
4.1 Variación	69
4.1.1 Traducción Automática Neuronal	69
4.1.2 Sistemas de preprocesamiento de unidades fraseológicas para TAN.....	74
4.1.3 Corpus lingüísticos	87
4.2 Manipulación fraseológica	93
4.2.1 Word Sketch	93
4.2.2 Glosarios	96

4.2.3	Análisis comparativo de distintos recursos y herramientas orientadas a la traducción de la manipulación fraseológica	108
CAPÍTULO V. DISCUSSION, CONCLUSION AND FUTURE LINES OF RESEARCH		
	111	
5.1	Summary of contributions.....	111
5.2	Future lines of research	118
BIBLIOGRAFÍA.....		120

AGRADECIMIENTOS

Vaya mi primera mención a la Dra. Gloria Corpas Pastor, por ser mi faro académico, una fuente inagotable de inspiración y todo un ejemplo de vocación docente e investigadora. Su magnífica labor como directora (o más bien como «madre académica»), sus constantes ánimos y sus valiosos consejos me han permitido crecer tanto personal como profesionalmente a lo largo de estos años de doctorado, así como han hecho posible mi participación en una gran variedad de proyectos, estancias y publicaciones gracias a los cuales ha visto la luz el presente proyecto de tesis.

Asimismo, me gustaría expresar mi gratitud a los doctores Natalie Kübler, Ruslan Mitkov y Johanna Monti por acogerme en el seno de sus equipos durante mis tres estancias de investigación y permitirme enriquecerme de primera mano con los interesantes proyectos, seminarios y conferencias a los que tuve la suerte de asistir y que supusieron un impulso decisivo en mi formación. También me gustaría agradecer a todos los compañeros y amigos que he conocido en este camino predoctoral y que han dejado en mí una huella imborrable, tanto aquí en España, como en Francia, en Reino Unido y en Italia.

Finalmente, a mis más queridos, familiares y amigos, por su inestimable apoyo en este arduo (pero gratificante) proceso. Gracias a mis tíos, Lola y Paco, por ofrecerme un espacio de trabajo en aquellos momentos en los que necesitaba un cambio de aires. Gracias a mi hermana, María Jesús, y a mis sobrinas, Silvia y Marta, por alumbrar mis días con su cariño y amor incondicional. Y, por supuesto, a mis dos pilares, mis padres, por estar siempre ahí, por desvivirse por mí y por animarme y apoyarme a alcanzar mi sueño de ser profesor universitario.

CAPÍTULO I. SUMMARY

1.1 Practical motivation

There are many reasons that have led us to undertake this PhD dissertation¹, whose main objective is the analysis and design of electronic tools and resources for the translation of somatisms, i.e., phraseological units (hereinafter, PUs) containing lexemes referring to human or animal body parts. On the one hand, phraseological units are ubiquitous in any type of discourse, register or textual genre. More specifically, it has been observed that over 50% of any text consists of PUs in the broad sense (cfr. Erman & Warren, 2000; Jackendoff, 1995; Mel'čuk, 1995; Sinclair, 1991).

In this context, phraseology plays a key role in the translation process. Therefore, one of the main motivations of the present research project is to enhance students' translation competence regarding phraseology, which inexorably entails learning heuristic and instrumental strategies in order to exploit all the possibilities offered by different tools such as corpora, machine translation systems and glossaries, inter alia, when establishing textual equivalences for somatisms in the TT.

The importance of teaching and promoting the use of these tools among students is reflected in recent surveys on the use of technologies by translators (Zaretskaya et al., 2015 & 2018), where it is reported that a very small number of the translators participating in the survey work with corpora (17%). In this study it was also possible to observe that only 36% of the interviewed translators currently use machine translation, although a majority (74%) consider that they could employ these systems more if they were of high quality. In this regard, the latest European Language Industry Surveys show that machine translation is already present in more than half of the European Language Service Companies (ESCs) participating in the study (ELIS,

¹ This dissertation has been carried out within the framework of different research projects in technologies applied to translation and interpreting (ref. FFI2016-75831-P, UMA18-FEDERJA-067, CEI-RIS3 and EUIN2017-87746). Furthermore, it has been funded by the Ministerio de Ciencia, Innovación y Universidades [Spanish Ministry of Science, Innovation and Universities] (FPU16/02032).

2018) and that 78% of them plan to start using or increase the use of MT or post-editing (ELIS, 2020). In this line, all sectors concerned (training institutes, buyers and language developers as well as independent professionals) also recognise MT and post-editing by far as the most prominent trend in the industry (ELIS, 2021). Against such a background, these and other essential tools in the translation task will be analysed along the present dissertation.

1.2 Hypothesis and objectives

In this context, we present the main hypothesis of this dissertation: *that it is possible to analyse and design electronic tools and resources for the translation of somatisms considering two of the main challenges they present as phraseological units: variation and manipulation*. From this initial hypothesis, the following objectives are pursued throughout this project.

The first objective (Paper 1) was to analyse the performance of current neural machine translation systems in the translation of somatisms, paying special attention to the main challenges presented by the variation and discontinuity of these units.

As a result of the difficulties detected, the second objective (Papers 2-7), the most ambitious one, was to design a text pre-processing system (*gApp*) for the automatic conversion of somatisms from their discontinuous into their continuous forms, in order to improve their automatic detection and translation.

The third objective (Papers 8, 9 and 10) sought to analyse the performance of corpora as information tools for the translation of somatisms under the challenge of idiom ambiguity (Paper 8), variation (Paper 9), and discontinuity (Paper 10).

The fourth objective (Paper 11) aimed at developing a corpus-based glossary of somatisms, called Glossomatic, for the creation of ad hoc phraseological equivalents in those cases where the

idiom manipulation in the source text along with the absence of one-to-one crosslinguistic phraseological correspondences could pose a challenge in the translation process. In this context, a teaching proposal was designed to provide trainee translators with different strategies and steps that could be implemented with the aid of the glossary to deal with these cases.

Finally, the fifth objective (Paper 12) was to contrast different electronic tools and resources (linguistic corpora, lexicographic resources, the web, NMT systems, among others) for the creation of textual equivalences in those cases with phraseological manipulation in the source text. To this end, in a case study carried out in the classroom, students' translation proposals were analysed in order to determine which resource and tool were most useful depending on the type of manipulation in the ST.

1.3 Methodology

In order to achieve the five objectives previously presented, all the publications in this PhD thesis follow a corpus-based methodology, either for the analysis and extraction of somatisms in their continuous and discontinuous forms (Papers 1-7); for the study of the possibilities offered by corpora in translation under the challenge of idiom ambiguity (Paper 8), variation (Paper 9) or manipulation (Papers 10, 11 and 12), or for the design of the glossary of somatisms Glossomatic (Paper 11).

In this way, for the first and second objectives (Papers 1-7), the PUs under analysis were extracted using Sketch Engine's Corpus Query Language (CQL) from esTenTen18 and Timestamped JSI web corpus 2014-2021 Spanish for Spanish, frTenTen17 and Timestamped JSI web corpus 2014-2021 French for French, and itTenTen16 and Timestamped JSI web corpus 2014-2021 Italian for Italian. After this, for both objectives, the performance of NMT systems

against the continuous and discontinuous form of the examined PUs was measured following an instance-based MT evaluation (Zaninello & Birch, 2020), with several possible candidates in the target text for somatisms in both their continuous and discontinuous forms. A corpus analysis was also carried out to examine the *relevant results* (those containing real instances of the discontinuous forms of the somatisms under study) and the *irrelevant results* (other concordances containing analogous patterns but unrelated to the idiomatic sequences), which allowed us to determine the necessary restrictions of *gApp*'s detection and conversion mechanism.

To achieve the third objective (Papers 8, 9 and 10), the search interface of different corpus management systems was analysed in order to explore the possibilities offered by each of them in querying somatisms under the challenges presented by these units: ambiguity (Paper 8), variation (Paper 9), and manipulation (Paper 10). In this context, both the simple concordance system of the corpora under study and additional functionalities were examined, such as the option *Coapariciones* of CORPES XXI and the annotated version of CREA (Papers 8 and 9), or, within Sketch Engine, the CQL code (Paper 9) and Word Sketch (Paper 10).

With regard to the fourth objective (Paper 11), the glossary Glossomatic was implemented in a relational database in Microsoft Access 2016, developed by Microsoft and belonging to the Microsoft Office suite of applications. In this regard, both parallel corpora, such as OPUS2 and Europarl77, and monolingual corpora in Spanish (CORPES XXI and esEuTenTen), English (BNC and enTenTen20), and German (deTenTen1813 and DWDS-Kernkorpus [1900-1999]) were analysed for the extraction of the somatisms constituting Glossomatic. The analysis of the contexts of real use of somatisms in the corpora also allowed the design of the different query tabs of Glossomatic (*collocations*, *polarity*, *register*, etc.) as well as useful contrastive information on translation equivalents in the Spanish-English-German language combination.

Finally, to achieve objective 5 (Paper 12), we analysed the performance of different tools, such as parallel and monolingual corpora, different lexicographic resources, Neural Machine Translation systems, search engines (Google, Bing, Metasearch, etc.) and rhymers, among others, in the translation of six cases with manipulated idioms in the ST: 3 of them belonging to a heterogeneous typology of PU and 3 of them being somatisms. The different translation proposals of the 45 students participating in the study were collected and analysed in the form creation tool Googleform, which allows an easy visualisation and processing of the stored results.

1.4 Results

The results in Papers 1-7 permitted to observe that idiom variation and discontinuity pose serious challenges for NMT systems. In this regard, in Paper 1 it was possible to determine that the two most robust NMT systems, DeepL and Google Translate, faced enormous difficulties for the translation of low-frequency somatisms, and other diatopically-marked variants which present higher frequency scores in widely spoken Latin-American national varieties. Regarding discontinuity, it meant a decline of DeepL's performance by 13% and of Google Translate's by 15%, when compared to the continuous form of these somatisms (Paper 1). In the light of these results, the text preprocessing system *gApp* was designed for the automatic detection and conversion of discontinuous somatisms into their continuous form. This system proved to provide an average improvement for DeepL by 10% and for Google Translate by 14.6% in the ES>EN directionality (Paper 2). In Papers 3 and 4, only DeepL's performance was analysed, for which *gApp* represented an improvement by 13.5% for the ES>EN, and by 9% for the ES>DE directionalities (Paper 3), as well as by 18% in the FR>EN and by 16.5% in the FR>ES directionalities (Paper 4). In Papers 5-7, new NMT systems were examined: ModernMT, in the

case of Paper 5 and 7, and VIP, in Paper 6. In these three Papers a new category of MWE was also studied: Verb + Prepositional Phrase (VPP) constructions. Against such a background, *gApp* proved to still make a decisive advance in NMT quality. In Paper 5, this system could improve ModernMT's performance by 7.5% and DeepL and Google Translate's by 6.5% in the ES>EN/DE/FR/IT/PT directionalities. In Paper 6, *gApp* could improve VIP's performance by 21.5%, DeepL's by 13.5% and Google Translate's by 8.5% in the ES>EN directionality. Finally, in Paper 7, we could observe how *gApp* was able to ameliorate ModernMT, DeepL and Google Translate's performance, on average, by 11.5% in the IT>EN and by 8.3% in the IT>DE directionality.

With regard to corpora, we examined the possibilities they offer in the translation of somatisms under the challenges posed by idiom ambiguity (Paper 8), variation (Paper 9) and manipulation (Paper 10). In Papers 8 and 9, we contrasted the performance of different corpora (CORPES XXI, Linguee, and Sketch Engine, in Paper 8, and CREA, in its traditional and annotated version, CORPES XXI, and Sketch Engine, in Paper 9). The results from both studies permitted to observe that Sketch Engine was the system offering a better performance both in quantitative and qualitative terms. In Paper 10, we analysed the possibilities of Sketch Engine's functionality, Word Sketch, in the creation of ad hoc phraseological equivalences for manipulated idioms in the ST, with very promising results: along the 4 translation scenarios, most of the trainee translators participating in the study were able to provide not only a manipulated idiom but more specifically a manipulated somatism for the TT (97% in translation scenario 1, 74% in scenario 2, 51% in scenario 3, and 71% in scenario 4).

Further resources for the translation of manipulated somatisms are presented in Paper 11, where the trilingual corpus-based glossary Glossomatic is described, and in Paper 12, in which we

contrasted the performance of different e-tools (language corpora, lexicographic resources, or the web, inter alia) for the creation of textual equivalences for manipulated idioms in the ST. In the light of these results, along the 6 translation scenarios presented, it was possible to observe a global predominant use of lexicographic resources (34.6%) and textual corpora (32.4%). The remaining tools were mainly employed by trainee translators in two specific translation scenarios: while in case 1, 79.3% of trainee translators consulted search engines to find toponyms and ethnonyms through which they could recreate an analogous manipulation to the one in the ST; in case 2, 25% employed rhymers to reproduce a specific phonostylistic effect in the TT.

1.5 List of associated publications

The list of associated publications is here presented in an order corresponding to the main objectives of the project, i.e., it is not necessarily a chronological order. In this way, once the performance of the different NMT systems in cases of idiom variation and discontinuity is analysed in Paper 1, Papers 2-7 present the text preprocessing system *gApp* for the automatic detection and conversion of discontinuous somatisms. Subsequently we will examine the performance of different corpus management systems under the challenge of phraseological ambiguity (Paper 8), variation (Paper 9), and manipulation (Paper 10). Further tools for the translation of manipulated somatisms are introduced in Papers 11 and 12.

Paper 1. Hidalgo-Ternero, C. M. (2020a). Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation. In P. Mogorrón Huerta (Ed.), *Análisis multidisciplinar del fenómeno de la variación en traducción e interpretación / Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting. MonTI Special Issue 6*, 154-177. <https://doi.org/10.6035/MonTI.2020.ne6.5> (Indexed and evaluated in: Web of Science: JCR (Q4 in Language & Linguistics in Emerging Sources Citation Index), Scopus (Q2 in Language & Linguistics), SJR SCImago Journal & Country Rank (Q2 in Language &

Linguistics), MLA - Modern Language Association Database, DIALNET, CARHUS Plus+ 2018, LATINDEX [32 criterios], Sello de calidad FECYT, ERIHPlus, MIAR [7.5/11])

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

Paper 2. Hidalgo-Ternero, C. M., & Corpas Pastor, G. (2020b). Bridging the ‘gApp’: improving neural machine translation systems for multiword expression detection. *Yearbook of Phraseology*, 11(1), 61–80. <https://doi.org/10.1515/phras-2020-0005> (Indexed and evaluated in: Web of Science: JCR (Q2 in Language & Linguistics in Emerging Sources Citation Index), Scopus (Q2 in Language & Linguistics), SJR SCImago Journal & Country Rank, IBZ Online, Linguistic Bibliography, Communication & Mass Media Index, MLA - Modern Language Association Database, ERIHPlus, MIAR [9.5/11])

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, concluding remarks on how to further optimise MWE-aware NMT systems will round off the chapter.

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

Paper 3. Hidalgo-Ternero C. M., & Corpas Pastor, G. (2022a/forthcoming). *ReGap: a text preprocessing algorithm to enhance MWE-aware neural machine translation systems*. In

J. Monti, G. Corpas Pastor & R. Mitkov (Eds.), *Recent Advances in MWU in Machine Translation and Translation technology*. John Benjamins Publishing Company. (in SPI 11th position in the general ranking and 5th position in Linguistics [international ranking])

Abstract

This research presents ReGap, a text preprocessing algorithm for the automatic token-based identification and conversion of discontinuous multiword expressions (MWEs) into their continuous form as a means to optimise neural machine translation (NMT) systems. To this purpose, an experiment with semi-fixed Verb-Noun Constructions (VNCs) is conducted so as to assess to what extent ReGap can enhance the performance of the most robust NMT system to date, DeepL, under the challenge of MWE discontinuity in the Spanish-into-English and the Spanish-into-German directionalities. In this regard, the promising results yielded for VNCs will shed some light on new avenues for enhancing MWE-aware NMT systems.

Keywords: text preprocessing algorithm, Neural Machine Translation (NMT), DeepL, token-based MWE identification, Verb-Noun Constructions (VNC), discontinuity

Paper 4. Hidalgo-Ternero C. M. (2021). El algoritmo ReGap para la mejora de la traducción automática neuronal de expresiones pluriverbales discontinuas (FR>EN/ES). In G. Corpas Pastor, M. R. Bautista Zambrana & C. M. Hidalgo-Ternero (Eds.), *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus* (pp. 253-270). Comares. (in SPI 11th position in the general ranking, and 14th position in Linguistics [Spanish ranking])

Abstract

In this paper we introduce the text-preprocessing algorithm ReGap designed for automatically detecting and converting discontinuous multiword expressions into their continuous forms so as to improve the performance of current neural machine translation systems (NMT). In order to test its effectiveness, an experiment with the most robust NMT system to date, DeepL, has been carried out for the translation of verbo-nominal somatisms in the FR>EN and FR>ES directionalities. The promising results yielded by the study for this typology of multiword expressions will open up new ways to optimise MWE-processing by NMT systems.

Keywords: text preprocessing algorithm, Neural Machine Translation (NMT), DeepL, verbo-nominal somatisms, discontinuity

Paper 5. Hidalgo-Ternero C. M., & Corpas Pastor, G. (2022b/forthcoming). Qué se traerá *gApp* entre manos... O cómo mejorar la traducción automática neuronal de variantes somáticas (ES>EN/DE/FR/IT/PT). In M. Seghiri & M. Pérez Carrasco (Eds.), *Aproximación a la traducción especializada*. Peter Lang. (in SPI 5th position in the general ranking and 2nd position in Linguistics [international ranking])

Abstract

In this paper we present research results with *gApp*, a text-preprocessing system designed for automatically detecting and converting discontinuous multiword expressions (MWEs) into their continuous forms so as to improve the performance of current neural machine translation systems (NMT) (see Hidalgo-Ternero, 2021, and Hidalgo-Ternero & Corpas Pastor, 2020 and 2021). In order to test its effectiveness, an experiment with some of the most robust NMT systems to date (ModernMT, DeepL and Google Translate) has been carried out in the ES>EN/DE/FR/IT/PT directionalities for the translation of somatisms, i. e. MWEs containing lexemes referring to human or animal body parts (Mellado Blanco, 2004). More specifically, we have analysed Verb + Prepositional Phrase (VPP) constructions: the lexical variants *traer(se) entre manos*, *tener entre manos* y *llevar(se) entre manos*. The promising results yielded by the study for this typology of multiword expressions open up new ways to optimise MWE-processing by NMT systems. This is, to the best of our knowledge, the first study analysing the performance of this system for the neural machine translation of VPP multiword expressions in five different language directionalities and with three different NMT systems.

Keywords: text preprocessing system, Neural Machine Translation (NMT), Verb + Prepositional Phrase constructions, somatic variants, discontinuity

Paper 6. Hidalgo-Ternero C. M., & Corpas Pastor, G. (2022c/forthcoming). A la cabeza de la traducción automática neuronal asistida por *gApp*: somatismos en VIP, DeepL y Google Translate. In G. Corpas Pastor y M. Seghiri (Eds.), *Aplicaciones didácticas de las tecnologías de la interpretación*. Comares. (in SPI 11th position in the general ranking, and 14th position in Linguistics [Spanish ranking])

Abstract

In this paper we present research results with *gApp*, a text-preprocessing system designed for automatically detecting and converting discontinuous multiword expressions (MWEs) into their continuous forms so as to improve the performance of current neural machine translation systems (NMT) (see Hidalgo-Ternero, 2021, and Hidalgo-Ternero & Corpas Pastor, 2020, 2022a and 2022b). In order to test its effectiveness, an experiment with the NMT systems of VIP, DeepL and Google Translate has been carried out in the ES>EN directionality for the translation of somatisms, i. e., MWEs containing lexemes referring to human or animal body parts (Mellado Blanco, 2004). More specifically, we have analysed Verb + Prepositional Phrase (VPP) constructions, such as *agradecer de corazón*, *ver con buenos ojos*, *ir de cabeza* and *ir en cabeza*. In this regard, some of the unexpected results yielded by the study of these multiword expressions will question the widely accepted conception of phraseological discontinuity as an unequivocal synonym of worse NMT performance.

Keywords: text preprocessing system, Neural Machine Translation (NMT), VPP constructions, somatisms, phraseological discontinuity

Paper 7. Hidalgo-Ternero C. M., Lista, F. & Corpas Pastor, G. (2022/under review). *gApp-assisted NMT: how to improve the neural machine translation of discontinuous multiword expressions (IT>EN/DE)*. Language Resources and Evaluation (Indexed and evaluated in: Web of Science: Science Citation Index Expanded, Scopus (Q1 in Language & Linguistics), SJR SCImago Journal & Country Rank (Q1 in Language & Linguistics), Academic Search Premier, FRANCIS, Aerospace Database, Applied Science & Technology Source, Civil Engineering Abstracts, Computer & Applied Sciences, Humanities Abstracts, INSPEC, Linguistic Bibliography, Metadex, Communication Abstracts, EBSCO Education Source, Educational research abstracts (ERA), Linguistics & Language Behavior Abstracts, MLA - Modern Language Association Database, MIAR [11/11])

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 + Prepositional Phrase (VPP) Constructions will be carried out in order to evaluate to what extent *gApp* can optimise the performance of three free open-source NMT systems —ModernMT, Google Translate and DeepL— under the challenge of MWE discontinuity in the Italian into English and Italian into German directionalities. In the light of our promising results, concluding remarks on how to further optimise *gApp*-assisted NMT will round off the chapter.

Keywords: text preprocessing system, Neural Machine Translation (NMT), Multiword Expression (MWE), Verb + Prepositional Phrase (VPP) constructions, discontinuity

Paper 8. Hidalgo Ternero, C. M., & Corpas Pastor, G. (2020c). Estrategias heurísticas con corpus para la enseñanza de la fraseología orientada a la traducción. In M. Seghiri (Ed.), *La lingüística de corpus aplicada al desarrollo de la competencia tecnológica en los estudios de traducción e interpretación y la enseñanza de segundas lenguas* (pp. 181-204). Peter Lang. <https://doi.org/10.3726/b14734> (in SPI 5th position in the general ranking and 2nd position in Linguistics [international ranking])

Abstract

This work presents a didactic proposal carried out in the subject *Lengua y cultura “B” aplicadas a la Traducción e Interpretación (II) – inglés*, taught in the first year of the Bachelor’s Degree in Translation and Interpreting, at the University of Malaga. The main objective of this proposal is to teach the possibilities that both monolingual and bilingual corpora can provide for the correct identification and interpretation of phraseological units with regard to their translation, paying special attention to those cases where the ambiguity of phraseological sequences may lead to multiple readings. We will focus on *somatisms* and will mainly use two Spanish monolingual corpora (CORPES XXI and esEuTenTen), an English monolingual corpus (enTenTen20) and two parallel corpora (Europarl7 and Linguee, more specifically its English-Spanish subcorpus). Against this background, this proposal is divided into several learning activities. After a first seminar where the concepts of *corpus*, *phraseology* and *translation* are introduced, in the learning activity 2 we will

use parallel corpora to find translation pairings that contain translation mistakes caused by problems with phraseological ambiguity. Then, in the third learning activity, we will teach some disambiguating elements that will facilitate a correct identification and interpretation of the phraseological unit, in order to be able to convey its pragmatic and semantic weight in the target text. It is in this step where corpora can play a decisive role as documentation tools. Nevertheless, the localisation and interpretation of phraseological units is not problem-free. Given the necessity to develop some techniques that will enable a more effective detection of phraseological units, in the fourth learning activity students will learn an array of heuristic strategies to refine their searches in the consulted corpora as well as to select adequate equivalences after a correct interpretation of the results produced by these corpora.

Keywords: didactic proposal, phraseological unit, corpus linguistic, equivalence and somatism

Paper 9. Hidalgo-Ternero, C. M., & Corpas Pastor, G. (2021a). La variación fraseológica: análisis del rendimiento de los corpus monolingües como recursos de traducción. *Études romanes de Brno*, 42(1), 359-379. <https://doi.org/10.5817/ERB2021-1-19> (Indexed and evaluated in: Web of Science: JCR [Q3 in Language & Linguistics in Emerging Sources Citation Index], Scopus, SJR SCImago Journal & Country Rank [Q3 in Literature and Literary Theory], IBZ Online, Linguistic Bibliography, Linguistics & Language Behavior Abstracts, MLA - Modern Language Association Database, DOAJ, DIALNET, Directory of Open Access Journals, ERIHPlus, MIAR [10/11])

Abstract

Idioms tend to vary significantly in discourse (variation, grammatical inflection, discontinuity...). This makes it especially difficult to create appropriate query patterns that obtain these units in all shapes and forms while avoiding excessive noise. In this context, this paper analyses the performance of different corpus management systems available for Spanish when searching phraseological variants such as *tener entre manos*, *traer entre manos* and *llevar entre manos*, as well as *ir al pelo* and *venir al pelo*. More specifically, we will examine two corpora created by the Real Academia Española (CREA, in its original and annotated version, and CORPES XXI), the Corpus del Español, and Sketch Engine. The results of our study will shed some light on which corpus management system can offer a better performance for translators under the challenge of idiom variation.

Keywords: idiom variation, discontinuity, query patterns, corpus management systems

Paper 10. Hidalgo-Ternero, C. M. (2019a). Translating manipulated idioms (EN>ES) in the Word Sketch scenario. In G. Corpas Pastor & R. Mitkov (Eds.), *Computational and Corpus-Based Phraseology* (pp. 173-186). Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-30135-4_13 (in SPI 4th position in the general ranking and 21th position in Linguistics)

Abstract

This paper presents a teaching proposal for translation/interpreting undergraduate students. Following a corpus-based methodology, the main objective of this proposal is to train students on how to exploit the possibilities offered by Sketch Engine for the creation of *ad hoc* phraseological equivalences. More precisely, we will examine the potential of its functionality Word Sketch as a convenient tool in those cases where the manipulation of idioms and the absence of one-to-one phraseological correspondences may pose problems to translation. The experimental setting comprised an introductory seminar on the convergence of corpora, phraseology and translation, followed by a hands-on session, where trainee translators were presented with some case-studies for which they had to create *ad hoc* phraseological equivalents for the manipulated idioms in the source text (ST) and had to justify both the translation *process* and *product*. Overall, the insights gained from analysing the results obtained will allow us to determine to what extent trainee translators are able to translate manipulated idioms in the Word Sketch scenario.

Keywords: Teaching proposal, translation, Word Sketch, idiom manipulation, *ad hoc* phraseological equivalences

- Paper 11.** Corpas Pastor, G., Hidalgo Ternero, C. M., & Bautista Zambrana, M. R. (2020). Teaching idioms for translation purposes: a trilingual corpus-based glossary applied to Phraseodidactics (ES/EN/DE). In F. Mena Martínez & C. Strohschen (Eds.), *Teaching Phraseology in the XXI Century: New Challenges* (pp. 75-93). Peter Lang. (in SPI 5th position in the general ranking and 2nd position in Linguistics [international ranking])

Abstract

Phraseology plays a pivotal role in the development of translation competence as well as in translation quality assessment. Thus far, however, there remains a paucity of research on how to best teach idioms for translation purposes. Against such a background, this study aims to shed some light on the multiple applications of phraseodidactics to translation training. We will follow a corpus-based methodology and, for the sake of the argument, the focus will be on somatisms in Spanish, English and German. The overall structure of this paper takes the form of four sections. Section one begins by laying out the theoretical dimensions of phraseology and its convergence with translation. In section two we examine the main components of a corpus-based glossary of somatisms, named Glossomatic, and how it can be employed to establish *ad hoc* phraseological equivalences in those cases (analysed in section three) where the manipulation of idioms and the absence of one-to-one phraseological correspondences may pose some problems to translation. In this regard, given the importance of accurately conveying the pragmatic, semantic and discursive load of an idiom into a TT and, concomitantly, conveying the manipulation depicted in the ST, section four presents a teaching proposal in which students are prompted with a set of strategies and steps to be implemented with the aid of the glossary in order to solve these issues. Overall, the insights gained from this research will prove useful not only in developing trainees' phraseological competence but also in giving centre stage to phraseodidactics in Translation Studies.

Keywords: phraseodidactics, *ad hoc* phraseological equivalences, idiom manipulation, corpus linguistics, glossary

Paper 12. Hidalgo Ternero C. M., & Corpas Pastor, G. (2021a). Herramientas y recursos electrónicos para la traducción de la manipulación fraseológica: un estudio de caso centrado en el estudiante. *CLINA, An Interdisciplinary Journal of Translation, Interpreting and Intercultural Communication*, 71-94. <https://doi.org/10.14201/clina2020627194> (Indexed and evaluated in: Index Islamicus, DIALNET, LATINDEX. Catálogo v2.0 [2018-], ERIHPlus, MIAR [3.8/11])

Abstract

This article presents a case study carried out with students of the subject Traducción General «BA-AB» (II) - Inglés-Español / Español-Inglés, taught in the first year of the Bachelor's Degree in Translation and Interpreting, at the University of Malaga. In this regard, at a first stage, students were trained on how to exploit the possibilities offered by different e-tools and resources (language corpora, lexicographic resources, or the web, inter alia) for the creation of textual equivalences in those cases where the manipulation of idioms and the absence of one-to-one phraseological correspondences may pose problems to translation. To this end, an introductory seminar on phraseological manipulation and translation was followed by a hands-on session, where trainee translators were presented with some scenarios including manipulated idioms in the source text (ST). Overall, the insights gained from analysing the results obtained will allow us to determine to what extent the different tools can help students walk the tightrope of translating phraseological manipulation.

Keywords: case study, phraseological manipulation, e-tools for translation, English-Spanish

In this context, a summary of the scientific production over my PhD studies can be observed in Table 1 (with the publisher in bold).

	Journal articles	Book chapters	Books	Reviews	Conferences
1	Hidalgo-Ternero (2020a). MONTI	Hidalgo-Ternero (2019a). Springer	Corpas Pastor, Bautista Zambrana & Hidalgo-Ternero (Eds.) (2021). Comares	Hidalgo-Ternero (2018). Hikma: Revista de Traducción	Hidalgo-Ternero (25-27 September 2019a). Europhras 2019
2	Hidalgo-Ternero & Corpas Pastor (2020b). Yearbook of Phraseology	Corpas Pastor, Hidalgo-Ternero & Bautista Zambrana (2020). Peter Lang		Hidalgo-Ternero (2019b). TRANS: Revista de Traductología	Hidalgo-Ternero (25-27 September 2019b).

					Europhras 2019
3	Hidalgo-Ternero & Corpas Pastor (2020c). CLINA	Hidalgo-Ternero & Corpas Pastor (2020a). Peter Lang		Hidalgo-Ternero (2020b). Yearbook of Phraseology	
4	Hidalgo-Ternero & Corpas Pastor (2021a). Études romanes de Brno	Hidalgo-Ternero (2021). Comares			
5	Hidalgo-Ternero, Lista & Corpas Pastor (2022/under review). Language Resources and Evaluation	Hidalgo-Ternero & Corpas Pastor (2022a). John Benjamins Publishing Company			
6		Hidalgo-Ternero & Corpas Pastor (2022b). Peter Lang			
7		Hidalgo-Ternero & Corpas Pastor (2022c). Comares			

Table 1. Summary of the scientific production over my PhD studies

1.6 Thesis structure

After this summary has been presented, the remainder of the dissertation is structured as follows.

In Chapter II, we will further elaborate on the motivation, hypothesis, objectives, and methodology of the present PhD thesis. In Chapter III, we will delve into its theoretical framework, by exploring different key notions such as *phraseology* and *phraseological unit*, *somatism*, *the translation of PUs*, and two of the main characteristics of PUs, i.e., *fixedness* and *idiomaticity*, from where the notions of *variation* and *manipulation* emanate. Subsequently, in Chapter V, we will present the different electronic tools and resources applied to the translation of somatisms under two of the main challenges presented by these units: variation and manipulation.

CAPÍTULO II. INTRODUCCIÓN

2.1 Motivación

Numerosas son las razones que nos han llevado a emprender el presente proyecto de tesis doctoral, cuyo objetivo principal consiste en el análisis y diseño de recursos documentales y tecnológicos para la traducción de somatismos. Por un lado, las unidades fraseológicas (en adelante, UF) en sí constituyen una parte esencial de las lenguas y, por ende, del proceso traslaticio. Lejos de ser excepcionales, las UF se presentan de forma ubicua en todo tipo de discurso, registro o género textual. Según el «principio idiomático» (*idiom principle*) de Sinclair (1991) la fraseología, en su sentido amplio, conforma en torno al 50 % de cualquier texto; así lo señalan también Jackendoff (1995) y, en un porcentaje aún mayor, Erman y Warren (2000) y Mel'čuk (1995). En este sentido, tal es la importancia de la fraseología que Ladmiral (2015) habla de una «fraseologización generalizada de la lengua».

Así, la fraseología desempeña un papel fundamental en el desarrollo de la competencia traductora y en la evaluación de la calidad de las traducciones. En general, los principales retos que aún hoy continúan presentándose en el proceso traslaticio son la identificación de la UF en el texto de origen y su correcta comprensión en contexto, la selección del equivalente o estrategia de traducción adecuados, y la correcta reproducción de la carga semántico-pragmática de la unidad en el texto meta. El fallo en alguno de estos pasos conlleva irremediablemente un error de traducción.

Dadas estas dificultades, resulta esencial poner el foco, por un lado, en cómo enseñar al traductor en formación una serie de estrategias heurísticas y documentales para seleccionar equivalentes de traducción correctos para las UF. En este sentido, aún queda un largo camino por recorrer en la enseñanza de la traducción de la fraseología. Así, si bien son numerosas las investigaciones que han puesto el foco en las aplicaciones de la fraseodidáctica para la enseñanza

de segundas lenguas o lenguas extranjeras (González-Rey, 2012 y 2017; Hallsteinsdóttir, 2011a; Hallsteinsdóttir et al., 2011; Admiral, 2015; Meunier y Granger, 2008, entre muchos otros), no se ha estudiado en la misma medida sus posibles contribuciones en el ámbito de la traducción e interpretación. Para este «nicho de mercado», se han propuesto disciplinas como la *fraseotraducción*, situada en el cruce entre la fraseología, la traducción, los estudios contrastivos y la fraseodidáctica (Sułkowska, 2018), pero aún es una incógnita su posible aceptación y consolidación como disciplina *per se* en el ámbito científico. Asimismo, no solo la fraseología es aquí la única desatendida: Hurtado Albir (1999) destaca la insuficiente investigación en el ámbito de la pedagogía de la traducción y la interpretación en general, así como, en la misma línea, Pintado Gutiérrez (2012) critica el abandono que sufre la enseñanza de la habilidad traductora:

Son muchos los teóricos de la traducción, como afirmaba Gibert (1989), que se han limitado a dar por supuesto que el alumno de traducción ha adquirido plenamente la competencia en las dos lenguas con las que trabaja (la lengua origen y la lengua meta), sin tener en cuenta otras cuestiones importantes, como puede ser la propia habilidad traductora. (p. 327)

En este contexto, una de las principales motivaciones del presente proyecto de investigación es no solo contribuir a desarrollar la competencia fraseológica del traductor en formación tanto en la L1 como en la L2, sino de forma más concreta potenciar su competencia traductora, la cual conlleva inexorablemente, entre otros muchos aspectos, el aprendizaje de estrategias heurísticas e instrumentales a fin de explotar todas las posibilidades documentales que ofrecen herramientas como los corpus, los sistemas de traducción automática y los glosarios en la búsqueda de equivalencias textuales para los somatismos.

La importancia de enseñar y potenciar el uso de estas herramientas entre los estudiantes queda reflejada en recientes encuestas sobre el uso de tecnologías por parte de los traductores (Zaretskaya et al., 2015 y 2018). En ellas, se extraen resultados tan interesantes como que solo un

36 % de los traductores encuestados hacen uso actualmente de la traducción automática, si bien una mayoría (el 74 %) considera que podría sacarles mayor partido a tales sistemas siempre que ofrecieran una alta calidad. Asimismo, se indica que un número muy reducido de los traductores encuestados trabajan con corpus (17 %) o, en un porcentaje aún menor, compilan sus propios corpus (15 %), dado que la consideran como una tarea demasiado laboriosa. Así, la explotación de los corpus parece no ir más allá de la fase de elaboración de diccionarios, quedando, por tanto, inaccesibles para los usuarios finales (cfr. Durán-Muñoz y Corpas Pastor, 2020). De todo ello surge, por tanto, la necesidad de, por un lado, enseñar a los traductores a cómo sacar el máximo partido de estas herramientas y, por otro, de mejorar la traducción automática, a fin de que pueda ser de provecho para los traductores.

En lo que respecta a la relevancia de las tecnologías en el ámbito empresarial, al analizar el último *Estudio de caracterización del sector de tecnologías del lenguaje en España* (ONTSI, 2018), se observan resultados tan reveladores como que un 61,8 % de las empresas consultadas cuenta con un departamento de I+D+i, en el que invierten de media un 6 % de su volumen de facturación. En lo que concierne de forma concreta a actividades del sector de las tecnologías del lenguaje, este representa un 10,1 % del volumen de facturación de las empresas (ONTSI, 2018). Asimismo, como veremos en el apartado 4.1.1. (*Traducción Automática Neuronal*), en el ámbito europeo, las recientes encuestas sobre el sector lingüístico de Europa (European Language Industry Survey) señalan que la traducción automática está ya presente en el día a día de más de la mitad de las empresas de servicios lingüísticos europeas (ESC) participantes en el estudio (ELIS, 2018) y que un 78 % de estas piensa comenzar a utilizar o incrementar el uso de la TA o la posesición (ELIS, 2020). En esta misma línea, todos los sectores implicados (institutos de formación, desarrolladores lingüísticos y consumidores, así como profesionales independientes) también

reconocen a la traducción automática y la posedición como la tendencia más destacada con diferencia en la industria (ELIS, 2021).

Finalmente, en lo concerniente a la tipología de UF, hemos escogido analizar, de forma concreta, la fraseología somática por diversas razones: además de su carácter universal, y su elevada aparición y frecuencia de uso, los somatismos representan un grupo de unidades fraseológicas con autonomía y cohesión figurativas, entre otras características ya señaladas por Olza Moreno (2011):

Como exponentes claros de este antropocentrismo lingüístico, se puede decir que los somatismos conforman, pues, una suerte de microsistema fraseológico, cuyo examen se hace imprescindible si se quiere abordar en profundidad el estudio de la configuración semántica específica de la fraseología en cualquier idioma (Cermak 2000, 55). [...] [S]e trata, además, de un tipo de corpus fraseológico -más o menos cerrado- que se articula en torno a unos criterios comunes y autónomos de recogida y que ofrece, además, garantías de poder ser estudiado de acuerdo con instrumentos y patrones analíticos ciertamente sistemáticos. (p. 38)

El especial interés que ha suscitado el estudio de los somatismos queda asimismo reflejado en el elevado número de tesis doctorales defendidas en España que abordan, de forma específica, esta tipología de UF, como puede observarse en Tabla 1.

Autora	Título	Año
Carmen Mellado Blanco	Los somatismos del alemán. Semántica y estructura	1997
Pablo Larreta Zulategui	Fraseología contrastiva del alemán y español. Teoría y práctica a partir de un corpus bilingüe de somatismos	2001
Gonzalo Guillén Monje	Fraseología contrastiva ruso española. Análisis de un corpus bilingüe de somatismos	2004
Virginia Sciutto	Fraseologismos somáticos del español de Argentina	2005
Elizabete Aparecida Marques	Análisis cognitivo-contrastivo de locuciones somáticas del español y del portugués	2007
Inés Olza Moreno	Aspectos de la Semántica de las unidades fraseológicas. La fraseología somática metalingüística del español	2009

Carolina Julià Luna	Estructura y variación en el léxico del cuerpo humano	2010
Gustave Voltaire Dioussé	Estudio contrastivo unilateral de unidades fraseológicas (UF) con somatismos del mancagne al español	2011
Mohamad Sameer Rayyan	Fraseología y Lingüística informatizada: Elaboración de una base de datos electrónica contrastiva árabe-español/español-árabe de fraseologismos basados en partes del cuerpo	2014
Marta Saracho Arnáiz	La fraseología del español: una propuesta de didactización para la clase de ele basada en los somatismos	2015
Iva Vuksan	Los somatismos mano, dedo, puño en la fraseología española, italiana y croata	2016
Sofía Cherouveim	Los somatismos de la cabeza en español y en griego moderno. Estudio comparativo	2017

Tabla 1. Tesis doctorales defendidas en España sobre el estudio de los somatismos

De esta manera, una vez presentada la relevancia y motivación en la elección del presente proyecto de tesis, pasamos ahora a desgranar los objetivos que pretendimos alcanzar con el mismo.

2.2 Hipótesis y objetivos

Así pues, en este contexto surge nuestra hipótesis de partida: *que es posible analizar y diseñar recursos documentales y tecnológicos para la traducción de somatismos atendiendo a dos de los principales desafíos que presentan como unidades fraseológicas: la variación y la manipulación.*

De esta hipótesis de partida, se desprenden los siguientes objetivos y subobjetivos perseguidos a lo largo del presente proyecto de tesis doctoral.

Como primer objetivo (Artículo 1), se pretendió analizar el rendimiento de los actuales sistemas de traducción automática neuronal en la traducción de somatismos, poniendo especial atención a los principales desafíos que presentan la variación y la discontinuidad de estas unidades.

De este primer objetivo se desprenden los siguientes subobjetivos:

1. Contrastar el rendimiento de los dos sistemas TAN más robustos hasta la fecha — Google Translate y DeepL— en la traducción de los somatismos *tomar el pelo* y *meter la pata*, sus variantes nominales *tomadura/tomada de pelo* y *metedura/metida de pata*

y otras variantes con un empleo casi exclusivo en la variedad de español europeo como *meter la pata hasta el corvejón*, *meter la gamba* y *metedura/metida de gamba*.

2. Evaluar si existe una correspondencia directa entre una mayor frecuencia de aparición de un somatismo concreto a lo largo de las variedades nacionales y un mejor rendimiento de la TAN o si, en cambio, determinadas variedades nacionales se ven privilegiadas sobre otras en la traducción de estos somatismos.
3. Analizar el rendimiento de estos sistemas TAN en la traducción de posibles palabras raras y fuera de vocabulario, como es el caso de las variantes con escasa frecuencia de aparición *meter la pata hasta el corvejón*, *meter la gamba*, y sus variantes nominales *metedura de gamba* y *metida de gamba*.

A raíz de las dificultades detectadas, el segundo objetivo (Artículos 2-7), el más ambicioso, consistió en diseñar un sistema de preprocesamiento de unidades fraseológicas (gApp) para la conversión automática de los somatismos desde sus formas discontinuas hacia sus formas continuas, a fin de mejorar su detección y traducción automáticas. De este objetivo global, se desglosan los siguientes subobjetivos:

1. Analizar el rendimiento de Google Translate y DeepL en la traducción (ES>EN) de 560 casos con somatismos verbonominales en forma discontinua y continua, contrastando la conversión automática con gApp frente a aquella manual (Artículo 2).
2. Examinar la mejora que gApp supone para DeepL en las combinaciones ES>EN/DE y FR>EN/ES con 400 casos con somatismos verbonominales en forma discontinua y continua, nuevamente contrastando la conversión con gApp frente a la manual (Artículos 3 y 4).

3. Estudiar 300 casos con la forma discontinua y continua de variantes somáticas con la estructura de verbo + sintagma preposicional (V+SP), a fin de demostrar la mejora que supone gApp en el rendimiento de ModernMT, DeepL y Google Translate en las direccionalidades ES>EN/DE/FR/IT/PT (Artículo 5).
4. Analizar 400 casos de somatismos de estructura V+SP con gApp para los sistemas TAN de VIP, DeepL y Google Translate en la direccionalidad ES>EN (Artículo 6).
5. Examinar 400 casos de somatismos de estructura V+SP con gApp para ModernMT, DeepL y Google Translate en las direccionalidades IT>EN/DE (Artículo 7)

El tercer objetivo (Artículos 8, 9 y 10) consistió en analizar el rendimiento de los corpus como herramientas documentales para la traducción de somatismos, para lo cual se establecieron los siguientes subobjetivos:

1. Elaborar una propuesta didáctica para enseñar las posibilidades que presentan los corpus en la correcta identificación e interpretación de las UF con vistas a la traducción, poniendo especial atención a las secuencias fraseológicas susceptibles de presentar múltiples lecturas. En este contexto, se examinaron de forma concreta dos corpus monolingües del español (CORPES XXI y esEuTenTen), un corpus monolingüe en inglés (enTenTen20) y dos paralelos multilingües (Europarl7 y Linguee, subcorpus inglés-español) (Artículo 8).
2. Contrastar el rendimiento de los corpus creados por la RAE (el CREA, en sus versiones tradicional y anotada, y el CORPES XXI), el Corpus del Español y Sketch Engine ante el desafío de la variación fraseológica (Artículo 9).
3. Analizar la funcionalidad Word Sketch de Sketch Engine como herramienta documental para aquellos casos de manipulación fraseológica en el TO en los que la

ausencia de correspondencias fraseológicas biunívocas puede presentar problemas para la traducción (Artículo 10).

En el cuarto objetivo (Artículo 11) se persiguió elaborar un glosario de somatismos basado en corpus, denominado Glossomatic, para la creación de equivalentes fraseológicos *ad hoc* en aquellos casos en los que la manipulación fraseológica en el texto origen unida a la ausencia de correspondencias biunívocas interlingües podía presentar un desafío en el proceso traslaticio. En este contexto, se diseñó una propuesta didáctica para enseñar a los traductores en formación las distintas estrategias y pasos que se pueden implementar con la ayuda del glosario para enfrentarse a estos casos.

Finalmente, el quinto objetivo (Artículo 12) pretendía contrastar diferentes recursos y herramientas documentales electrónicos (corpus lingüísticos, recursos lexicográficos, la web, o los sistemas TAN, entre otros) para la creación de equivalencias textuales en aquellos casos con manipulación fraseológica en el texto origen. Para ello, en un estudio de caso llevado a cabo en el aula, se analizaron las propuestas de traducción por parte de los estudiantes a fin de determinar qué recurso y herramienta era de mayor utilidad según el tipo de manipulación en el TO.

2.3 Metodología

La metodología que hemos seguido en la elaboración de esta tesis pretende alcanzar los cinco objetivos expuestos anteriormente, todos los cuales comparten como eje vertebrador el diseño y elaboración de recursos y herramientas documentales en la traducción de somatismos atendiendo a dos de los principales desafíos que estos presentan: la variación y la manipulación. Con este fin, todas las publicaciones de la presente tesis siguen una metodología basada en corpus, ya sea para el análisis y extracción de somatismos en sus formas continuas y discontinuas (Artículos 1-7); para el estudio de las posibilidades que ofrecen los sistemas de corpus en la traducción ante el desafío

de la ambigüedad (Artículo 8), la variación (Artículo 9) o la manipulación de las UF (Artículos 10 y 12), o para el diseño del glosario de somatismos Glossomatic (Artículo 11).

Así, a lo largo del presente estudio, se analizaron tanto corpus monolingües (en español, inglés, francés, italiano, portugués y alemán) como paralelos. Entre los corpus monolingües para el español, se estudiaron dos creados por la Real Academia Española: el Corpus de Referencia del Español Actual (CREA), con más de 123 millones de formas ortográficas en su versión 0.4 (publicada en octubre de 2021) y que se encuentra accesible tanto en su versión tradicional como en la anotada, y el Corpus del Español del Siglo XXI (CORPES XXI), con más de 350 millones de formas en su versión 0.94 (publicada en julio de 2021). Se estudió asimismo el Corpus del Español de Mark Davies, y, entre los corpus que incluye en el sistema, se examinó de forma concreta el NOW (News on the Web) (Davies, 2018), que contiene más de 7 mil millones de palabras. Finalmente, los restantes corpus monolingües en español examinados se encuentran disponibles en Sketch Engine (Kilgarriif et al., 2003), un sistema de gestión y consulta que contiene 400 corpus en más de cuarenta lenguas distintas. Dentro de este, se analizaron corpus de la familia TenTen: esTenTen18 (español general), con más de 17 000 millones de palabras, y las subvariantes esEuTenTen (español europeo) y esAmTenTen (español de América). También se estudió el Timestamped JSI web corpus 2014-2021 Spanish, con más de 16 400 millones de palabras.

En lo que respecta a corpus monolingües en inglés, se analizaron tres también disponibles en Sketch Engine: por un lado, el British National Corpus (BNC), que cuenta con 100 millones de palabras, así como el enTenTen20, con 38 100 millones de palabras y el Timestamped JSI web corpus 2014-2021 English (60 400 millones). Para el francés, se analizaron dos corpus: el frTenTen17, que incluye más de 5 700 millones de palabras, y el Timestamped JSI web corpus

2014-2021 French, con 7 000 millones. En lo que respecta al alemán, se estudió el corpus deTenTen1818, que contiene 5 300 millones, así como el Timestamped JSI web corpus 2014-2021 German (7 100 millones) y el DWDS-Kernkorpus [1900–1999] (con más de 121 millones de palabras), perteneciente a la familia de corpus Digitales Wörterbuch der Deutschen Sprache (Geyken, 2007). En el caso del italiano, se examinó el corpus itTenTen16 (5 000 millones) y Timestamped JSI web corpus 2014-2021 Italian (8 700 millones). Finalmente, para el portugués, se analizó el corpus ptTenTen11 (3 900 millones) y Timestamped JSI web corpus 2014-2021 Portuguese (4 700 millones).

En lo que concierne a los paralelos, se empleó, por una parte, Linguae, una herramienta de corpus compilado a partir de traducciones alineadas con sus respectivos textos originales que se encuentran en páginas web multilingües (cuenta con más de 1000 millones de traducciones). Dentro de Sketch Engine, se emplearon los corpus OPUS2, que cuenta con 40 lenguas para una consulta tanto bilingüe como multilingüe, y Europarl7, que incluye aproximadamente 60 millones de palabras para cada una de las lenguas oficiales de la Unión Europea. En Tabla 2, se ofrece un esquema con los principales corpus empleados.

CORPUS	IDIOMAS	TAMAÑO
CREA	ES	123 M
CORPES XXI	ES	350 M
Corpus del Español NOW corpus (News on the Web)	ES	7 000 M
esTenTen18	ES	17 000 M
Timestamped JSI web corpus 2014-2021 Spanish	ES	16 400 M
British National Corpus (BNC)	EN	100 M
enTenTen20	EN	38 100 M
Timestamped JSI web corpus 2014-2021 English	EN	60 400 M
frTenTen17	FR	5 700 M

Timestamped JSI web corpus 2014-2021 French	FR	7 000 M
deTenTen18	DE	5 300 M
Timestamped JSI web corpus 2014-2021 German	DE	7 100 M
DWDS-Kernkorpus [1900–1999]	DE	121 M
itTenTen16	IT	5 M
Timestamped JSI web corpus 2014-2021 Italian	IT	8 700 M
ptTenTen11	PT	3 900 M
Timestamped JSI web corpus 2014-2021 Portuguese	PT	4 700 M
OPUS2	ES	700 M
	EN	1100 M
	FR	766 M
	DE	125 M
	IT	180 M
	PT	297 M
Europarl7	ES/EN/FR/DE/IT/PT	50-60 M per language

Tabla 2. Tamaño (en millones de palabras) de los distintos corpus analizados

Asimismo, para el análisis del significado de los somatismos en las lenguas abordadas a lo largo del estudio (español, inglés, francés, alemán, italiano y portugués) empleamos diccionarios multilingües como *Tureng Multilingual Dictionary* (TURENG, 2018), *Wordreference* (WR, n. d.), *Collins Dictionary* (COL, 2021), *Cambridge Dictionary* (CD, 2021) y *Oxford Dictionary* (OD, 2021) así como monolingües en español (como el *Diccionario de la Lengua Española* [DEL, n. d.], el *Diccionario de uso del español* [DUE, 2010] o el *Diccionario fraseológico documentado del español actual* [DFDEA, 2017]), en francés (*Larousse* [LAR, 2018]), alemán (*Duden* [DUDEN, 2006]), italiano (Treccani [TREC, n. d.] y *Corriere Della Sera/Dizionari* [CDSD, 2021]) y portugués (*Diccionario Priberam da Língua Portuguesa* [DPLP, 2021]).

Una vez presentados los principales instrumentos empleados a lo largo de la investigación para la extracción y análisis de los somatismos, pasamos ahora a describir qué metodología específica se ha seguido para alcanzar cada uno de los objetivos propuestos para el presente proyecto de tesis.

Así, para el primer y segundo objetivo (Artículos 1-7), siguiendo una metodología basada en corpus, las UF analizadas se extrajeron de esTenTen18 y Timestamped JSI web corpus 2014-2021 Spanish (para el español), frTenTen17 y Timestamped JSI web corpus 2014-2021 French (para el francés) y itTenTen16 y Timestamped JSI web corpus 2014-2021 Italian (para el italiano), todos ellos disponibles en Sketch Engine. Para dicha extracción, se empleó el lenguaje de búsqueda CQL (Corpus Query Language), mediante el cual se pueden crear patrones lexicogramaticales específicos con un alto nivel de precisión y exhaustividad. A modo de ejemplo, para obtener la forma continua y discontinua de las UF analizadas en el Artículo 1, estos fueron los patrones CQL empleados (Tabla 3).

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"]

Tabla 3. Patrones CQL para la extracción de la forma continua y discontinua de los somatismos

Tras esto, para ambos objetivos se midió el rendimiento de los sistemas TAN ante la forma continua y discontinua de las UF examinadas, siguiendo una evaluación para TA basada en ejemplo (*instance-based MT evaluation*, Zaninello y Birch, 2020), con varios posibles candidatos en el texto meta para los somatismos tanto en sus formas continuas como discontinuas. Si bien los

siete primeros artículos comparten una base metodológica análoga, existen notables diferencias en los mismos tanto en el tipo de conversión realizada (mientras que en el Artículo 1 se analizó exclusivamente la manual, en los Artículos 2-7 se contrastó la manual frente a la de gApp), en los sistemas TAN analizados (en los Artículos 1 y 2 se contrastó el rendimiento de DeepL frente al de Google Translate, en los Artículos 3 y 4 se analizó únicamente el de DeepL, en los Artículos 5 y 7 se comparó el rendimiento de DeepL, Google Translate y ModernMT, y, finalmente, en Artículo 6 los sistemas TAN examinados fueron DeepL, Google Translate y VIP), en la combinación lingüística (ES>EN en los Artículos 1, 2 y 6, ES>EN/DE en el Artículo 3, FR>ES/EN en el Artículo 4, ES>EN/DE/FR/IT/PT en el Artículo 5 y IT>EN/DE en el Artículo 7) así como en la tipología de los somatismos objeto de estudio (verbonominales en los Artículos 1-4 y con estructura de Verbo + Sintagma Preposicional en los Artículos 5-7), entre otras diferencias.

Con respecto a ese segundo objetivo (Artículos 2-7), para el diseño del algoritmo ReGap, fue de vital importancia la estancia de investigación que realicé con el Research Group in Computational Linguistics de la University of Wolverhampton (Reino Unido) entre los meses de septiembre y diciembre de 2019. Con ellos aprendí nociones fundamentales sobre Procesamiento de Lenguaje Natural mediante el lenguaje de programación Python, que empleé para el diseño del algoritmo de detección y conversión automáticas de UF discontinuas hacia sus formas continuas ReGap. Para la redacción de los patrones de detección del lexicón que alberga ReGap, se analizó la forma discontinua de las UF en los corpus esTenTen18 y Timestamped JSI web corpus 2014-2021 Spanish para la versión de ReGap en español (Artículos 2, 3, 5 y 6), frTenTen17 y Timestamped JSI web corpus 2014-2021 French para el francés (Artículo 4) e itTenTen16 y Timestamped JSI web corpus 2014-2021 Italian para el italiano (Artículo 7). Así, se analizaron tanto las concordancias con resultados relevantes (es decir, aquellas que realmente coincidían con

las UF objeto de consulta) como los irrelevantes (con patrones análogos, pero sin coincidir con la forma discontinua de estas UF), lo que nos sirvió para determinar las restricciones necesarias en los patrones de detección a fin de optimizar la precisión y exhaustividad de ReGap.

Este algoritmo Regap sirvió de base para la implementación de la aplicación gApp, para la que se empleó la herramienta Python 3.7, más la librería Spacy, especializada en la ejecución de diversas tareas avanzadas de PLN tales como tokenización no destructiva, etiquetado gramatical, análisis de dependencias, lematización y coincidencias basadas en reglas, entre otras (Honnibal y Montani, 2017). De forma más específica, se usó el modelo estadístico *es_core_news_sm* para español, *fr_core_news_sm* para el francés, e *it_core_news_sm*, para el italiano. Tras el diseño de la herramienta, con la asistencia de Javier Alejandro Fernández Sola, personal técnico de Lexytrad, se desarrolló la interfaz web utilizando las tecnologías HTML, CSS y Javascript. Esta se conecta, a través de PHP, con el script Python que está alojado en el servidor. gApp está asimismo registrado por la Universidad de Málaga (registro de la propiedad intelectual) en Safe Creative (información disponible en el siguiente enlace <https://www.safecreative.org/work/2011165898461-gapp>).

Para alcanzar el tercer objetivo (Artículos 8, 9 y 10), se analizó la interfaz de búsqueda de distintos sistemas de corpus a fin de explorar las posibilidades que ofrece cada uno de ellos en la consulta de somatismos ante los retos que presentan estas unidades: ambigüedad (Artículo 8), variación (Artículo 9) y manipulación (Artículo 10). En este contexto, se examinó tanto el sistema de concordancias simple de los corpus objeto de estudio como funcionalidades adicionales, como es el caso de la opción *Coapariciones* del CORPES XXI y de la versión anotada del CREA (Artículos 8 y 9), en la que es posible consultar somatismos discontinuos cuyos elementos constituyentes pueden estar separados por hasta nueve *tokens*, o, dentro de Sketch Engine, el

código CQL, también para la obtención de somatismos discontinuos (Artículo 9) y *Word Sketch*, para la extracción de posibles candidatos a UF a partir de un determinado somatónimo, de especial utilidad para la creación de equivalentes fraseológicos *ad hoc* a fin de conseguir una determinada manipulación en el texto meta (Artículo 10). Los artículos 8 y 10 se presentaron en forma de propuesta didáctica en la que se evaluó en qué medida los distintos sistemas de corpus eran herramientas documentales útiles para los traductores en formación ante diversos escenarios con somatismos en el texto origen.

Con respecto al cuarto objetivo (Artículo 11), para la elaboración del glosario de somatismos Glossomatic, resultó esencial la estancia de investigación que llevé a cabo con el equipo CLILLAC-ARP, de la Université Paris Diderot (Francia), entre los meses de septiembre y diciembre de 2018. Así, con la asistencia de las doctoras Natalie Kübler y Mojca Pecman, pude tomar como referencia el diccionario fraseológico bilingüe online e interactivo que este equipo diseñó fruto del proyecto ARTES², así como aprendí a cómo configurar un glosario a partir de una base de datos relacional en el sistema de gestión de bases de datos Microsoft Access 2016, desarrollado por Microsoft y perteneciente al paquete de aplicaciones de Microsoft Office.

A este respecto, para la extracción de los somatismos que constituyen Glossomatic se analizaron tanto corpus paralelos, como OPUS2 y Europarl77, y corpus monolingües en español (CORPES XXI y esEuTenTen), inglés (BNC y enTenTen20), y alemán (deTenTen1813 y DWDS-Kernkorpus [1900–1999]). El análisis de los contextos de uso real de los somatismos en los corpus empleados permitió asimismo el diseño de las diferentes pestañas de consulta de Glossomatic (*colocados, polaridad, registro*, etc.) así como útil información contrastiva sobre los equivalentes de traducción en la combinación de lenguas español-inglés-alemán. El Artículo 11 también se

² El diccionario ARTES se encuentra disponible en la siguiente dirección URL: <https://artes.eila.univ-paris-diderot.fr/>

presentó en forma de propuesta didáctica, a fin de determinar en qué medida Glossomatic podía asistir a los traductores en formación en la creación de equivalentes fraseológicos *ad hoc* ante casos de manipulación fraseológica en el texto origen.

Finalmente, para lograr el objetivo 5 (Artículo 12), se seleccionaron seis casos de traducción publicitaria con distintos tipos de UF manipuladas en el texto origen: por un lado, 3 con una tipología heterogénea de UF frente a otros 3 con somatismos. El objetivo de tal distinción era determinar en qué medida la herramienta documental y el tipo de estrategia de traducción escogidos por el estudiante venían determinados por la aparición o no de un somatismo. Entre las herramientas documentales se analizaron corpus paralelos y monolingües disponibles en Sketch Engine, distintos recursos lexicográficos, los sistemas de Traducción Automática Neuronal, los motores de búsqueda (Google, Bing, Metasearch, etc.) y los rimadores, entre otros. Las distintas propuestas de traducción de los 45 estudiantes participantes en el estudio se recogieron y analizaron en la herramienta de creación de formularios Googleform, que permite una fácil visualización y procesamiento de los resultados almacenados.

En este contexto, una vez detallados la motivación, hipótesis, objetivos y metodología, pasamos ahora a desglosar el marco teórico en el que se circunscribe la presente tesis doctoral.

CAPÍTULO III. MARCO TEÓRICO: LOS SOMATISMOS Y SU TRADUCCIÓN

3.1 Sobre el concepto de *fraseología* y *unidad fraseológica*

Nacida como disciplina científica en la URSS de la década de los cincuenta, la fraseología entró en el panorama del español de la mano de autores como Alberto Zuluaga y Julio Casares, cuya impronta quedaría reflejada en los trabajos de los estudiosos que les sucederían y que contribuirían a su definitiva consolidación. Hablamos de investigadores de la talla de Antonio Pamies Bertrán, Jesús Cantera Ortiz de Urbina, M.^a Inmaculada Penadés Martínez, Mario García-Page, María Isabel González Rey, Gerd y Barbara Wotjak o Gloria Corpas Pastor, entre muchos otros.

A fin de comenzar a delimitar el objeto de estudio de la fraseología, partiremos de la conocida dicotomía propuesta por Sinclair (1991): el *principio de libre elección* (*open-choice principle*) frente al *principio idiomático* (*idiom principle*). Así, en todo acto de habla, es posible detectar, por un lado, unidades léxicas simples con significado autónomo, cada una de las cuales constituye una elección independiente (*principio de libre elección*) y, por otro, sintagmas prefabricados cuyos elementos integrantes no son elegidos libremente por el hablante sino que están predefinidos (*principio idiomático*). En esta misma línea, contrasta Zuluaga (1980) las combinaciones libres frente a las expresiones fijas:

Las combinaciones libres son producidas en cada acto de habla mediante las reglas de la gramática actual, son productos de procedimientos propios de la sincronía de la lengua. Las expresiones fijas, en cambio, son construcciones anteriores al acto de habla que no las crea sino que las repite; en otras palabras, las expresiones fijas son producto de procesos de repetición en la diacronía de la lengua. (p.16)

Así, del estudio de estas «expresiones fijas» (que, en adelante, denominaremos *unidades fraseológicas*) se encarga la disciplina fraseológica. A este respecto, es necesario precisar que gran parte de las combinaciones de palabras que componen el discurso no presentan la libertad de

elección o la fijación en términos absolutos, sino que conforman un extenso contínuum entre ambos parámetros, de ahí, como veremos, la dificultad de establecer unos límites nítidos.

En este sentido, la diversidad de estructuras englobadas bajo el término genérico de *unidad fraseológica* ha dificultado enormemente el establecimiento de límites, tipologías o denominaciones aceptadas por toda la comunidad científica. Prueba de ello es la amplia heterogeneidad terminológica existente en la literatura científica para, con distintos matices, hacer referencia a este tipo de estructuras: unidades fraseológicas, fraseologismos, frasemas, expresiones fijas, expresiones pluriverbales, unidades pluriverbales, unidades plurilexemáticas o unidades poliléxicas, entre muchos otros. Por cuestiones de espacio, no trataremos aquí de forma pormenorizada este asunto, baste indicar que a lo largo de del presente proyecto de tesis doctoral³ emplearemos la definición de *unidad fraseológica* propuesta por Corpas Pastor, en su obra *Manual de fraseología española* (1996), que sigue siendo hoy día referencia obligada para la disciplina:

Unidades léxicas formadas por más de dos palabras gráficas en su límite inferior, cuyo límite superior se sitúa en el nivel de la oración compuesta. Dichas unidades se caracterizan por su alta frecuencia de uso, y de coaparición de sus elementos integrantes; por su institucionalización, entendida en términos de fijación y especialización semántica; por su idiomatización y variación potenciales; así como por el grado en el cual se dan todos estos aspectos en los distintos tipos. (Corpas Pastor, 1996, p.20).

Así, Corpas Pastor (1996, pp. 50-53) establece una taxonomía de las UF en torno a dos criterios principales: el enunciado y la fijación. La autora entiende por *enunciado* una «unidad de comunicación mínima, producto de un acto de habla, que corresponde a una oración simple o compuesta, pero que también puede constar de un sintagma o una palabra» (p. 51). Así, distingue

³ De forma excepcional, en los Artículos 1-7 empleamos *expresión pluriverbal* (del inglés *multiword expression* o *MWE*) para referirnos a estas unidades, dado que este término presenta mayor frecuencia de uso en el campo de la lingüística computacional, en el que se circunscriben las mencionadas publicaciones.

en un primer momento aquellas unidades que no conforman enunciados ni actos de habla, esto es, que necesitan combinarse con otros signos lingüísticos, ya que equivaldrían a sintagmas. Dentro de esta división, diferencia asimismo la esfera I con aquellas unidades fijadas solo en la norma, las *colocaciones*, de la esfera II, las UF fijadas en el sistema, las *locuciones*. Finalmente, en la esfera III, incluye aquellas UF fijadas en el habla que conforman actos de habla a través de enunciados completos y que denomina *enunciados fraseológicos*. Con esta taxonomía, Corpas Pastor incorpora unidades que ni Casares (1950) ni Zuluaga (1980) habían considerado como UF, tales como las paremias, los compuestos de dos o más palabras gráficas y las colocaciones. Es esta, por tanto, una perspectiva ancha de la fraseología, que emplearemos asimismo en nuestra investigación. La estrecha, defendida por investigadores como Ruiz Gurillo (1997) y García-Page (2008), incluye exclusivamente a las locuciones como unidades propias del ámbito fraseológico.

A este respecto, han surgido nuevas categorías complementarias a las taxonomías clásicas fruto de la reciente imbricación de la fraseología con la Gramática de Construcciones, que ha supuesto un auténtico cambio de paradigma en los estudios fraseológicos (cfr. Corpas Pastor, 2021; Dobrovol'skij, 2011a; González Rey, 2015; Ziem, 2018, entre otros). Así, la Gramática de Construcciones concibe la lengua como una red compleja de construcciones, es decir, de uniones de forma y significado en diferentes niveles de abstracción y esquematicidad (cfr. Goldberg, 1995 y 2006; Stefanowitsch, 2011), por lo que, en lugar de situar a las UF en una esfera diferente a otro tipo de unidades lingüísticas, propone la existencia de un contínuum entre léxico y sintaxis: en el polo léxico se encontrarían los elementos saturados léxicamente (morfemas, palabras simples y compuestas, locuciones, paremias, etc.), en un espacio intermedio se situarían secuencias parcialmente saturadas (por ejemplo, *dar por* $X_{AGG/PART. PRES./PART. PASS}$) o mínimamente saturadas (*cuanto*

más..., *más*) y en el polo gramatical se situarían las construcciones abstractas, tales como la construcción ditransitiva o la pasiva, entre otras (Mellado Blanco, 2020).

En ese espacio intermedio surgiría una categoría complementaria que no había sido abordada en profundidad en la investigación tradicional clásica: las *construcciones fraseológicas* o *idiomáticas*⁴⁵, correspondiente a los conceptos de *constructional idioms* (Taylor, 2016) o *partially lexically filled phrasal patterns* (Goldberg, 2006), en inglés, y *Phrasem-Konstruktionen* (Dobrovól'skij, 2011a) o *Phraseoschlabonen* (Ziem, 2018), en alemán, términos bajo los que se engloban las siguientes estructuras:

Constructional idioms are syntactic constructions with a (partially or fully) noncompositional meaning contributed by the construction, in which —unlike idioms in the traditional sense— only a subset (possibly empty) of the terminal elements is fixed. (Booij, 2002, p. 320)

En este contexto, es necesario señalar que, además de las construcciones fraseológicas, desde la Gramática de Construcciones también es posible abordar el estudio de las otras esferas de la fraseología anteriormente descritas: locuciones, colocaciones, pemiias, etc. No obstante, a lo largo del presente proyecto de tesis doctoral hemos optado por seguir el marco de la fraseología tradicional, dado que, en el ámbito hispano, es el que ofrece un mayor número de publicaciones

⁴ Como apunta Mellado Blanco (2015), el término «construcción fraseológica» coincidiría parcialmente con lo que en la tradición fraseológica española ha recibido el nombre de «esquemas fraseológicos» (Zuluaga, 1980, pp. 159-160, y Ruiz Gurillo, 1997, pp. 117-120), «esquemas sintácticos» (García-Page, 2007), o «locuciones de casillas vacías» (Montoro del Arco, 2008). No obstante, existen algunas diferencias importantes entre ellos:

(1) el término esquema está asociado, en la fraseología española, a la herencia estructuralista y no contempla en primer término el análisis del significado global de estas unidades [...], (2) en [...] la Gramática de Construcciones, el término “construcciones” encuentra mejor cabida [y] (3) la “construcción fraseológica” tiene un espectro de aplicación más amplio que los esquemas fraseológicos, que, por lo general, son de valor sintagmático, suboracional. A diferencia de estos, las construcciones fraseológicas engloban también unidades de valor oracional, e incluso pueden entenderse a nivel supraoracional (textual). (Mellado Blanco, 2020, pp. 14-15)

en torno a las locuciones, en las que se encuadran la mayor parte de los somatismos. Esto nos ha facilitado enormemente la tarea de construir un marco teórico sólido sobre el que orientar el enfoque eminentemente práctico del presente proyecto. No obstante, como indicamos en el Capítulo V (*Discussion, conclusion and future lines of research*), en posteriores investigaciones pretendemos aplicar el prisma construccionista para analizar UF no abordadas en el presente estudio.

De esta manera, una vez delimitado el objeto de estudio de la disciplina fraseológica, en el próximo apartado abordaremos, de forma específica, una tipología de UF predominante en el acervo fraseológico de las lenguas: los somatismos.

3.2 Sobre el concepto de *somatismo* y su lugar en el universo fraseológico

Siguiendo la definición de Mellado Blanco (2004, p. 22), referencia obligada en este campo, «los SO [somatismos] son FR [fraseologismos] que contienen un lexema referido a un órgano o parte del cuerpo humano, a veces animal». Además de su carácter universal, la autora destaca su elevada aparición y frecuencia de uso, tanto en el registro oral como en el escrito: en alemán, por ejemplo, los somatismos constituyen un 13,8 % del total de las unidades fraseológicas de dicha lengua (Mellado Blanco, 2004, p. 23). Otros autores han señalado asimismo este predominio de los somatismos en los sistemas fraseológicos de las distintas lenguas:

No parece haber duda de que en la elección de los dominios que sirven como bases para el desarrollo del contenido idiomático de las expresiones fijas, los hablantes muestran un claro *antropocentrismo* y que, por ello, las UF que no están relacionadas de algún modo con el hombre o con lo humano [...] constituyen un grupo menor dentro de la fraseología de cada lengua (Makkai 1993, 320; Fernando 1996; Čermák, 2000, 57-58). Es más, dentro ya del conjunto de expresiones fraseológicas relacionadas con algún dominio de lo humano, los somatismos constituyen un grupo llamativamente nutrido en términos

absolutos (cf. Beinhauer 1941, *apud* Polo 2004, 121; Wotjak, 1988, 542; Fernando 1996, 19; Čermák, 2000, 57; Kövecses 2002, 16; Mellado Blanco 2004, 22-23; Nissen 2006, 95; Julià Luna, 2010). (Olza Moreno, 2011, p. 38)

En lo que respecta a los principales rasgos de las unidades somáticas, Mellado Blanco (2004) ofrece una pormenorizada clasificación de la que podemos destacar los siguientes parámetros. Así, desde un punto de vista categorial-funcional, los somatismos sintagmáticos suelen ser verbales (*tomar el pelo, meter la pata, arrimar el hombro*) y adverbiales (*de la cabeza a los pies, de cabo a rabo*, etc.). En menor proporción, existen también somatismos nominales (*tomadura de pelo, metedura de pata*, etc.) y adjetivales (*cara dura, ojo avizor*, etc.). Desde una perspectiva semántico-funcional, los somatismos «se definen en general por resaltar los aspectos negativos de estos estados humanos, como agotamiento físico, falta de inteligencia o de destreza, inclinaciones desmedidas a hablar o comer, poca sutileza en el trato con los demás, etc. [...] Tal es el grado de negatividad que denotan y connotan que muchos de los SO sólo son usados normalmente en la segunda y en especial en tercera persona verbal» (Mellado Blanco, 2004, pp. 23-24). Finalmente, desde un punto de vista semántico-onomasiológico, los somatismos son definibles en 3 niveles: por el significado directo de los componentes por separado, en virtud del cual todos los somatismos pertenecen al campo léxico-semántico de «partes y órganos del cuerpo» (nivel 1); por el significado literal de la secuencia completa, a partir del cual pueden encuadrarse en distintos campos conceptuales como «limpieza» (*lavar el cerebro*), «suciedad» (*mancharse las manos*), etc. (nivel 2), y por el significado fraseológico global, según el cual los somatismos pueden clasificarse en diferentes campos conceptuales como «pensamiento» (*rondar por la cabeza, pasar por la cabeza*, etc.), «ayuda» (*echar una mano, arrimar el hombro*, etc.) (nivel 3) (Mellado Blanco, 2004, p. 25). Como veremos en el apartado 4.2.2. (*Glosarios*), esta potencial clasificación de los distintos

somatismos en campos conceptuales facilitará enormemente la repertorización onomasiológica de los mismos en recursos lexicográficos como el glosario Glossomatic (Artículo 11).

En esta misma obra, la autora ofrece una clasificación adicional de los somatismos atendiendo a la metáfora y metonimia conceptual (Lakoff y Johnson, 1980) que los originan. Así, partiendo de la metáfora espacial de horizontalidad LO BUENO ES «ARRIBA»/LO MALO ES «ABAJO», encontramos distintos subgrupos como ESTADO DE ÁNIMO POSITIVO ES «ARRIBA» (*levantar cabeza, con la cabeza alta...*), y ESTADO DE ÁNIMO NEGATIVO ES «ABAJO» (*con la cabeza baja, caérsele el alma a los pies...*); TENER CONTROL O FUERZA ES «ARRIBA» (*mirar por encima del hombro, subírsele a la chepa [a alguien]...*) y ESTAR SUJETO A CONTROL ES «ABAJO» (*[ponerse] a los pies [de alguien], [ponerse] de rodillas [ante alguien]...*). Una misma dicotomía se presenta en la metáfora espacial de verticalidad, donde LO BUENO ES «DELANTE»/LO MALO ES «DETRÁS». Así, «DELANTE» está vinculado a distintos semas como ‘ánimo y resolución’ (*plantar cara, verse las caras, sacar la cara por alguien...*), ‘abiertamente, sin tapujos’ (*decir a la cara, reírse en la cara [de alguien]...*). Por su parte, podemos asociar «DETRÁS» a semas como ‘rechazo’ (*dar la espalda*), ‘traición’ (*dar una puñalada por la espalda, a espaldas de*), ‘carga o responsabilidad’ (*echarse a las espaldas*), entre otros (Mellado Blanco, 2004, pp. 56-71).

A este respecto, el hecho de que buena parte de estas metáforas y metonimias conceptuales sean subyacentes a los acervos somáticos de las distintas lenguas aquí estudiadas permite que puedan establecerse diferentes posibilidades de equivalencia en traducción entre las mismas, como puede observarse en la siguiente tipología de Gutiérrez Pérez (2010, pp. 65-66):

- 1) Una misma metáfora o metonimia conceptual ha lexicalizado de forma análoga en ambas lenguas: *perder la cabeza: to lose one's head*.

- 2) Ambas lenguas comparten una misma proyección metafórica o metonímica, pero esta ha seguido distintos procesos de convencionalización en su lexicalización (si bien se trata de partes del cuerpo contiguas o interrelacionadas): *morderse la lengua: to bite one's lips*.
- 3) Una misma metáfora o metonimia conceptual ha producido, en algunas ocasiones, diferentes extensiones lingüísticas. Así, la metáfora «CONTROL ES TENER ALGO EN LA MANO» sí se encuentra en ambas lenguas en el somatismo *in one's hands* (*my life was in his hands: mi vida estaba en sus manos*), pero no en el caso de *in the hollow of one's hand* (generalmente empleada en la expresión *May God hold/keep you in the hollow of His hand*, FD, n. d.), para la que no existe un equivalente somático en español.
- 4) No existe una misma proyección metafórica o metonímica en ambas lenguas, como el caso del dedo pulgar vinculado a la torpeza en inglés: *to be all (fingers and) thumbs: ser un manazas*.
- 5) Los acervos somáticos de ambas lenguas cuentan con UF de similar significado pero conceptualizadas en distintas partes del cuerpo no interrelacionadas: *by the skin of one's teeth: por los pelos*.

En el estudio de caso presentado en Artículo 12, podremos comprobar como este elevado número de metáforas y metonimias conceptuales que comparten las lenguas facilita enormemente al traductor la creación de equivalentes fraseológicos *ad hoc* para somatismos del texto origen sin correspondencia primaria en la lengua meta. A este respecto, en el próximo apartado profundizaremos aún más en las distintas posibilidades existentes para la traducción de las UF.

3.3 Sobre la traducción de las unidades fraseológicas

Corpas Pastor (2003, pp. 216-222) ha descrito por extenso los principales pasos en la traducción de unidades fraseológicas: identificación de la UF en texto origen, interpretación en contexto, y búsqueda y establecimiento de sus correspondencias en el plano léxico, textual y discursivo. En relación con este último paso en la traducción de UF, numerosos autores (Corpas Pastor, 2003; Koller, 2007; Pedersen, 1997; Roberts, 1998; Svensén, 2009, entre otros) afirman asimismo que, en la medida de lo posible, las UF deben ser trasladadas en el TM por medio de un equivalente fraseológico, dada la riqueza que proporcionan al discurso:

Idioms give flavour to the text, and their absence consequently impoverishes it. That an idiom sometimes has no equivalent poses no immediate problem; it can be rendered by an unidiomatic expression, or by a single word. But if this solution is embraced regularly in a given text; the translation becomes flatter than the original. (Pedersen, 1997, p. 109)

En este sentido, Roberts (1998, pp. 75-76) reconoce que existen tres maneras de traducir una UF: mediante otra UF equivalente en la LM, mediante una sola palabra, y mediante una paráfrasis de su contenido semántico. No obstante, señala que un equivalente fraseológico, si lo hay, debe ser la opción preferente para el TM. Así lo defiende también Svensén (2009, p. 156), «[i]dioms in the source language must as far as possible be paralleled in the target language by idioms with the same content», y Dobrovolskij (2011b, p. 14), «in texts where they [idioms] are communicatively relevant, their loss in translation leads to an informational and aesthetic decline». Sobre esta difícil permuta de una UF por un término neutro incide asimismo Mogorrón Huerta (2012):

Además conviene señalar y destacar que las UF se utilizan con un objetivo discursivo muy claro. Representan el deseo del locutor de utilizar en el acto de comunicación una determinada fórmula refrendada por la mayor parte de la comunidad lingüística, sumándole de esa forma un matiz cultural, social, generacional, etc., en vez de utilizar un término neutro o una perífrasis verbal. (p. 1)

En esta dicotomía UF o término neutro, Zuluaga (2001, pp. 72-73) arroja algo más de luz al resaltar las cinco funciones que aportan las unidades fraseológicas al discurso. Por un lado, encontramos la «función fraseológica», como garantía de comunicabilidad y eficiencia, dado que simplifica el mensaje con una construcción ya prefabricada y conocida en la comunidad lingüística. Por su parte, la «función de realce» permite poner en relieve ese fragmento de discurso dada la anomalía morfosintáctica y semántica que suelen presentar las UF en contraste con el resto de unidades léxicas simples. Aunque no exclusiva, otra función en las UF es la «función de connotación», al poseer una concreta marca diasistemática con la que evocan a una comunidad lingüística, nivel sociocultural o entorno determinados al insertarse en el discurso. Por otro lado, la «función icónica», especialmente presente en las UF idiomáticas y semiidiomáticas, se define por mostrar un contenido a través de la fuerza expresiva que aporta una imagen visual, lo que las hace poseer tanto un sentido literal como figurado (*meter la pata, arrimar el hombro, dejarse la piel*, etc.). De este carácter diédrico de las UF (plano literal + plano figurado) así como de los procedimientos fonoestilísticos aparecientes en muchas de ellas (en especial, en las paremias) nace la «función lúdico-poética», con la que, además de la sonoridad que aportan determinadas UF, muchas de ellas se presentan en el discurso con algún tipo de manipulación, como podremos observar por extenso en el apartado 3.4.5.2 (*Manipulación*) de la presente tesis doctoral. Así, para Zuluaga (2001), estas funciones inherentes de las UF hacen que resulte esencial encontrar, en la medida de lo posible, equivalentes fraseológicos para el texto meta:

Debido a las funciones inherentes de las UF, concedemos prioridad a la búsqueda de equivalencias también fraseológicas en la traducción de textos en que se presenten estas unidades. [...] Por lo tanto, consideramos como ejemplos loables de creatividad y de lealtad para con el original aquéllos en que el traductor, a falta de correspondencias en la lengua meta, construye *ad hoc* una estructura equivalente con forma material propia de UF; así salva las funciones mencionadas en nuestro análisis. (p. 79)

Así, en los dos próximos subapartados, profundizaremos en las distintas posibilidades existentes para afrontar ese último paso en el proceso traslaticio de UF (Corpas Pastor, 2003, pp. 216-222), es decir, la búsqueda y establecimiento de correspondencias en el plano léxico, textual y discursivo; unas correspondencias que, como veremos, no siempre podrán (o deberán) ser necesariamente fraseológicas.

3.3.1 *La equivalencia translémica*

Al abordar la traducción de las unidades fraseológicas de una lengua a otra, resulta clave el concepto de equivalencia translémica. Un translema es una «unidad bitextual de cualquier tipo o nivel constituida por un mismo contenido y dos manifestaciones formales diferenciadas pero solidarias y cuya existencia depende de la relación global de equivalencia subyacente a cada binomio textual TO-TM» (Santoyo y Rabadán, 1991, p. 322). En este sentido, Corpas Pastor (2003, p. 206) diferencia cuatro tipos de equivalencia translémica: *equivalencia total*, *equivalencia parcial*, *equivalencia nula* y *equivalencia aparente* («falsos amigos»). Para la autora, la *equivalencia total* tiene lugar cuando ambas locuciones conforman un translema en todo momento, esto es, «cuando ambas se traducen mutuamente mediante la técnica de transferencia denominada *sustitución*» (Corpas Pastor, 2003, p. 206), como por ejemplo *no tener ni pies ni cabeza* con significado equivalente a *not to make head nor tail*. En la *equivalencia parcial*, la locución de la lengua origen (LO) y la de la meta (LM) presentan divergencias en sus significados denotativos o connotativos o en las variedades diastráticas, diafásicas o diatópicas a las que pertenecen, o la de la LO «es traducida por una unidad léxica simple carente de los valores expresivos de la locución en cuestión» (Corpas Pastor, 2003, p. 207). La no compensación de dicha equivalencia en contexto puede provocar *infratraducción* o *sobretraducción*. Un ejemplo de equivalencia parcial lo encontramos en la UF *keep your hair on!* ('said to tell someone to stop being so angry or upset',

CD, 2021) que podría traducirse por *¡mantén la calma!* (WR, n. d.), una UFM carente de los valores expresivos de la UFO. Un tercer tipo de equivalencia sería la *equivalencia nula*, que incluye aquellas UFO que no poseen una UFM equivalente, de ahí que en estos casos el traductor deba decantarse por la *paráfrasis* u otra técnica de *transferencia*, lo que por lo general implica una pérdida de las connotaciones correspondientes (Corpas Pastor, 2003, p. 208). Ejemplo de ello es la UF procedente del ámbito taurino *ponerse el mundo por montera* ('despreocuparse de las opiniones ajenas o de las conveniencias sociales', DFDEA, 2017, p. 531), sin UFM equivalente en inglés. Finalmente, con el concepto de *equivalencia aparente*, Corpas Pastor (2003) hace referencia a aquellas UFO y UFM con análoga forma, pero diferente significado («falsos amigos»), como por ejemplo la UF *to keep one's head down* que no es equivalente a *estar cabizbajo* sino a *pasar desapercibido*. Asimismo, incide en que, de todas ellas, la que con mayor frecuencia se presenta es la *equivalencia parcial*:

La mayor parte del frásico de las lenguas puebla el vasto territorio que se extiende desde un extremo (la equivalencia nula) al otro (la equivalencia total). Esto es, las correspondencias fraseológicas que se establecen entre dos lenguas distintas suelen ser relaciones de equivalencia parcial por cuanto las UFS implicadas presentan ciertas diferencias y solapamientos con respecto a uno o varios de los siguientes aspectos: contenido semántico, base figurativa, características morfosintácticas, restricciones diasistemáticas y rasgos pragmáticos. (Corpas Pastor, 2003, p. 254)

En esta misma línea, Koller (2007, p. 205) presenta una clasificación también con cuatro tipos de equivalencia: *equivalencia 1:1* (equivalencia total), caracterizada por la existencia en la LM de un equivalente semántico, con análoga forma léxica y mínima o ninguna diferencia connotativa; *equivalencia 1:sustitución*, es decir, equivalente semántico con mínima o ninguna diferencia connotativa pero distinta forma léxica; *equivalencia 1:parte* (equivalencia parcial), con equivalente semántico, y pequeñas diferencias en el plano léxico o sintáctico o en la connotación,

y, finalmente, *equivalencia 1:0* (equivalencia nula), en la que la LM carece de un equivalente semántico. Por tanto, las principales diferencias entre la clasificación de Corpas Pastor y la de Koller, es que este último no incluye el concepto de *equivalente aparente* y desglosa un nuevo subtipo (*equivalencia 1:sustitución*) que pertenecería a la categoría de *equivalencia parcial* en la clasificación de Corpas Pastor.

A los equivalentes de este último subtipo (*equivalencia 1:sustitución*), Dobrovol'skij (2011) los denomina *paralelos fraseológicos*, es decir, UF de la LO y de la LM que se corresponden en su significado nuclear pero difieren en su componente de imagen como, por ejemplo, *dar (cien/cincuenta/mil...) vueltas [a alguien]* en español (es decir, 'ser muy superior [a alguien]', DFDEA, 2017, p. 915), y *to be head and shoulders [above someone]* en inglés ('to be far superior [to someone]', OD, 2021).

A diferencia de las propuestas aquí presentadas, no todos los autores han concedido la misma importancia a la imagen como factor determinante para establecer el grado de equivalencia transléfica entre dos UF. Para Mellado Blanco (2015), esta tendencia se debe principalmente al fenómeno de la *desautomatización* que, como veremos en el apartado 3.4.5.2 (*Manipulación*), implica una (re)activación y una (re)motivación del plano literal de una UF, de ahí la importancia de la imagen. A este respecto, la autora incide en que «parece que no tiene mucho sentido restar equivalencia interlingüística a dos unidades con similar significado y función, solo por el hecho de que se apoyen en imágenes distintas» (Mellado Blanco, 2015, p. 163), especialmente, cuando no se adopta el mismo criterio para las unidades léxicas simples. Por ejemplo, el sustantivo *saltamontes* puede considerarse un equivalente total (al menos, en el plano sistémico) de *grasshopper*, aun cuando ambos presentan imágenes distintas: en el primer caso al animal se le atribuiría la habilidad de *saltar montes* y, en el segundo, de *saltar la hierba*, lo que no supondría,

por ello, un menor grado de equivalencia. Así, según Mellado Blanco y de forma análoga a las unidades léxicas simples, el hecho de compartir o no una misma imagen entre dos UF no debería influir en el grado de equivalencia transléfica entre las mismas.

3.3.2 *La equivalencia textual*

En torno al concepto de *equivalencia fraseológica*, Dobrovol'skij (2013) considera necesario distinguir dos aspectos distintos. Por un lado, encontramos la *equivalencia en el sistema* (abordada en el apartado anterior), es decir, la relación entre la UF de la LO y la de la LM en el nivel sistémico, y, por otro, la *equivalencia en traducción*, esto es, la relación entre ambas UF (LO y LM) pero esta vez en el nivel textual. Así, considera la equivalencia en traducción como una relación *unilateral* (*unívoca*, en la terminología de Mellado Blanco, 2015, p. 154) y la equivalencia sistémica como una relación *bilateral* entre las UF. Con respecto a esta dicotomía, existe un amplio consenso en que, en el proceso traslaticio, deben establecerse equivalencias funcionales o textuales antes que equivalencias sistémicas (cfr. Corpas Pastor, 2003; Dobrovol'skij, 2013; Hallsteinsdóttir, 2011b):

- 1) In the case of phraseology, the text defines the role of a phraseme and hence, the text dominates the decision of an equivalent in L2 (cf. Mellado Blanco 2009; cf. also Hallstendóttir/Faro 2010). The translation of phrasemes is dominated by the textual equivalence and the L2-text textuality as defined in Neubert and Shreve (1992), and not by the potential equivalence in the language systems in L1 and L2 (cfr. Hallsteinsdóttir 2006; Mellado Blanco 2009). (Hallsteinsdóttir, 2011b, p. 286)
- 2) La traducción de la fraseología no se reduce, evidentemente, al mero establecimiento de relaciones de (in)equivalencia entre sus respectivos frásicos. La traducción es un fenómeno eminentemente textual. (Corpas Pastor, 2003, p. 253)

Así, en todo momento, debe primar la reproducción de la carga semántica, pragmática y discursiva de la UF origen en el texto meta, por encima de traducir cada UF por su correspondencia primaria, a fin de no caer en lo que Farø (2006) denomina *dogmatismo fraseológico*, esto es, el intento desmedido de traducir cada UF del TO por una UF en el TM, sea apropiado o no.

A este respecto, Corpas Pastor (2003, pp. 263-269) menciona cinco procedimientos principales en la traducción de una UF. El primer procedimiento sería la *équivalence* (Vinay y Dalbérnet, 1995), que consistiría en la búsqueda de un translema equivalente en el repertorio fraseológico bilingüe y el reemplazo de la UFO por la UFM, como puede observarse en las siguientes concordancias⁶ con las UF y su traducción en negrita:

[ES] Al **abrir la mente** y hacer nuevas cosas es cuando descubres hasta donde puedes llegar.

[EN] When you **open your mind** and do new things you discover how far you can go.

En caso de que el traductor no consiga encontrar un translema equivalente en el plano textual (o, aun existiendo, decida no decantarse por el mismo), un segundo procedimiento consistiría en la paráfrasis del contenido semántico-pragmático de la UFO, como en el siguiente caso:

[ES] Mr President, I have a sense of **déjà vu** here, having drafted the opinion of the Committee on Employment and Social Affairs on a number of these issues.

[EN] Señor Presidente, tengo cierta sensación de **haber vivido ya esta situación**, pues he sido ponente de opinión de la Comisión de Empleo y Asuntos Sociales en algunos de estos temas.

Un tercer procedimiento sería la omisión de la UF bien porque el traductor no encuentre un equivalente en la lengua meta bien porque no lo considere relevante en la comprensión del discurso. Como señala la autora, esta omisión conlleva normalmente pérdidas de carácter semántico o

⁶ Las concordancias de este apartado se han extraído de Linguee.com (LING, 2008)

pragmático en el TM, como el caso que aquí analizamos, en el que la omisión de la UF inglesa *for goodness sake* ('an expression of annoyance', CD) hace que el texto meta en español pierda el matiz expresivo y enfático del texto origen:

[EN] But **for goodness sakes**⁷, let the nuclear industry itself decide if it wants to make the investment.

[ES] Pero que sea la misma industria de la energía atómica la que calcule con detalle si decide llevar a cabo tal inversión.

A fin de compensar esta omisión, un cuarto procedimiento consistiría en incluir unidades en otras partes del TM aun no existiendo en el TO con el fin de hacer más natural el texto, como podemos observar en el mismo binomio textual anterior (en subrayado el fragmento introducido y en negrita la UF):

[EN] But for goodness sakes, let the nuclear industry itself decide if it wants to make the investment.

[ES] Pero que sea la misma industria de la energía atómica la que calcule **con detalle** si decide llevar a cabo tal inversión.

Finalmente, la autora menciona un quinto procedimiento: el calco. Así, en el siguiente ejemplo (analizado asimismo en Artículo 8), podemos detectar en el TO la UF *keep someone's head down* con el significado de 'to remain inconspicuous in difficult or dangerous times' (EOLD, 2018). En el TM se ha optado por calcar la estructura hacia *bajemos la cabeza*, una UF que posee un significado completamente distinto al de la UFO ('obedecer y ejecutar sin réplica lo que se manda' [DLE, n. d.] y 'conformarse, tener paciencia cuando no hay otro remedio' [DLE, n. d.]), de ahí el error de traducción:

⁷ Si bien la forma más usual es en singular (*for goodness sake*) con una frecuencia de 0,18 por millón en enTenTen20, también encontramos instancias de esta UF en plural (*for goodness sakes*) con una frecuencia de 0,03 por millón en el mismo corpus.

[EN] There is still to this day a culture of 'let's **keep our heads down** and not report wrongdoings' in the European Commission.

[ES] En estos momentos, en la Comisión Europea aún impera una cultura de "**bajemos la cabeza** y no informemos de los errores".

3.4 Características principales de las unidades fraseológicas

Una vez examinadas las diferentes posibilidades en la traducción de unidades fraseológicas, a lo largo del presente apartado analizaremos dos de los principales retos que las UF presentan en el proceso traslaticio (la variación y la manipulación), derivados de sus dos características principales: la fijación y la idiomatidad.

3.4.3 Fijación

Se entiende por *fijación* «la propiedad que tienen ciertas expresiones de ser reproducidas en el hablar como combinaciones previamente hechas» (Zuluaga, 1975, p. 230). Para Thun (1978)⁸, las UF pueden presentar fijación externa o interna. En lo que respecta a la externa, es posible distinguir cuatro tipos: la situacional, la analítica, la pasemática y la posicional. En este sentido, Corpas Pastor (1996, p. 24) explica que la primera hace referencia a las unidades empleadas en situaciones sociales concretas como por ejemplo *buenos días*, *¿qué tal?*, *hasta luego*⁹. Por su parte, con *fijación analítica* se refiere a aquellas combinaciones de palabras empleadas para analizar los hechos del mundo, en contraste con otras igualmente posibles, en teoría. Con respecto a la tercera, la *pasemática*, se presenta en el uso de unas unidades lingüísticas dependientes del papel que ocupe el hablante en el acto comunicativo. Por ejemplo, como se señala en DFDEA (2017), la UF *no*

⁸ Op. cit. en Corpas Pastor (1996, pp. 23-24)

⁹ A este tipo de fijación, Zuluaga (1980) la denomina *fijación pragmática*, con la que hace referencia a aquellas unidades que no solo están fijadas en el plano sintáctico, sino también en el pragmático, es decir, aquellas unidades restringidas a una situación social específica tales como *¿qué tal todo?*

tener ojos en la cara «se usa en constrs como ¿no tienes ojos en la cara? [...] para comentar la incapacidad para percibir algo que salta a la vista» (p. 571), es decir, solo suele emplearse en la segunda persona. Finalmente, en lo concerniente a la fijación posicional, se entiende la predilección de empleo de algunas unidades en ciertas posiciones del discurso como, por ejemplo, la fórmula *[reciba] un cordial saludo* en la despedida de las cartas formales.

Por otro lado, en lo que respecta a la fijación interna, Thun (1978) hace referencia a la *fijación material* (imposibilidad de reordenar los elementos integrantes, manifestación fonética fija, limitación en la selección de componentes) y la *fijación de contenido* (o particularidades semánticas). Sobre la fijación interna material, podemos resumir aquí los principales rasgos que la caracterizan, con una ejemplificación con somatismos (cfr. Gross, 1984; Mendívil-Giró, 1990; Ruiz Gurillo, 1997; Zuluaga, 1980, entre otros):

- a) Inalterabilidad del orden de los componentes: *de pies a cabeza* (**de cabeza a pies*), *uña y carne* (**carne y uña*).
- b) Invariabilidad de alguna categoría gramatical: *tomar el pelo* (**tomar los pelos*), *si [alguien] levantara la cabeza...* (**si [alguien] levanta la cabeza...*).
- c) Inalterabilidad del inventario de los componentes: *en cuerpo y alma* (**en cuerpo y en alma*), *poner ojitos* (**poner ambos ojitos*).
- d) Inconmutabilidad de los elementos componentes: *meter mano* (**introducir mano*), *con la cabeza fría* (**con la cabeza helada*).
- e) Componentes no extraíbles (no pronominalizables): *¿quién dices que sentó cabeza?*
La sentó Ana.*/increíble el ojo que pegué ayer.*
- f) Imposibilidad de inserción, modificación (adjetival, adverbial...) o supresión de un componente: **dar la cara redonda*, **tomarse a pecho robusto*.

g) Fijación transformativa: imposibilidad de pasivización (**los codos fueron hincados por el estudiante*) o cambio de categoría gramatical (*cara dura > *la dureza de la cara > *endurecer la cara*).

Junto a la *fijación formal*, autores como Mendívil-Giró (2009) y Penadés Martínez (2014) inciden en la relevancia de la *fijación estructural* en la caracterización de buena parte de las unidades fraseológicas y, de forma concreta, de las locuciones:

...nisto, xustamente, radica a fización dunha locución, no feito de que estes non establecen as relación sintácticas que se dan entre os constituíntes dos grupos sintácticos libres. Unha locución presenta, logo, a propiedade da fixación estrutural, é dicir, constitúe una estrutura fixa na que os seus elementos constituíntes non desempeñan funcións específicas, co que non se poden analizar asignándolles algunha.¹⁰ (Penadés Martínez, 2014, pp. 282-283)

A este respecto, conviene señalar que ninguna de las fijaciones anteriormente descritas (formal, estructural, etc.) se presenta, a lo largo del acervo fraseológico de la lengua, de forma absoluta, sino de manera relativa o gradual. Así, encontramos UF que pueden sufrir cambio de categoría gramatical (*tomar el pelo > tomadura/tomada de pelo; meter la pata > metedura/metida de pata*), modificación adjetival/adverbial (*tener la cara [muy] dura, tener [mucha, bastante, demasiada] cara*), conmutación de sus elementos componentes (*meter la nariz/el hocico*), variación de la categoría gramatical de los elementos integrantes (*meter la nariz/las narices, meter el hocico/los hocicos*), entre muchas otras.

¹⁰ «En esto, justamente, radica la fijación de una locución, en el hecho de que estas no establecen las relaciones sintácticas que se dan entre los constituyentes de los grupos sintácticos libres. Una locución presenta, entonces, la propiedad de la fijación estructural, es decir, constituye una estructura fija en la que los elementos constituyentes no desempeñan funciones específicas, con lo que no se pueden analizar asignándoles alguna» (La traducción es nuestra)

3.4.4 *Idiomaticidad*

La fijación de una determinada secuencia puede llevar a su idiomaticidad, es decir, «aquella propiedad semántica que presentan ciertas unidades fraseológicas, por la cual el significado global de dicha unidad no es deducible del significado aislado de cada uno de sus elementos constitutivos» (Corpas Pastor, 1996, p. 26). Hace, por tanto, referencia a la no composicionalidad del significado de una UF. De forma análoga a la fijación, la idiomaticidad se presenta de forma gradual a lo largo del acervo fraseológico de una lengua. A fin de delimitar las UF (y, de forma más concreta, los somatismos) según su grado de idiomaticidad, Mellado Blanco (2004) parte de la concepción de centro y periferia de Fleischer (1983), representando el centro «más estabilidad (más restricciones y menos variantes) y más idiomaticidad (menos composicionalidad, más opacidad y menor divisibilidad)» (Mellado Blanco, 2004, p. 47). En esta misma obra, la autora menciona distintas escalas de idiomaticidad (Fleischer, 1983 y 1997; Telija, 1975; Wotjak, 1986), todas las cuales emplean como parámetro de medida la distancia existente entre el significado global de la UF y el de su homófono libre:

Si [los hablantes de nivel medio] consiguen inferir el significado fraseológico a partir de la imagen del significado recto, entonces se trataría de un FR transparente; por el contrario, si lo representado mediante esta imagen ha desaparecido por completo de la memoria colectiva, entonces estaríamos ante un FR opaco. (Mellado Blanco, 2004, p. 49).

Así, mientras que la pluriverbalidad y la fijación son características indispensables para la consideración de una unidad léxica como UF, la idiomaticidad es un rasgo facultativo adicional, cuya mayor o menor presencia hará que estas unidades se sitúen más cerca de la periferia o del núcleo prototípico de la fraseología.

3.4.5 La variabilidad en la fraseología

Aunque en principio pueda parecer contradictorio hablar de *la variabilidad de las expresiones fijas*, toda variación requiere en esencia que exista de esa forma previa estable, reconocible y bien delimitada que conforman las UF. Por todo ello, debemos considerar la fijación como una propiedad relativa de las UF:

Si en un principio las UFs se definieron como unidades estables y fijas que mostraban rechazo a cualquier alteración léxica, semántica y morfosintáctica, en vista de la existencia de cambios reales y potenciales, no cabe hablar de la fijación como una propiedad absoluta (Hausermann, 1977; Burger, 1998) sino como una cualidad relativa (Fleischer, 1982; Glaser, 1986; Wotjak, 1992; Corpas Pastor, 1996; Burger, 1998) y, nunca mejor dicho, variable. (Corpas Pastor y Mena Martínez, 2003, p. 183)

De esta manera, la variabilidad es un rasgo intrínseco a las UF no *a pesar de* sino *gracias a* su carácter estable: «Die Polilexicalität ist ein Appell an die Fragmentierung, die Fixiertheit an die Variabilität, die Figuriertheit an die Litteralisierung»¹¹ (Gréciano, 1987, p. 196). En este contexto, para Corpas Pastor y Mena Martínez (2003) la variabilidad puede presentarse de dos formas principales, como variabilidad estable (variantes y variaciones) y como variabilidad ocasional (modificaciones):

Cuando la causa de la alteración o variabilidad de una UF proviene de las motivaciones conscientes de los hablantes que persiguen de esta manera alguna finalidad, hablamos de modificaciones¹². Si, por el contrario, los cambios que se perciben en la unidad en cuestión no son fruto de las ansias innovadoras de los hablantes, sino meramente la realización de

¹¹ «La polilexicalidad es una llamada a la fragmentación; la fijación, a la variabilidad; la figuratividad, a la literalidad» (la traducción es nuestra)

¹² Para este proceso de «alteración o variabilidad de una UF según las motivaciones conscientes de los hablantes que persiguen de esta manera alguna finalidad», Zuluaga (1980) acuñó el nombre de *desautomatización de unidades fraseológicas*.

las posibilidades que el sistema lingüístico en general y el fraseológico en concreto ofrecen, estamos ante variantes institucionalizadas. (p. 184)

En los próximos apartados, profundizaremos así en los conceptos de *variabilidad estable* (que, en el presente estudio, denominaremos *variación*) y *variabilidad ocasional* (aquí designada como *manipulación*).

3.4.5.1 Variación

A lo largo del presente proyecto de tesis, emplearemos los conceptos de *variación* y *variante* para hacer referencia a aquellas alteraciones obtenidas de cambios estables e institucionalizados fruto de las opciones que ofrece el sistema lingüístico (y, concretamente, el fraseológico) (cfr. Burger et al., 1982; Corpas Pastor, 1996; Corpas Pastor y Mena Martínez, 2003; García-Page, 2008; Palm, 1995; Zuluaga, 1980, entre otros), distinguiéndolos de aquellos casos producidos ocasionalmente por la voluntad creativa del hablante (en este caso, hablaremos de *manipulación*, fenómeno que abordaremos de forma pormenorizada en el apartado 3.4.5.2).

A este respecto, Zuluaga (1975, pp. 239-241) ofrece una clasificación inicial diferenciando las «variantes en sentido estricto» de las «pseudovariantes» (o «variantes en sentido amplio»). Así empieza por delimitar cuáles a su juicio pertenecen a este último grupo de pseudovariantes:

- a) aquellas producidas por mero cambio categorial: *meter la pata* > *metedura de pata*, *quebrarse la cabeza* > *quebradero de cabeza*;
- b) las denominadas como *series idiomáticas* (Burger, en Zuluaga 1975, p. 238), dado que su significado léxico es diferente: *ans Ruder kommen* (*ofrecer la mano* ‘brindar ayuda’) - *am Ruder sein* (*dar la mano* ‘ayudar’);

- c) las distintas realizaciones en el discurso de las expresiones fijas con casillas vacías: *por mi cara bonita - por tu cara bonita - por su cara bonita...*;
- d) las expresiones fijas que, aun siendo sinónimas, no comparten una estructura o componentes similares: *tomar las de Villadiego* y *poner pies en polvorosa*, ambas sinónimas de *huir*;
- e) las variantes diatópicas¹³: *hacer vaca* (Perú) y *hacer novillos* (España);
- f) las variantes diastráticas: *me importa un chorizo*, variante popular en Colombia de *me importa un bleo*, y
- g) las variantes diafásicas: *sentidísima condolencia* y *sentido pésame*.

Así, tras la delimitación de las pseudovariantes, Zuluaga (1975, p. 241) procede a describir qué rasgos sí deben reunir las auténticas variantes (o variantes en sentido estricto):

- a) deben pertenecer a la misma lengua funcional: *tomar las de Villadiego* y *coger las de Villadiego* son variantes en España y Colombia, pero no en Argentina y Uruguay;
- b) no pueden tener significados distintos;
- c) deben ser libres, independientes de los contextos;
- d) son parciales, es decir, solo se diferencian en alguno/algunos de sus componentes: por ello *tomar las de Villadiego* y *poner pies en polvorosa* no son consideradas como variantes, pero, *tomar las de Villadiego* y *coger las de Villadiego*, sí.
- e) sufren una sustitución fija: en *meter la pata*, el componente *pata* puede únicamente ser sustituido por *gamba* y no por *pierna* (*meter la gamba*, **meter la pierna*).

¹³ Las variantes diatópicas y diafásicas, en un principio, no son reconocidas como variantes en sentido estricto por Zuluaga (1975 y 1980), pero sí posteriormente (Zuluaga, 1992).

A lo largo de nuestro proyecto de tesis, sin embargo, hemos optado por una perspectiva ancha de la variación, en la que incluimos como variantes auténticas las que, además de los rasgos anteriores, reúnan también las siguientes características:

- a) las variantes diasistemáticas: así, variantes diatópicas como *metedura de pata* – *metida de pata* o *tomadura de pelo* – *tomada de pelo* son consideradas, a nuestro criterio, como auténticas variantes. Hemos tomado esta decisión por tres razones principales. En primer lugar, y como señala Montoro del Arco (2005, p. 142), estas nominalizaciones están fijadas, no son simples inflexiones propias del discurso libre. Así, por ejemplo, no es posible decir *toma de pelo*. En segundo lugar, los propios sustantivos *metedura*, *tomadura* y *tomada* son idiomáticos, con un uso restringido a estas secuencias. La última razón se debe a que los límites entre una lengua funcional y otra son difusos. Así, como observamos en Artículo 1, el estudio en corpus realizado en esTenTen18 nos permitió comprobar que las dicotomías *metedura de pata* – *metida de pata* y *tomadura de pelo* – *tomada de pelo* coexisten con una similar frecuencia de uso en algunas variedades del español (0,23 vs. 0,25 en Méjico y 0,8 vs. 0,85 en Uruguay en el caso de *tomadura/tomada de pelo*, o 0,18 vs. 0,14 en Ecuador para *metedura/metida de pata*) pero en otras variedades diatópicas presentan un marcado contraste en la frecuencia de uso (1,43 vs. 0,01 en España; 0,51 vs. 0,01 en Ecuador, o 0,12 vs. 0,4 en Argentina para las variantes *tomadura/tomada de pelo*, y 0,84 vs. 0,09 en España, 0,06 vs. 1,43 en

Perú, o 0.06 vs. 1,58 en Guatemala, en el caso de *metedura/metida de pata*), como podemos observar en Tabla 4, extraída de Artículo 1.

	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

Tabla 4. Frecuencias normalizadas y brutas de los somatismos objeto de estudio en las 19 variedades nacionales del corpus esTenTen18

- b) los casos de *series fraseológicas*, dado que, aun siendo antonímicas las variantes, constituyen una misma unidad realizada de diversas formas según el discurso en el que se insertan: *poner buena/mala cara*, *de buena/mala gana*.
- c) cambios de categoría gramatical. Existe, a este respecto, un amplio debate en la literatura científica. Por un lado, si bien, como hemos visto anteriormente, Zuluaga (1975) en un principio no consideraba las nominalizaciones como variantes en sentido estricto, posteriormente sí (Zuluaga, 1992). García-Page (2008, p. 258) por su parte las delimita como «realizaciones discursivas de carácter eventual de las locuciones verbales, pero no variantes institucionalizadas o codificadas ni locuciones en sí mismas». Montoro del Arco (2005, p. 142) sí las define como variantes (de forma más concreta, como «variantes estructurales de orden morfosintáctico»), en la misma línea que Mellado Blanco (2004) que las delimita como «variantes conversivas» y Koike

(2007), como «variantes por relación sintagmática». Así, de forma análoga, en nuestro estudio también hemos definido estas nominalizaciones como *variantes*, dado que consideramos que son fruto de la articulación de un mismo núcleo semántico en UF que se diferencian únicamente por el cambio de uno de sus componentes. Así, al igual que *todo queda en casa* y *todo queda en familia* son consideradas variantes por conservar el mismo significado y solo presentar cambio en uno de sus componentes (en este caso el léxico, por lo que podrían denominarse como «variantes por relación paradigmática», siguiendo la terminología de Koike [2007]) *tomar el pelo* y *tomadura de pelo* poseen asimismo un núcleo semántico común en UF que presentan en este caso un cambio morfosintáctico (por lo que podrían considerarse como «variantes por relación sintagmática», [Koike, 2007]). Asimismo, en ambos binomios, el cambio es fijo: al igual que *todo queda en casa* solo puede sustituirse por *todo queda en familia* (y no **todo queda en hogar*), *tomar el pelo* tiene como únicas variantes por relación sintagmática a *tomadura/tomada de pelo* (y no **toma de pelo*). Como hemos visto anteriormente, los sustantivos *tomadura* y *tomada de pelo* son idiomáticos, con un uso restringido a estas UF.

3.4.5.2 Manipulación

Una vez abordado en el anterior apartado el fenómeno de la *variación* (o *variabilidad estable*), examinaremos aquí el fenómeno de la *manipulación* (o *variabilidad ocasional*). Con *manipulación fraseológica*, también conocida como *desautomatización* (Mena Martínez, 2003; Ruiz Gurillo, 1997; Zuluaga, 1997;) o *modificación* (Burger et al., 1982; Corpas Pastor, 1996; García-Page,

1996), se hace referencia a diferentes tipos de modificaciones intencionadas¹⁴ y fácilmente reconocibles que afectan al plano formal, semántico o discursivo de una UF, con el objetivo de conseguir determinados efectos ilocutivos o perlocutivos, tales como humor, ironía, sorpresa o persuasión, entre otros (cfr. Corpas Pastor, 2003; Mena Martínez, 2003; Omazić, 2008; Timofeeva, 2009; Zuluaga, 2001).

Lejos de ser un fenómeno excepcional, las UF que han sufrido algún tipo de manipulación se presentan de forma ubicua en el discurso:

Wotjak (1992), por ejemplo, ha calculado entre un 30 % y un 50 % para el alemán. Un porcentaje similar (40 %) deduce Corpas Pastor (1998) para las paremias en el análisis que realiza a partir de un corpus de español actual peninsular. Lo mismo se puede decir con respecto a otras lenguas como el inglés y el francés, a tenor de los resultados obtenidos por Arnaud (1993) y Moon (1998). (Corpas Pastor, 2003, p. 319)

Como indica Martí Sánchez (2015), esta ansia por desautomatizar lo ya automatizado se encuentra en la razón de ser de toda lengua, motivada por una intuición de los hablantes: «significante y significado son solidarios, de modo que toda modificación en una de las dos caras conlleva una modificación en la otra. Con ello, se reaviva la atención y se aumenta la comprensión» (Martí Sánchez, 2015, pp. 118-119). A este respecto, Philip (2008) menciona que la manipulación fraseológica se puede incluso considerar como la reafirmación del *principio de libre elección* (Sinclair, 1991) esta vez dentro las UF (la cuales *a priori* están regidas por el *principio de idiomatidad* [Sinclair, 1991]). Philip (2008) asimismo describe este tipo de modificaciones (y su activación de una doble lectura literal y figurada) como un *efecto palimpsesto*:

¹⁴ A este respecto, Martí Sánchez (2015) considera que la manipulación fraseológica también puede ser inintencional, con ejemplos como **destornillarse de risa* (por *desternillarse de risa*) o **nadar en la ambulancia* (por *nadar en la abundancia*), dado que, a su parecer, y al igual que la intencional, esta también surge de la necesidad inconsciente del hablante de motivar aquello que, en principio, parece no motivado. No obstante, en este proyecto de tesis con *manipulación* haremos únicamente referencia a aquella intencional.

Just as vellum was re-used in medieval times by over-writing the pages of old books — hiding but not erasing the original text— puns constitute a linguistic palimpsest in which the new meaning is written over the old one, but fails to cancel it out completely. (p. 104)

En lo que respecta a los tipos de manipulación, en Artículo 12 analizaremos en profundidad diferentes clasificaciones propuestas y matices a las mismas. De ellas, a lo largo del proyecto de investigación seguiremos principalmente la de Timofeeva (2009), que supone una revisión de la de Corpas Pastor (2003). Así, Timofeeva distingue, por un lado, entre *desautomatización formal*, cuando las modificaciones afectan a la estructura formal de la unidad fraseológica (coincidiría, por tanto, con la *manipulación interna* en la clasificación de Corpas Pastor); *desautomatización semántica*, cuando la manipulación surge del potencial semántico de los elementos constituyentes de la UF, y *desautomatización discursiva*, cuando la UF se inserta, de forma intencionada, en un discurso al que no pertenece, por corresponder, por ejemplo, a una variedad diafásica distinta (estos dos últimos tipos coincidirían con la *manipulación externa* en la terminología de Corpas Pastor [2003]).

Así, una vez delimitado el marco teórico del presente proyecto de tesis doctoral, en el próximo capítulo presentaremos los distintos recursos y herramientas electrónicos aplicados a la traducción de somatismos.

***CAPÍTULO IV. RECURSOS Y
HERRAMIENTAS ELECTRÓNICOS
APLICADOS A LA TRADUCCIÓN DE
SOMATISMOS SEGÚN LOS DISTINTOS
RETOS***

En este apartado nos adentramos en el terreno de la *fraseología computacional*, una disciplina de reciente consolidación (cfr. Corpas Pastor y Colson, 2020) fruto de la imbricación de los campos de la fraseología y la lingüística computacional. En lo que concierne a su objeto de estudio, para Heid (2008, p. 337) la fraseología computacional hace referencia a dos tipos principales de actividades computacionales en relación con las UF: por un lado, métodos de lingüística computacional aplicados a la fraseología (es decir, el ordenador asistiendo al humano en la descripción, [re]presentación o clasificación de las UF); y, por otro, métodos de procesamiento del lenguaje natural (PLN) para el tratamiento de las UF durante el análisis de textos (es decir, el ordenador procesando textos que incluyen UF). Así, la fraseología computacional realiza importantes aportaciones a un amplio abanico de aplicaciones de PLN tales como la recuperación y extracción de información, el análisis de sentimientos, el resumen automático de textos o la traducción automática, entre muchas otras.

En este contexto, a lo largo de los artículos que vertebran la presente tesis doctoral haremos uso de ambos métodos para el análisis del tratamiento de los somatismos por parte de los sistemas de Traducción Automática Neuronal (Artículo 1) y, a la luz de dichos resultados, el diseño e implementación del sistema de preprocesamiento de unidades fraseológicas gApp para la mejora del rendimiento de la TAN (Artículos 2-7); el desarrollo de estrategias con corpus (y otros recursos electrónicos) aplicadas a la traducción de somatismos en casos de ambigüedad (Artículo 8), variación (Artículo 9) o manipulación fraseológica (Artículos 10 y 12), y, finalmente, para la repertorización lexicográfica de estos somatismos en Glossomatic (Artículo 11). Así, pasamos a continuación a la presentación de los distintos recursos y herramientas electrónicos aplicados a la traducción de somatismos según dos de los desafíos principales que presentan estas unidades: la variación y la manipulación.

4.1 Variación

Para el caso de la variación de los somatismos, analizaremos tres recursos y herramientas documentales aplicados a la traducción: los sistemas de Traducción Automática Neuronal (apartado 4.1.1), el sistema de preprocesamiento de unidades fraseológicas gApp (apartado 4.1.2) y finalmente los corpus lingüísticos (apartado 4.1.3).

4.1.1 Traducción Automática Neuronal

La reciente aparición de los sistemas neuronales ha supuesto una auténtica revolución en el campo de la traducción automática, dando lugar a la TAN, que representa un enorme salto cualitativo en comparación con el predecesor sistema de traducción automática estadística basada en segmentos o frases (*Phrase-Based Statistical Machine Translation* o *PBSMT*) (Bojar et al., 2016; Castilho et al., 2017; Sennrich et al., 2016). Así lo señalan asimismo diversos estudios que han cuantificado esta mejora tanto con evaluación automática (Bahdanau et al., 2015; Sennrich et al., 2016; Stasimioti et al., 2020) como con humana (Shterionov et al., 2018; Wu et al., 2016), y que han observado una reducción en el esfuerzo de posesición (Bentivogli et al., 2016) y una mejora en la calidad del resultado final de la posesición (Yamada, 2019) en comparación con la traducción automática estadística. Este cambio de paradigma que ha supuesto la TAN también puede observarse fácilmente en los últimos estudios sobre el sector lingüístico europeo (European Language Industry Survey): por primera vez en 2018 más de la mitad de las empresas de servicios lingüísticos (ESC) en Europa afirma emplear la traducción automática (ELIS, 2018) y un 78 % de las ESC participantes en el estudio tienen previsto incorporar o aumentar el empleo de la traducción automática o la posesición (ELIS, 2020). En esta misma línea, todos los sectores implicados (institutos de formación, desarrolladores lingüísticos y consumidores, así como

profesionales independientes) también reconocen a la traducción automática y la posedición como la tendencia más destacada con diferencia en la industria (ELIS, 2021).

A pesar de estos avances, los sistemas TAN aún tienen un importante talón de Aquiles: las unidades fraseológicas. A lo largo de mi investigación he detectado numerosos errores en los resultados arrojados por la TAN en relación con las UF, principalmente causados por los múltiples desafíos que ellas conllevan: ambigüedad, no composicionalidad, manipulación, variación, flexibilidad morfosintáctica, discontinuidad, o anisomorfismo interlingüe, entre otros (cfr. Colson, 2019; Monti et al., 2018 y 2020; Popović, 2017; Rohanian et al., 2019, entre otros). Estos errores se deben principalmente a la propia naturaleza de estos sistemas. La TAN consiste en una arquitectura de codificador-decodificador, en la que en el codificador es una red neuronal recurrente que examina el texto origen tanto hacia adelante como hacia atrás y lo transforma en una representación vectorial. Tras esto, el decodificador consta de otra red neuronal recurrente que genera el texto meta palabra por palabra, teniendo en cuenta tanto la palabra anteriormente producida como la representación vectorial completa del texto origen, todo ello empleando un mecanismo de atención que le permite al decodificador enfocarse en distintas partes de la entrada a fin de calcular la probabilidad de la nueva palabra del TM (Casacuberta y Peris Abril, 2017). Así, la TAN es, en esencia, un sistema basado en palabras o caracteres y, por tanto, no considera a las UF como unidades de traducción, de ahí su dificultad para abordarlas (Wang et al., 2017; Zaninello y Birch, 2020). En este sentido, se han acometido diversos esfuerzos a fin de optimizar el rendimiento de la TAN ante estas unidades. Partiendo de los modelos PBSMT, que poseen un repertorio de traducciones de UF, diversos autores han presentado sistemas TAN basados en segmentos o frases (Huang et al., 2017; Tang et al., 2016) mediante la incorporación de memorias externas con UF. Métodos adicionales se han centrado en la extracción de candidatos de UF

paralelos tanto de forma automática (Fadaee et al., 2018; Rikters y Bojar, 2017) como manual (Zaninello y Birch, 2020) orientados a proporcionar datos de entrenamiento complementarios para la TAN, o en la explotación de datos monolingües a fin de incorporar un mejorado modelo de la lengua en el codificador (Sennrich et al., 2016).

Precisamente de la disponibilidad de estos datos de entrenamiento surge el segundo gran problema de la TAN ante estas unidades: pese a la relevancia de los somatismos dentro del acervo fraseológico del español, en el conjunto de la lengua su presencia es relativamente escasa, de ahí que se necesiten ingentes cantidades de datos de entrenamiento para un apropiado rendimiento de la TAN; más aún si cabe teniendo en cuenta que gran parte de estos somatismos conviven en los corpus con sus homófonos literales (*bajar los brazos, abrir la boca, abrir los ojos*, etc.), lo que incrementa aún más las dificultades para su correcta desambiguación semántica y posterior traducción. Mayores son los retos para la TAN si consideramos que buena parte de estas unidades suelen presentarse en textos informales (como el Contenido Generado por el Usuario o CGU), en los que abundan las palabras raras y fuera de vocabulario que, por tanto, no suelen formar parte de los datos de entrenamiento canónicos de la TAN (Derczynski et al., 2013; Gui et al., 2017; Neunerdt et al., 2013).

En este contexto, en Artículo 1, analizaremos el rendimiento de los dos sistemas TAN más robustos del mercado (Google Translate y DeepL) en la traducción de la forma continua y discontinua (Anastasiou, 2010) de somatismos con distintas frecuencias de aparición en los corpus, tales como *tomar el pelo* y *meter la pata*; sus variantes nominales *tomadura/metadura de pelo* y *tomada/metida de pata*, y otras variantes con menor frecuencia de uso como *meter la pata hasta el corvejón*, *meter la gamba* y *metadura/metida de gamba*.

A este respecto es necesario indicar que, salvo *tomar el pelo* y *meter la pata*, las restantes variantes presentan una frecuencia de aparición desigual a lo largo de las distintas variedades diatópicas nacionales del español, algunas de ellas (como en el caso de *meter la pata hasta el corvejón*, *meter la gamba* y *metedura/metida de gamba*) con un empleo casi exclusivo en la variedad de español europeo. Así, uno de los objetivos de la investigación será evaluar si existe una correspondencia directa entre una mayor frecuencia de uso de un determinado somatismo a lo largo de las distintas variedades nacionales y un mejor rendimiento de la TAN o si entran en juego otros factores, como determinadas variedades nacionales viéndose privilegiadas sobre otras en la traducción de estos somatismos. Otro de los objetivos del estudio consistirá en arrojar algo más de luz sobre el rendimiento de estos sistemas TAN en la traducción de posibles palabras raras y fuera de vocabulario, como es el caso de las variantes de escasa frecuencia *meter la pata hasta el corvejón*, *meter la gamba*, y sus variantes nominales *metedura de gamba* y *metida de gamba*. Todo ello nos servirá en última instancia para determinar en qué casos los sistemas TAN de Google Translate o DeepL ofrecen un mejor rendimiento ante el desafío de la variación y la discontinuidad fraseológica

Artículo 1

Hidalgo-Ternero, C. M. (2020a). Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation. En P. Mogorrón Huerta (Ed.), *Análisis multidisciplinar del fenómeno de la variación en traducción e interpretación / Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting*. *MonTI Special Issue* 6, 154-177. <https://doi.org/10.6035/MonTI.2020.ne6.5>

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.

4.1.2 *Sistemas de preprocesamiento de unidades fraseológicas para TAN*

Como hemos podido analizar en el apartado anterior, a los problemas que, por antonomasia, presentan las UF para la TAN, tales como anomalía sintáctica, no composicionalidad, variación diasistemática o ambigüedad, entre muchos otros, se suma un importante desafío: las unidades fraseológicas no siempre están compuestas por palabras contiguas (por ejemplo, *nos tomaron una vez más el pelo*), lo que supone enormes dificultades para su detección y traducción automáticas (Corpas Pastor, 2013; Constant et al., 2017; Foufi et al., 2019; Ramisch y Villavicencio, 2018; Rohanian et al., 2019). De forma más concreta, en Artículo 1, hemos podido observar que dicha discontinuidad supone un declive del 13 % en el rendimiento de DeepL y, más aún, del 15 % para Google Translate, dos de los principales sistemas TAN en la actualidad.

Con el fin de superar los retos que supone la discontinuidad incluso para los sistemas TAN más robustos (cfr. Colson, 2019; Zaninello y Birch, 2020), hemos ideado gApp, un sistema para la detección y conversión automáticas de UF discontinuas hacia sus formas continuas. A este respecto, a lo largo de los Artículos 2-7, haremos un pormenorizado análisis del estado de la cuestión en lo que concierne a otros sistemas de preprocesamiento de unidades fraseológicas ya existentes, así como describiremos tanto las diferentes etapas de detección y conversión del algoritmo ReGap como el sistema en el que se integra, gApp.

Para comprobar la eficacia de este algoritmo, se han llevado a cabo seis experimentos que presentaremos de forma pormenorizada a lo largo de las seis publicaciones que completan este apartado. En el primero (Artículo 2), se analizó el rendimiento de los dos sistemas de traducción automática neuronal más avanzados en la actualidad (Google Translate y DeepL) en la traducción (ES>EN) de 560 casos de expresiones pluriverbales en forma discontinua y continua, contrastando la conversión automática con gApp frente a aquella realizada de forma manual. Dicho experimento

arrojó resultados tan interesantes como que el algoritmo consiguió mejorar el rendimiento final de DeepL en un 10 % y el de Google Translate en un 14,6 %, con una media de 12,3 % entre ambos sistemas TAN. Asimismo, los resultados con gApp solo presentaron una diferencia de 2,7 % con respecto a la conversión manual, que, por su parte, obtuvo una mejora media del 15 %. Este rendimiento cuasianálogo se debe en gran medida al alto nivel de precisión (94,8 %) y exhaustividad (96,8 %) del algoritmo, lo que significa que pudo identificar casi todos los resultados relevantes y, al mismo tiempo, descartar la práctica totalidad de resultados irrelevantes de forma automática.

En el segundo experimento (Artículo 3) se analizó la mejora que gApp suponía para DeepL no solo hacia el inglés (ES>EN) sino también hacia el alemán (ES>DE) con 400 casos de expresiones pluriverbales en forma discontinua y continua, nuevamente contrastando la conversión con gApp frente a la manual. Los resultados arrojados muestran como gApp mejoró el rendimiento de DeepL en un 13,5 % en la direccionalidad ES>EN y en un 9 % en ES>DE, solo un 0,5 % por debajo de los resultados obtenidos con la conversión manual (14 % y 9,5 % respectivamente). Esta mejora casi idéntica entre ambas conversiones se debe nuevamente el alto nivel de precisión (95.1 %) y exhaustividad (97,5 %) de gApp.

En el tercer experimento (Artículo 4), con 400 casos con la forma discontinua y continua de somatismos verbonominales en francés, se demostró que el algoritmo ReGap consiguió ahora alcanzar unos resultados análogos a los de la conversión manual para los somatismos objeto de estudio, con una mejora en el rendimiento de DeepL del 18 % en la direccionalidad FR>EN y del 16,5 % en la direccionalidad FR>ES.

En el cuarto (Artículo 5), el sistema gApp demostró seguir teniendo un impacto decisivo en la calidad de la traducción automática neuronal de las expresiones pluriverbales, ahora para una

nueva tipología de somatismos (construcciones con la estructura de verbo + sintagma preposicional [V+SP]), nuevos sistemas TAN (ModernMT, además de DeepL y Google Translate) y adicionales direccionalidades de traducción (ES>EN/DE/FR/IT/PT), con una mejora media en el rendimiento de ModernMT en un 7,5 %, y de DeepL y Google Translate en un 6,5 %, es decir, una mejora global de 6,8 % a lo largo de los distintos sistemas TAN y las cinco diferentes direccionalidades.

En el quinto experimento (Artículo 6), se contrastó la mejora que suponía gApp para el sistema TAN de VIP, diseñado por nuestro equipo de investigación Lexytrad, con respecto a DeepL y Google Translate en la direccionalidad ES>EN. En este contexto, los resultados globales mostraron una mejora media en el rendimiento de VIP en un 21,5 %, de DeepL en un 13,5 % y de Google Translate en un 8,5 %, es decir, una mejora media global del 14,5 % a lo largo de los distintos sistemas TAN.

Finalmente, para el experimento 6 (Artículo 7), se adaptó el sistema gApp al italiano como lengua de partida, gracias a la estancia de investigación que llevé a cabo en el seno del equipo UNIOR NLP, en la Università degli Studi di Napoli L'Orientale (Nápoles, Italia) y, posteriormente, se analizó la mejora que suponía en el rendimiento de los sistemas TAN de ModernMT, Google Translate y DeepL en las direccionalidades IT>EN y IT>DE. De esta manera, pudimos observar como los resultados globales en estas nuevas combinaciones lingüísticas continuaban en la línea de los experimentos previos, con una mejora media del 11,5 % en la direccionalidad IT>EN y del 8,3 % en IT>DE, es decir, una mejora media final del 9,9 % a lo largo de los distintos sistemas TAN para ambas direccionalidades.

En Tabla 5 se describen, de forma esquematizada, los seis experimentos llevados a cabo con gApp.

	Tipología de somatismos	Número de casos	Sistema TAN	Direccionalidad	Mejora
1. Hidalgo-Ternero y Corpas Pastor, 2020	verbo-nominal	560	DeepL	ES>EN	10 %
			Google Translate		14,6 %
2. Hidalgo-Ternero y Corpas Pastor, 2022a	verbo-nominal	400	DeepL	ES>EN	13,5 %
				ES>DE	9 %
3. Hidalgo-Ternero, 2021	verbo-nominal	400	DeepL	FR>EN	18 %
				FR>ES	16,5 %
4. Hidalgo-Ternero y Corpas Pastor, 2022b	verbo + sintagma preposicional (V+SP)	300	ModernMT	ES>EN	10 %
				ES>DE	10 %
				ES>FR	10,7 %
				ES>IT	4 %
				ES>PT	2,7 %
			DeepL	ES>EN	10,7 %
				ES>DE	7,3 %
				ES>FR	4,7 %
				ES>IT	4 %
				ES>PT	6 %
			Google Translate	ES>EN	9,3 %
				ES>DE	8 %
				ES>FR	8 %
				ES>IT	0,7 %
				ES>PT	6,7 %
5. Hidalgo-Ternero y Corpas Pastor, 2022c	V+SP	400	VIP	ES>EN	21,5 %
			DeepL	ES>EN	8,5 %
			Google Translate	ES>EN	13,5 %
6. Hidalgo-Ternero, Lista y Corpas Pastor, 2022	V+SP	400	ModernMT	IT>EN	16,5 %
				IT>DE	9 %
			Google Translate	IT>EN	18 %
				IT>DE	10,5 %
			DeepL	IT>EN	0 %
				IT>DE	5,5 %
Total		2460			12,5 %

Tabla 5. Experimentos con gApp

A continuación, presentamos en mayor profundidad información adicional sobre el sistema gApp y sus resultados en la mejora de la TAN ante el desafío de la discontinuidad fraseológica a lo largo de los artículos 2-7

Artículo 2

Hidalgo-Ternero, C. M., y Corpas Pastor, G. (2020b). Bridging the ‘gApp’: improving neural machine translation systems for multiword expression detection. *Yearbook of Phraseology*, 11(1), 61–80. <https://doi.org/10.1515/phras-2020-0005>

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, concluding remarks on how to further optimise MWE-aware NMT systems will round off the chapter.

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

Resumen

En este artículo se presenta *gApp*, un sistema para la detección y conversión automáticas de UF discontinuas hacia sus formas continuas. En este contexto, se analizó el rendimiento de los dos sistemas de traducción automática neuronal más avanzados en la actualidad (Google Translate y DeepL) en la traducción (ES>EN) de 560 casos de UF en forma discontinua y continua, contrastando la conversión automática con *gApp* frente a aquella realizada de forma manual. Dicho experimento arrojó resultados tan interesantes como que el algoritmo consiguió mejorar el rendimiento final de DeepL en un 10 % y el de Google Translate en un 14,6 %, con una media de 12,3 % entre ambos sistemas TAN. Asimismo, los resultados con *gApp* solo presentaron una diferencia de 2,7 % con respecto a la conversión manual, que, por su parte, obtuvo una mejora media del 15 %. Este rendimiento cuasianálogo se debe en gran medida al alto nivel de precisión (94,8 %) y exhaustividad (96,8 %) del algoritmo, lo que significa que pudo identificar casi todos los resultados relevantes y, al mismo tiempo, descartar la práctica totalidad de resultados irrelevantes de forma automática.

Palabras clave: Sistema de preprocesamiento de texto, Traducción Automática Neuronal (TAN), expresión pluriverbal, construcciones verbonominales, discontinuidad

Artículo 3

Hidalgo Ternero C. M., y Corpas Pastor, G. (2022c, en prensa). *ReGap: a text preprocessing algorithm to enhance MWE-aware neural machine translation systems*. En J. Monti, G. Corpas Pastor y R. Mitkov (Eds.), *Recent Advances in MWU in Machine Translation and Translation*

Abstract

This research presents ReGap, a text preprocessing algorithm for the automatic token-based identification and conversion of discontinuous multiword expressions (MWEs) into their continuous form as a means to optimise neural machine translation (NMT) systems. To this purpose, an experiment with semi-fixed Verb-Noun Constructions (VNCs) is conducted so as to assess to what extent ReGap can enhance the performance of the most robust NMT system to date, DeepL, under the challenge of MWE discontinuity in the Spanish-into-English and the Spanish-into-German directionalities. In this regard, the promising results yielded for VNCs will shed some light on new avenues for enhancing MWE-aware NMT systems.

Keywords: text preprocessing algorithm, Neural Machine Translation (NMT), DeepL, token-based MWE identification, Verb-Noun Constructions (VNC), discontinuity

Resumen

Esta investigación presenta ReGap, un algoritmo de preprocesamiento de texto para la identificación y la conversión automática basada en tokens de expresiones pluriverbales discontinuas hacia sus formas continuas a fin de optimizar los sistemas de traducción automática neuronal (TAN). Para ello, llevamos a cabo un experimento con construcciones verbonominales semifijas con el fin de evaluar en qué medida puede mejorar ReGap el rendimiento del sistema TAN más robusto hasta la fecha, DeepL, ante el reto de la discontinuidad fraseológica en las direccionalidades español-inglés y español-alemán. En este sentido, los prometedores resultados obtenidos con las construcciones verbonominales arrojarán algo de luz sobre las nuevas vías para mejorar el tratamiento de estas expresiones por parte de los sistemas TAN.

Palabras clave: algoritmo de preprocesamiento de texto, Traducción Automática Neuronal (TAN), DeepL, identificación de expresiones pluriverbales basada en *tokens*, construcciones verbonominales, discontinuidad

Artículo 4

Hidalgo-Ternero C. M. (2021). El algoritmo ReGap para la mejora de la traducción automática neuronal de expresiones pluriverbales discontinuas (FR>EN/ES). En G. Corpas Pastor, M. R. Bautista Zambrana y C. M. Hidalgo-Ternero (Eds.), *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus* (pp. 253-270). Comares.

Abstract

In this paper we introduce the text-preprocessing algorithm ReGap designed for automatically detecting and converting discontinuous multiword expressions into their continuous forms so as to improve the performance of current neural machine translation systems (NMT). In order to test its effectiveness, an experiment with the most robust NMT system to date, DeepL, has been carried out for the translation of verbo-nominal somatisms in the FR>EN and FR>ES directionalities. The promising results yielded by the study for this typology of multiword expressions will open up new ways to optimise MWE-processing by NMT systems.

Keywords: text preprocessing algorithm, Neural Machine Translation (NMT), DeepL, verbo-nominal somatisms, discontinuity

Resumen

En este artículo se presenta el algoritmo de preprocesamiento de texto ReGap, diseñado para la identificación y conversión automáticas de expresiones pluriverbales desde sus formas discontinuas hacia sus formas continuas con el objetivo de mejorar el rendimiento de los actuales sistemas de traducción automática neuronal (TAN). A fin de demostrar su eficacia, se ha llevado a cabo un experimento con el sistema TAN más robusto hasta la fecha, DeepL, en la traducción de somatismos verbonominales en las direccionalidades FR>EN y FR>ES. De esta manera, los prometedores resultados obtenidos con este tipo de expresiones podrán abrir camino hacia nuevas formas de optimizar el tratamiento de las expresiones pluriverbales por parte de los sistemas TAN.

Palabras clave: algoritmo de preprocesamiento de texto, traducción automática neuronal (TAN), DeepL, somatismos verbonominales, discontinuidad

Artículo 5

Hidalgo-Ternero C. M., y Corpas Pastor, G. (2022c/en prensa). Qué se traerá *gApp* entre manos... O cómo mejorar la traducción automática neuronal de variantes somáticas (ES>EN/DE/FR/IT/PT). En M. Seghiri y M. Pérez Carrasco (Eds.), *Aproximación a la traducción especializada*. Peter Lang.

Abstract

In this paper we present research results with *gApp*, a text-preprocessing system designed for automatically detecting and converting discontinuous multiword expressions (MWEs) into their continuous forms so as to improve the performance of current neural machine translation systems (NMT) (see Hidalgo-Ternero, 2021, and Hidalgo-Ternero & Corpas Pastor, 2020 and 2021). In order to test its effectiveness, an experiment with some of most robusts NMT system to date (ModernMT, DeepL and Google Translate) has been carried out in the ES>EN/DE/FR/IT/PT directionalities for the translation of somatisms, i. e. MWEs containing lexemes referring to human or animal body parts (Mellado Blanco, 2004). More specifically, we have analysed Verb + Prepositional Phrase (VPP) constructions: the lexical variants *traer(se) entre manos*, *tener entre manos* y *llevar(se) entre manos*. The promising results yielded by the study for this typology of multiword expressions open up new ways to optimise MWE-processing by NMT systems. This is, to the best of our knowledge, the first study analysing the performance of this system for the neural machine translation of VPP multiword expressions in five different language directionalities and with three different NMT systems.

Keywords: text preprocessing system, Neural Machine Translation (NMT), Verb + Prepositional Phrase constructions, somatic variants, discontinuity

Resumen

En este artículo presentamos los resultados de una investigación con *gApp*, un sistema de preprocesamiento de texto, diseñado para la identificación y conversión automáticas de expresiones pluriverbales desde sus formas discontinuas hacia sus formas continuas con el objetivo de mejorar el rendimiento de los actuales sistemas de traducción automática neuronal (TAN) (véase Hidalgo-Ternero, 2021, Hidalgo-Ternero y Corpas Pastor, 2020 y 2021). A fin de demostrar su eficacia, se ha llevado a cabo un experimento con algunos de los sistemas TAN más robustos hasta la fecha (ModernMT, DeepL y Google Translate) en las direccionalidades ES>EN/DE/FR/IT/PT para la traducción de somatismos, es decir, expresiones pluriverbales que contienen un lexema que hace referencia a un órgano o parte del cuerpo humano o animal (Mellado Blanco, 2004). De forma más concreta, se han analizado construcciones con la estructura de verbo más sintagma preposicional [V+SP]: las variantes léxicas *traer(se) entre manos*, *tener entre manos* y *llevar(se) entre manos*. De esta manera, los prometedores resultados obtenidos con este tipo de expresiones abren el camino hacia nuevas formas de optimizar el tratamiento de las expresiones pluriverbales por parte de los sistemas TAN. Hasta donde sabemos, se trata del primer estudio que

analiza el rendimiento de este sistema para la traducción automática neuronal de expresiones con la estructura V+SP en cinco direccionalidades diferentes y con tres sistemas TAN distintos.

Palabras clave: sistema de preprocesamiento de texto, traducción automática neuronal (TAN), variantes somáticas, construcciones V+SP, discontinuidad

Artículo 6

Hidalgo-Ternero C. M., y Corpas Pastor, G. (2022d/en prensa). A la cabeza de la traducción automática neuronal asistida por *gApp*: somatismos en VIP, DeepL y Google Translate. En G. Corpas Pastor y M. Seghiri (Eds.), *Aplicaciones didácticas de las tecnologías de la interpretación*. Comares.

Abstract

In this paper we present research results with *gApp*, a text-preprocessing system designed for automatically detecting and converting discontinuous multiword expressions (MWEs) into their continuous forms so as to improve the performance of current neural machine translation systems (NMT) (see Hidalgo-Ternero, 2021, and Hidalgo-Ternero & Corpas Pastor, 2020, 2022a and 2022b). In order to test its effectiveness, an experiment with the NMT systems of VIP, DeepL and Google Translate has been carried out in the ES>EN directionality for the translation of somatismos, i. e. MWEs containing lexemes referring to human or animal body parts (Mellado Blanco, 2004). More specifically, we have analysed Verb + Prepositional Phrase (VPP) constructions, such as *agradecer de corazón*, *ver con buenos ojos*, *ir de cabeza* and *ir en cabeza*. In this regard, some of the unexpected results yielded by the study of these multiword expressions will question the widely accepted conception of phraseological discontinuity as an unequivocal synonym of worse NMT performance.

Keywords: text preprocessing system, Neural Machine Translation (NMT), VPP constructions, somatismos, phraseological discontinuity

Resumen

En este artículo presentamos los resultados de una investigación con *gApp*, un sistema de preprocesamiento de texto diseñado para la identificación y conversión automáticas de expresiones pluriverbales desde sus formas discontinuas hacia sus formas continuas con el objetivo de mejorar el rendimiento de los actuales sistemas de traducción automática neuronal (TAN) (véase Hidalgo-Ternero, 2021, Hidalgo-Ternero y Corpas Pastor, 2020, 2022a y 2022b). A fin de demostrar su eficacia, se ha llevado a cabo un experimento con los sistemas TAN de VIP, DeepL y Google Translate en la direccionalidad ES>EN para la traducción de somatismos, es decir, expresiones pluriverbales que contienen un lexema que hace referencia a un órgano o parte del cuerpo humano o animal (Mellado Blanco, 2004). De forma más concreta, se han analizado construcciones con la estructura de verbo más sintagma preposicional [V+SP], tales como *agradecer de corazón*, *ver con buenos ojos*, *ir de cabeza* e *ir en cabeza*. A este respecto, algunos de los inesperados resultados obtenidos con estas expresiones pluriverbales pondrán en cuestión la concepción ampliamente

aceptada de la discontinuidad fraseológica como sinónimo inequívoco de peor rendimiento de los sistemas TAN.

Palabras clave: sistema de preprocesamiento de texto, traducción automática neuronal (TAN), somatismos, construcciones V+SP, discontinuidad fraseológica

Artículo 7

Hidalgo-Ternero C. M., Lista, F. y Corpas Pastor, G. (2022/en revisión). *gApp-assisted NMT: how to improve the neural machine translation of discontinuous multiword expressions (IT>EN/DE)*. Language Resources and Evaluation.

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 + Prepositional Phrase (VPP) Constructions will be carried out in order to evaluate to what extent *gApp* can optimise the performance of three free open-source NMT systems —ModernMT, Google Translate and DeepL— under the challenge of MWE discontinuity in the Italian into English and Italian into German directionalities. In the light of our promising results, concluding remarks on how to further optimise *gApp*-assisted NMT will round off the chapter.

Keywords: text preprocessing system, Neural Machine Translation (NMT), Multiword Expression (MWE), Verb + Prepositional Phrase (VPP) constructions, discontinuity

Resumen

La presente investigación presenta la herramienta *gApp*, un sistema de preprocesamiento de texto basado en Python para la identificación y conversión automática de expresiones pluriverbales discontinuas hacia sus formas continuas con el fin de mejorar la traducción automática neuronal (TAN). Con este fin, llevaremos a cabo un experimento con construcciones semifijas con la estructura verbo + sintagma preposicional (V+SP) con el fin de evaluar en qué medida puede optimizar *gApp* el rendimiento de tres sistemas TAN de código abierto (ModernMT, Google Translate y DeepL) ante el desafío de la discontinuidad fraseológica en las direccionalidades italiano-inglés e italiano-alemán. A la luz de estos prometedores resultados, completarán el capítulo algunas observaciones finales sobre cómo continuar optimizando la TAN asistida por *gApp*.

Palabras clave: sistema de preprocesamiento de texto, Traducción Automática Neuronal (TAN), expresión pluriverbal, construcción de estructura verbo + sintagma preposicional (V+SP), discontinuidad

4.1.3 Corpus lingüísticos

Numerosas son las posibilidades documentales que ofrecen los corpus en la consulta de UF con vistas a su traducción, ya sea para detectar posibles equivalentes en otras lenguas (en el caso de los corpus bilingües/multilingües) o analizar el uso real de estas unidades en contexto, su rango colocacional, su paradigma flexivo, sus posibles variantes, así como la variedad diasistemática a la que pertenecen, entre muchas otras opciones (cfr. Corpas Pastor, 2008, 2015 y 2018; Fantinuoli y Zanettin, 2015; Sánchez Ramos, 2019; Recio Ariza, 2021, entre otros).

Pese a este potencial documental, la consulta de UF en corpus no está aún exenta de problemas, por la propia naturaleza de estas unidades. Así, por un lado, el carácter pluriforme de las UF dificulta enormemente la creación de patrones de búsqueda en corpus lo suficientemente exhaustivos como para poder recuperarlas en todo su esplendor discursivo, y precisos, para evitar que se disipen entre el ruido documental. Variación, flexión gramatical, discontinuidad... numerosos son los accidentes que pueden alterar la apariencia de estas unidades, por lo que una consulta exclusiva de sus formas *canónicas* en corpus no es suficiente para observar todo su potencial textual.

Al mismo tiempo resulta de especial importancia que, durante este proceso documental, el traductor sea consciente de los propios límites de la variabilidad fraseológica. Así, además de las variantes y manipulaciones, que no suponen una nueva UF, en los corpus se dan con frecuencia casos en los que ante una pequeña alteración formal sí nos encontramos ante UF completamente distintas, de ahí la importancia de elaborar patrones con la adecuada precisión y exhaustividad. Sirva de ejemplo el caso de la UF *de narices* ('extraordinario o sumamente grande', DFDEA, 2017, p. 550): será necesario redactar un patrón de consulta lo suficientemente incluyente como para abarcar asimismo todas sus variantes (*de tres/dos/veinte pares de narices y de [tres pares de]*

cojones, DFDEA, 2017, p. 181), y, al mismo tiempo, excluyente, a fin de descartar UF con similar forma pero significados completamente diferentes, como *de las narices* («Se emplea siguiendo a un n [nombre] o a un adj sustantivado para marcarlos despectivamente», DFDEA, 2017, p. 550) o *de los cojones* («Se emplea, siguiendo a un n de pers o cosa, para referirse a ellas despectivamente», DFDEA, 2017, p. 181).

A la dificultad intrínseca de esta tarea se suma un reto adicional: la eterna presión temporal a la que está sometida la labor traductora hace imprescindible reducir al máximo el esfuerzo heurístico del traductor, a fin de que pueda recuperar la información necesaria sobre la UF objeto de consulta (y sus variantes) con el menor número de búsquedas posible y sin verse *infoxicado* por un elevado ruido documental. Con este objetivo, en los dos artículos que vertebran el presente apartado (Artículos 8 y 9) se contrastarán distintos sistemas de corpus, a fin de determinar cuál ofrece un mejor rendimiento tanto en términos cualitativos como cuantitativos ante el desafío que presentan la ambigüedad y variación fraseológicas en el proceso traslaticio.

Así, por un lado, en Artículo 8 se presenta una propuesta didáctica en la que se les enseñó a los traductores en formación una serie de estrategias heurísticas para la consulta de UF en corpus con vistas a su traducción, poniendo especial atención a las secuencias fraseológicas susceptibles de presentar múltiples lecturas tales como *[ir] a pelo* vs. *ir al pelo*, o *de narices* vs. *de las narices*, entre otras. En este contexto, a lo largo de la investigación se analizan dos corpus monolingües del español (CORPES XXI y eseuTenTen), un corpus monolingüe de inglés (enTenTen20) y dos paralelos bilingües (Europarl7 y Linguee, subcorpus inglés-español). Los corpus eseuTenTen, enTenten y Europarl7 se encuentran disponibles dentro del sistema Sketch Engine, que, como podremos comprobar en Artículo 8, es el sistema que presenta un mejor rendimiento en la consulta de UF con similar forma pero distinto significado.

Por su parte, en Artículo 9 se examinan las posibilidades que ofrece la interfaz de búsqueda de los corpus creados por la RAE (el CREA, en sus versiones tradicional y anotada, y el CORPES XXI), el Corpus del Español y Sketch Engine para la recuperación de UF en todas sus manifestaciones formales (variación léxica, flexión gramatical, discontinuidad...). De forma concreta, los somatismos objeto de estudio están conformados por las variantes *tener entre manos*, *traer entre manos* y *llevar entre manos*, así como *ir al pelo* y *venir al pelo*. En este contexto, los resultados obtenidos por el presente estudio han permitido vislumbrar como Sketch Engine es el corpus que nuevamente ofrece un mejor rendimiento tanto en términos cuantitativos como cualitativos en esta ardua labor heurística de obtener el mayor (y mejor) número de resultados con el menor número de búsquedas posible.

Artículo 8

Hidalgo Ternero, C. M., y Corpas Pastor, G. (2020a). Estrategias heurísticas con corpus para la enseñanza de la fraseología orientada a la traducción. En M. Seghiri (Ed.), *La lingüística de corpus aplicada al desarrollo de la competencia tecnológica en los estudios de traducción e interpretación y la enseñanza de segundas lenguas* (pp. 181-204). Peter Lang. <https://doi.org/10.3726/b14734>

Abstract

This work presents a didactic proposal carried out in the subject *Lengua y cultura “B” aplicadas a la Traducción e Interpretación (II) – inglés*, taught in the first year of the Bachelor’s Degree in Translation and Interpreting, at the University of Malaga. The main objective of this proposal is to teach the possibilities that both monolingual and bilingual corpora can provide for the correct identification and interpretation of phraseological units with regard to their translation, paying special attention to those cases where the ambiguity of phraseological sequences may lead to multiple interpretations. We will focus on *somatisms* and will mainly use two Spanish monolingual corpora (CORPES XXI and esEuTenTen), an English monolingual corpus (enTenTen20) and two parallel corpora (Europarl7 and Linguee, more specifically its English-Spanish subcorpus). Against this background, this proposal is divided into several learning activities. After a first seminar where the concepts of *corpus*, *phraseology* and *translation* are introduced, in the learning activity 2 we will use parallel corpora to find translation pairings that contain translation mistakes caused by problems with phraseological ambiguity. Then, in the third learning activity, we will teach some disambiguating elements that will facilitate a correct identification and interpretation of the phraseological unit, in order to be able to convey its pragmatic and semantic weight in the target text. It is in this step where corpora can play a decisive role as documentation tools. Nevertheless, the localisation and interpretation of phraseological units is not problem-free. Given the necessity to develop some techniques that will enable a more effective detection of phraseological units, in the fourth learning activity students will learn an array of heuristic strategies to refine their searches in the consulted corpora as well as to select adequate equivalences after a correct interpretation of the results produced by these corpora.

Keywords: didactic proposal, phraseological unit, corpus linguistic, equivalence and somatism

Resumen

En este trabajo se presenta una propuesta didáctica llevada a cabo en la asignatura *Lengua y cultura “B” aplicadas a la Traducción e Interpretación (II) – inglés*, impartida en el primer año del Grado en Traducción e Interpretación de la Universidad de Málaga. El principal objetivo de esta propuesta didáctica consistirá en enseñar las posibilidades que presentan los corpus (monolingües y bilingües) para la identificación y correcta interpretación de las unidades fraseológicas (UF) de cara a la traducción, poniendo especial atención a las secuencias fraseológicas susceptibles de presentar múltiples lecturas. Nos vamos a ceñir al caso de los somatismos y nos centraremos concretamente en dos corpus monolingües del español (CORPES XXI y esEuTenTen), un corpus monolingüe en inglés (enTenTen20) y dos paralelos multilingües (Europarl7 y Linguee, subcorpus inglés-español). Con este telón de fondo, la propuesta queda dividida en varias actividades

formativas. Tras un primer seminario de introducción a los conceptos de *corpus*, *fraseología* y *traducción*, en las actividades formativas 2 y 3 localizaremos en los corpus paralelos binomios textuales que presentan errores de traducción debido a problemas de ambigüedad fraseológica y mostraremos una serie de elementos desambiguadores que permitan una correcta identificación e interpretación de la UF, a fin de poder reproducir en el texto meta la carga semántico-pragmática de la UF. Es en este decisivo paso donde desempeña un papel fundamental el corpus como herramienta de documentación. Ahora bien, la localización e interpretación de UF en los corpus no está exenta de problemas. Por ello, y dada la necesidad desarrollar mecanismos que permitan detectar UF de una manera más efectiva, en la última actividad formativa, los estudiantes aprenderán una serie de estrategias heurísticas para refinar sus búsquedas en los corpus consultados así como para seleccionar equivalencias adecuadas tras interpretar correctamente los resultados arrojados en dichos corpus.

Palabras clave: propuesta didáctica, unidad fraseológica, lingüística de corpus, equivalencia, somatismo.

Artículo 9

Hidalgo-Tenero, C. M., & Corpas Pastor, G. (2021a). La variación fraseológica: análisis del rendimiento de los corpus monolingües como recursos de traducción. *Études romanes de Brno*, 42(1), 359-379. <https://doi.org/10.5817/ERB2021-1-19>

Abstract

Idioms tend to vary significantly in discourse (variation, grammatical inflection, discontinuity...). This makes it especially difficult to create appropriate query patterns that obtain these units in all shapes and forms while avoiding excessive noise. In this context, this paper analyses the performance of different corpus management systems available for Spanish when searching phraseological variants such as *tener entre manos*, *traer entre manos* and *llevar entre manos*, as well as *ir al pelo* and *venir al pelo*. More specifically, we will examine two corpora created by the Real Academia Española (CREA, in its original and annotated version, and CORPES XXI), the Corpus del Español, and Sketch Engine. The results of our study will shed some light on which corpus management system can offer a better performance for translators under the challenge of idiom variation.

Keywords: idiom variation, discontinuity, query patterns, corpus management systems

Resumen

Las múltiples manifestaciones con las que se pueden presentar las unidades fraseológicas en el discurso (variación, flexión gramatical, discontinuidad...) hacen especialmente compleja la creación de patrones de búsqueda apropiados que permitan recuperarlas en todo su esplendor discursivo sin que ello implique un excesivo ruido documental. En este contexto, a lo largo del presente estudio se analiza el rendimiento de diferentes sistemas de gestión de corpus disponibles para el español en la consulta de las variantes fraseológicas *tener entre manos*, *traer entre manos* y *llevar entre manos*, e *ir al pelo* y *venir al pelo*. De forma concreta, se someterán a examen dos corpus creados por la RAE (el CREA, en sus versiones tradicional y anotada, y el CORPES XXI), el Corpus del Español y Sketch Engine. Los resultados arrojados por este análisis permitirán vislumbrar qué sistema de gestión de corpus ofrece un mejor rendimiento para los traductores ante el desafío de la variación fraseológica.

Palabras clave: variación fraseológica, discontinuidad, patrones de búsqueda, sistemas de gestión de corpus

4.2 Manipulación fraseológica

En este apartado, presentaremos diferentes recursos y herramientas documentales para la traducción de somatismos manipulados en el TO, tales como la funcionalidad Word Sketch, de Sketch Engine (subapartado 4.2.1) o el glosario de somatismos Glossomatic (4.2.2). En 4.2.3, además de los corpus lingüísticos, los recursos lexicográficos y los traductores automáticos, se analizarán asimismo dos herramientas documentales adicionales: los motores de búsqueda y los *rimadores*.

4.2.1 Word Sketch

Si en el anterior capítulo (4.1.) analizamos el rendimiento de distintos sistemas de gestión de corpus ante el desafío de la ambigüedad y la variación fraseológica (siendo Sketch Engine el que ofrecía mejores resultados tanto en términos cuantitativos como cualitativos), en Artículo 10 examinaremos la utilidad de una de las funcionalidades que ofrece este sistema (Word Sketch) en aquellos casos en los que exista manipulación fraseológica en el texto origen y el traductor, ante la ausencia de una correspondencia primaria en la lengua meta que pueda recrear una manipulación análoga, decida crear un equivalente fraseológico *ad hoc* en el texto meta a partir de un determinado lema. Con el fin de evaluar su eficacia, se elaboró una propuesta didáctica en la que los estudiantes, en primer lugar, recibieron un seminario teórico sobre los conceptos de manipulación fraseológica, corpus y traducción, así como aprendieron las distintas funcionalidades de Word Sketch en su aplicación a la traducción de UF manipuladas. A continuación, tuvieron que enfrentarse a distintos escenarios con UF manipuladas en el TO, en los que se incluyeron somatónimos con distinto nivel de productividad¹⁵ (*head* [cabeza], *mind* [mente] y *back* [espalda]),

¹⁵ Sobre la productividad de los distintos somatónimos en la lengua española, Olza Moreno (2011) señala que «a partir de los datos extraídos del DFDEA, por ejemplo se puede decir que las bases somáticas más productivas del español son *mano* (145 UF registradas bajo esta voz), *pie(s)* (135 UFS), *ojo(s)* (91 UFS), *cabeza* (74 UFS), *boca* (70 UFS) o

a fin de determinar si para los distintos casos los estudiantes eran capaces de ofrecer una equivalente somático manipulado para el TO mediante el uso de Word Sketch. En otros TO, se ofreció un mismo somatónimo (*back*) acompañado por distintas imágenes, con el objetivo de evaluar si la creación de equivalentes fraseológicos *ad hoc* venía mayormente determinada por el texto origen o la imagen acompañante, o ambos. Así, a la luz de estos reveladores resultados, en Artículo 10 podremos comprobar en qué medida Word Sketch pudo asistir a los traductores en formación ante el reto de buscar equivalentes para los somatismos manipulados del TO.

cara (70 UFS)» (p. 104). Díaz (en Mellado Blanco, 2004, p. 29) menciona asimismo un similar orden de productividad de los lexemas somáticos: *mano, ojo, cabeza/pie, alma*.

Artículo 10

Hidalgo-Ternero, C. M. (2019a). Translating manipulated idioms (EN>ES) in the Word Sketch scenario. En G. Corpas Pastor y R. Mitkov (Eds.), *Computational and Corpus-Based Phraseology* (pp. 173-186). Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-30135-4_13

Abstract

This paper presents a teaching proposal for translation/interpreting undergraduate students. Following a corpus-based methodology, the main objective of this proposal is to train students on how to exploit the possibilities offered by Sketch Engine for the creation of *ad hoc* phraseological equivalences. More precisely, we will examine the potential of its functionality Word Sketch as a convenient tool in those cases where the manipulation of idioms and the absence of one-to-one phraseological correspondences may pose problems to translation. The experimental setting comprised an introductory seminar on the convergence of corpora, phraseology and translation, followed by a hands-on session, where trainee translators were presented with some case-studies for which they had to create *ad hoc* phraseological equivalents for the manipulated idioms in the source text (ST) and had to justify both the translation *process* and *product*. Overall, the insights gained from analysing the results obtained will allow us to determine to what extent trainee translators are able to translate manipulated idioms in the Word Sketch scenario.

Keywords: teaching proposal · translation · Word Sketch · idiom manipulation · *ad hoc* phraseological equivalences

Resumen

En este artículo, examinamos la utilidad de una de las funcionalidades que ofrece Sketch Engine (Word Sketch) en aquellos casos en los que exista manipulación fraseológica en el texto origen y el traductor, ante la ausencia de una correspondencia primaria en la lengua meta que pueda recrear una manipulación análoga, decida crear un equivalente fraseológico *ad hoc* en el texto meta, a partir de un determinado lema. Con el fin de evaluar su eficacia, se elaboró una propuesta didáctica en la que los estudiantes, en primer lugar, recibieron un seminario teórico sobre los conceptos de manipulación fraseológica, corpus y traducción así como aprendieron las distintas funcionalidades de Word Sketch en su aplicación a la traducción de UF manipuladas. A continuación, tuvieron que enfrentarse a distintos escenarios con UF manipuladas en el TO, en los que se incluyeron somatónimos con distinto nivel de productividad (*head* [cabeza], *mind* [mente] y *back* [espalda]), a fin de determinar si para los distintos casos los estudiantes eran capaces de ofrecer una equivalente somático manipulado para el TO mediante el uso de Word Sketch. En otros TO, se ofreció un mismo somatónimo (*back*) acompañado por distintas imágenes, con el objetivo de evaluar si la creación de equivalentes fraseológicos *ad hoc* venía mayormente determinada por el texto origen o la imagen acompañante, o ambos.

Palabras clave: propuesta didáctica, traducción, Word Sketch, manipulación fraseológica, equivalencias fraseológicas *ad hoc*

4.2.2 Glosarios

En este apartado, presentaremos las posibilidades que ofrece el glosario de somatismos basado en corpus denominado Glossomatic, que diseñamos para el establecimiento de equivalentes fraseológicos *ad hoc* en aquellos casos en los que la manipulación fraseológica en el texto origen unida a la ausencia de correspondencias biunívocas interlingües pudo plantear problemas para la labor traductora. Así, en un primer momento, haremos un breve repaso por el estado de la cuestión con otros recursos lexicográficos de somatismos ya existentes, así como introduciremos aquellos recursos y obras que han servido de referencia en el diseño de Glossomatic.

En este punto, nos adentramos así en el campo de la fraseografía, definida de la siguiente manera por una de las principales referencias en el ámbito fraseográfico para el español, la doctora Olímpio de Oliveira Silva (2007).

La fraseografía es una disciplina lingüística que se ocupa, por una parte, de los principios teóricos y prácticos que rigen la inclusión de la fraseología en compilaciones léxicas (diccionarios, léxicos, vocabularios, glosarios, concordancias, etc.), tanto restringidas como generales, y, por otra, del estudio crítico y descriptivo de estas compilaciones, en lo que al tratamiento de la fraseología se refiere, lo que significa decir que el ámbito de interés de la fraseografía comprende desde la presentación tipográfica seguida en la obra hasta la adecuación a los usuarios. (p. 27)

A lo largo de los próximos subapartados, nos centraremos, de forma más concreta, en la *fraseografía práctica*, que hace referencia a «la actividad de elaboración de los diccionarios fraseológicos [así como] el establecimiento de la técnica o la metodología de elaboración de los diccionarios fraseológicos» (Olímpio de Oliveira Silva, 2007, p. 28).

4.2.2.1 Estado de la cuestión

En la literatura ya es posible encontrar algunas propuestas de recursos lexicográficos que se centran de forma exclusiva en el estudio de los somatismos y que incluyen el español entre las lenguas abordadas. Una de ellas es la de Corrêa Rocha (2014), quien, en *A elaboração de um repertório semibilingüe de somatismos fraseológicos do português brasileiro para aprendizes argentinos*, presenta un diccionario semibilingüe de somatismos del portugués (variedad diatópica de Brasil) con equivalencias en español (variedad diatópica de Argentina). Para cada entrada, se ofrece la siguiente información lexicográfica:

Expresión idiomática: + definición lingüístico-conceptual perifrástica; definición lingüístico conceptual sinonímica. @ Ejemplo. # Información lexicultural. = Equivalente en español de Argentina. (Corrêa Rocha, 2014, p. 139)

Ejemplo:

Perder a cabeça: + Perder o juízo, descontrolar-se; **perder a estribeira.** @ A pessoa errada te faz perder a cabeça, perder a hora, morrer de amor (VERÍSSIMO, 2013). # Relacionam-se à palavra cabeça os conceitos de juízo, prudência, inteligência, raciocínio e imaginação, 144 posturas que inexistem na conduta daquele que perde a cabeça (CHEVALIER, 2001). Sob o ponto de vista da analogia, à cabeça associam-se o bom senso, o entendimento, a razão e a racionalidade, de modo que, aquele que perde a cabeça, age desprovido de tais aspectos (AZEVEDO, 1983). = **Perder la cabeza. Perder la chaveta.** (Corrêa Rocha, 2014, pp. 142-143)

Así, siguiendo un enfoque semasiológico, es posible consultar cada somatismo ordenado por orden alfabético según el primer somatónimo que incluya (en el ejemplo, *cabeza*). Para cada entrada, se ofrece una perífrasis explicativa del significado de la UF, un sinónimo (si lo hay), un ejemplo de uso (extraído de la web), información lexicocultural (con datos sobre etimología popular, metáfora

conceptual de la que parte el somatismo, etc.) y equivalente(s) en el plano léxico en el español de Argentina.

Baran à Nkoum (2015), en *La cabeza en las locuciones verbales españolas. Locuciones somáticas y correspondencias francesas*, ofrece nuevamente un modelo de diccionario semasiológico, en el que, para cada entrada, clasificada según el somatónimo, se ofrecen las distintas posibles acepciones de la UF, ejemplos de uso extraídos de diferentes obras lexicográficas y sinónimos. Para cada una de estas acepciones se ofrecen equivalentes en francés, también con ejemplos de uso y sinónimos. Un ejemplo de la entrada «perder la cabeza» puede observarse en la siguiente Figura 1.

95	PERDER LA CABEZA
(ES)	(FR)
<i>Perder algo la cabeza</i>	<i>Perdre qqn la tête</i>
[1] - Perder la serenidad y el control por un ataque de miedo, de cólera, de pasión, etc. (DUE).	[1] - Perdre son sang-froid, s'affoler (TLF).
«Todos sabemos que has sufrido mucho, pero no debes perder la cabeza, sino trabajar mucho y olvidarlo poco a poco» (DELE).	«Julien ne remarqua pas cette nuance, ce tutoiement lui fit perdre la tête, ou du moins des soupçons s'évanouirent; il osa serrer dans ses bras cette fille si belle, et qui lui inspirait tant de respect» (DEL).
[2] - Dejar de comportarse con juicio, volverse loco (DLVEE). Actuar una persona irreflexivamente, sin razonar (DSLE).	[2] - Être troublé, ne plus avoir tout son bon sens; devenir fou, ne plus avoir de cohérence dans son comportement (TLF). Ne pas avoir toutes ses facultés mentales (DCF).
«A raíz de la tragedia perdió la cabeza y tuvieron que ingresarlo en un manicomio» (DELE).	«Libre à toi de courir tes risques. Mais lier un autre être à ta destinée, en un moment pareil? c'est monstrueux, allons! Tu as totalement perdu la tête! Tu as cédé à un enfantillage qui ne supporte pas une minute l'examen» (DEF).
Sin.: perder algo el juicio (DLVEE). Descomponerse la cabeza (DRAE, DBLLE). Ofuscarse algo la cabeza a algo (DICC). Embolarse algo la cabeza (DUE). Irse/ volverse algo la cabeza (DBLLE).	Sin.: perdre son la raison (DEL). Avoir qqn la cervelle filé/démené/timbé, avoir perdu son la raison (DEF). Avoir son la tête filée (DEL). Être tombé(e) qqn sur la tête (DEL, L). Devenir qqn fou, perdre qqn la raison, tomber qqn en déraison (DBLLE).
[3] - Desmayarse o perder el conocimiento (DFDEA).	[3] - S'évanouir, perdre connaissance (L).
«Un religioso... le dio la extremaunción por los pasillos. Luciano no perdió la cabeza, dio cuenta del grupo sanguíneo a que pertenecía y mientras se desangraba chorros dijo que no era alcohólico» (DFDEA).	«Étourdi, Franck ne parvenait plus à se concentrer. Il comprenait maintenant ce qui avait empêché Miro de déclencher sa bombe. Cette musique lui avait donné le vertige et le prof avait perdu la tête avant d'avoir pu actionner le détonateur» (Gaëtan Picard, 2015, <i>Le crime de la foca échouée</i> , Montréal, Québec: NUM Éditeur).
[4] - [por algo] Estar ciegamente enamorado (DFEM).	[4] - [pour qqn]: tomber éperdument amoureux de cette personne (Expresso.fr).
«Ha perdido la cabeza por una antigua compañera de clase y ahora casi nunca sale con nosotros» (DELE).	Il prétend que Julie lui fait perdre la tête: elle se fait tant servir qu'il en deviendra fou» (DEL).

Figura 1. Ejemplo de la entrada perder la cabeza

Por su parte, Rayyan (2014), en su proyecto de tesis *Fraseología y lingüística informatizada: elaboración de una base de datos electrónica contrastiva árabe-español/español-árabe de*

fraseologismos basados en partes del cuerpo, presenta cuatro tipos de diccionarios somáticos: un *Diccionario de Equivalencias* (en el que incluye la *equivalencia total* y *parcial* según la terminología de Corpas Pastor, 2003), un *Diccionario de Uso Incorrecto* (en el que incluye la *equivalencia aparente*), un *Diccionario Temático de los Somatismos* (ordenado según el dominio meta y el subdominio meta) y, finalmente, un *Diccionario Axiológico de los Somatismos* (según el valor connotativo de los somatismos). En Figura 2, mostramos un ejemplo de entrada para «perder la cabeza» en el *Diccionario Temático de Somatismos*, en el que se incluye asimismo información sobre el tipo de equivalencia. Siguiendo la terminología del autor, este caso es considerado «equivalencia total (B)» dado que ambas UF coinciden en significado (tanto connotativo como denotativo) así como en el registro, pero el somatónimo es distinto (en español es *cabeza* y en árabe es *mente*). En Figura 3, mostramos asimismo un ejemplo de diccionario axiológico para distintas UF con los somatónimos «ala» y «alma».

SUBDOMINIO META:		EXPRESIÓN DE LAS EMOCIONES		
PARTE DEL CUERPO :		CABEZA		
EQUIVALENCIA EN EL CASTELLANO	ÁRABE	TRANSCRIPCIÓN	SIGNIFICADO (ARB)	TIPO DE EQUIVALENCIA
Perder la cabeza	فقد (فلان) عقله	Fāqādā (fūlan) ‘āqlāh	Para mostrar la locura y el desequilibrio de la mente de alguien	TOTAL (B)
Perder la cabeza	مور يعقله	mw bī ‘āqlīh	Perder alguien el control y no actuar con normalidad	TOTAL (B)

Figura 2. Fragmento del diccionario temático de somatismos

DICCIONARIO AXIOLÓGICO ÁRABE - ESPAÑOL

VALOR:		NEGATIVO	
PARTE DEL CUERPO:		(ALA) جناح	
ÁRABE	TRADUCCIÓN LITERAL	SIGNIFICADO	EQUIVALENCIA EN EL CASTELLANO
مكسور (مقصوع) الجناح	Roto el ala	Persona incapaz de hacer algo Desdichado y débil	Con las alas rotas
PARTE DEL CUERPO:		(ALMA) روح	
ÁRABE	TRADUCCIÓN LITERAL	SIGNIFICADO	EQUIVALENCIA EN EL CASTELLANO
(أمر) لا روح فيه	(Algo) No tiene alma.	Se usa para mostrar la ineficacia de algo y para denotar a la falta de importancia y la inutilidad	Sin alma (no confundir con desalmado)
أسلم (فلان) الروح	Entregó (fulano) el alma	Morir	Entregar el alma a Dios

Figura 3. Fragmento del diccionario axiológico árabe-español

En lo que respecta a otras bases de datos fraseológicas generales (no exclusivas de somatismos) que han servido de referencia para Glossomatic, podemos destacar, por un lado, la aplicación creada por el grupo de investigación FRASYTRAM¹⁶, en las que se incluyen *construcciones verbales fijas* (CVF) (Mogorrón, 2008 y 2010) tales como locuciones verbales, colocaciones verbales, verbos soporte y construcciones verbales comparativas. Esta base de datos incluye expresiones en 10 lenguas distintas: español (8642 UF), catalán (1610 UF), francés (2704 UF), italiano (470 UF), árabe (449 UF), polaco (469 UF), ruso (330 UF), inglés (1585 UF), alemán (1879 UF) y chino (2 UF). A fin de analizar la información lexicográfica recogida en cada entrada, podemos observar la Figura 4 con la UF *perder la cabeza*.

¹⁶ La aplicación de la base de datos multilingüe está disponible en la siguiente dirección URL: <http://84.127.230.137:6263/phrasology/>

Expresión:

perder la cabeza

Fuente de la expresión: DEUEM

Estructura completa de la expresión:

perder (alguien) la cabeza

Variantes:

perder alguien [el juicio, el seso, la cabeza, la chaveta, la razón]

Definición de la expresión:

Perder la facultad de razonar o dejar de comportarse con cordura

Contexto de la expresión:

Ante tanta novedad positiva, el consumidor no debe perder la cabeza.

Tipo de fuente del contexto: magazine

Autor de fuente del contexto: Prensa

Título de fuente del contexto: Revista Nutrición XXI, nº 9, 01-02/2003 : Desde Alemania

Año de fuente del contexto: 2003

Web de fuente del contexto: <http://corpus.rae.es/cgi-bin/crpsrvEx.dll?visualizar?tipo1=5&tipo2=0&initem=0&ordenar1=0&ordenar2=0&FID=040115020C000004012015203011561.1184.1180&desc={B}+>

[{I}+perder+la+cabeza{I},+en+todos+los+medios,+en+{I}CREA+{I}+{B}{BR}&marcas=0](#)

Niveles de uso: Popular/Familiar/Coloquial

Marcas dialectales: General

Frecuencia de uso:Frecuente

Figura 4. Ejemplo de entrada para la UF perder la cabeza

Así, es posible encontrar la UF *perder la cabeza* dentro del campo semántico «Salud-vida-muerte >> salud mental >> perder la facultad de razonar o dejar de comportarse con cordura». Dentro de esta entrada, es posible consultar diferente información lexicográfica tal como *fuerza de la expresión, estructura completa de la UF* (incluyendo aquí los posibles actantes), *variantes, definición, contexto* (con información sobre el mismo), *nivel de uso* (popular/familiar/coloquial), *variedad diatópica y frecuencia de uso*. Asimismo, en este subcampo «perder la facultad de razonar o dejar de comportarse con cordura», es posible consultar todas las UF parasinónimas (*cruzarse los cables, perder la razón*, etc.) en español, y equivalentes en catalán (*perdre el seny, beure's l'enteniment*, etc.), inglés (*not to be in one's right mind, to fly off the handle...*), y francés (*[ne pas] avoir toute sa tête, perdre le nord, péter un câble*, etc.).

Otra obra de referencia fue el Diccionario de Locuciones Idiomáticas del Español Actual (DiLEA), diseñado por Penadés Martínez (2019), en el que se incluyen principalmente locuciones verbales. En él, para cada entrada se ofrece distinta información lexicográfica relevante tal y como *categoría* (en el caso de *perder la cabeza* es una locución verbal intransitiva), *marcación diafásica*, *frecuencia de uso*, *combinatoria de la UF* («[alguien] pierde la cabeza»), *definición*, *ejemplos y corpus con ejemplos*, como podemos comprobar en Figura 5.

The screenshot shows the DiLEA dictionary interface. At the top, there is a navigation bar with links: Presentación, Instrucciones, Diccionario, Fuentes, and Enviar comentarios. On the left, a sidebar contains a search bar with 'perder la cabeza' and a list of entries under the letter 'P', including 'perder la cabeza'. The main content area displays the entry for 'perder la cabeza' with the following details:

- 1.** (Numbered entry)
- Categoría:** intr.
- Marcación:** infor.
- Frecuencia:** + f.
- Combinatoria:** [alguen]
- Definición:** Perder el juicio o la cordura. 🔍
- Ejemplos:**

Ya sabes cómo se ponen al ver un uniforme... – Como locas [...] – Exacto –y él sonrió a su vez–, es como si **perdieran la cabeza**.

cuando se enamoraba, estaba comprobado que **perdía la cabeza**, como les pasaba a los hombres con ella.

La clave del partido, según el técnico blanquiazul, fue no **perder la cabeza** en ningún momento y mantener la serenidad a pesar de que los minutos pasaban y las ocasiones se desperdiciaban.
- Corpus de ejemplos:** Información aún no disponible.

Figura 5. Ejemplo de la entrada perder la cabeza en el diccionario DiLEA

También sirvió como base PHRASEOTEXT (González Rey, 2017), un diccionario fraseodidáctico monolingüe del francés con UF extraídas de un corpus literario de 18 autores contemporáneos (siglos XX y XXI) (González Rey, 2017, p. 33). En lo que respecta a la microestructura del diccionario, como podemos observar en Figura 6, PHRASEOTEXT ofrece diferente información lexicográfica para cada entrada, tal como *UF*, *contexto*, *definición*, *palabra clave*, *concepto*, *autor* del texto, *obra* en la que aparece y *nivel* del MCERL.

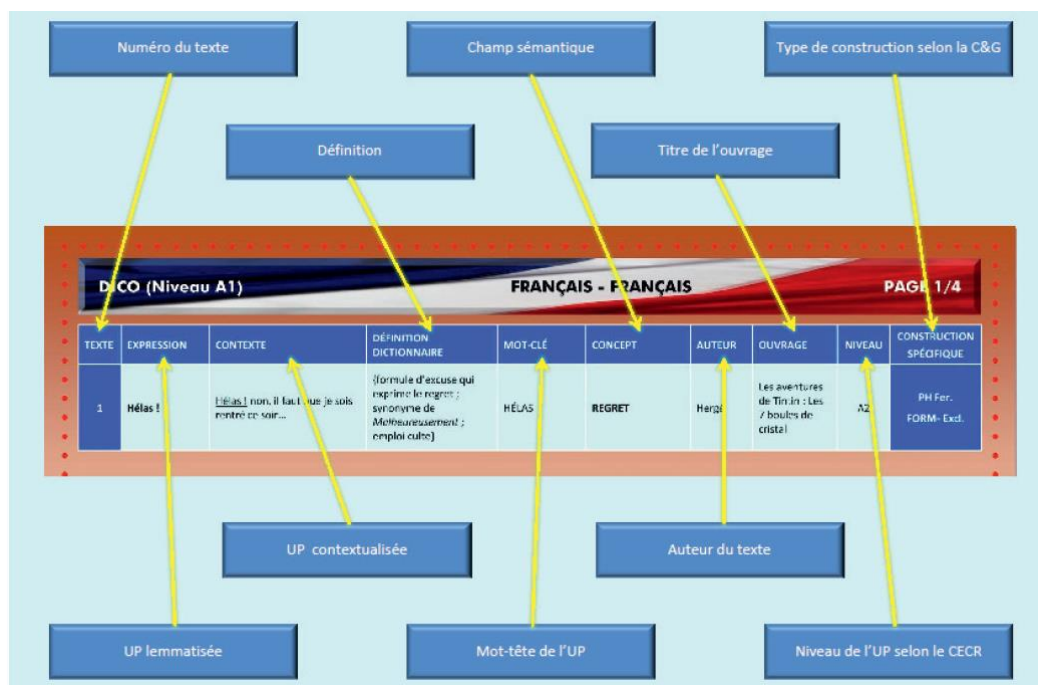


Figura 6. Ejemplo de entrada hélas en el diccionario PHRASEOTEXT

Finalmente, en lo que respecta a obras que abordan, desde un punto de vista teórico, los criterios de diseño para una base de datos fraseológica, podemos destacar el capítulo titulado *Diseño de una base de datos fraseológica para la traducción asistida por ordenador (TAO)* (Corpas Pastor, 2003). En él, la autora establece los elementos de información que debe tener una base de datos fraseológica para la TAO: [1] idioma, [2] tipo de unidad, [3] neologismo (sí/no), [4] Identificación de la unidad, [5] Características semánticas, [6] Características sintácticas, [7] Restricciones estilísticas y expresivas, [8] Características pragmáticas y discursivas, [9] Marcos de situación (para fórmulas), [10] Manipulaciones, [11] Unidades pluriverbales relacionadas, [12] Unidades monoverbales relacionadas, y [13] Equivalentes de traducción (Corpas Pastor, 2003, pp. 194-196). Asimismo, la autora, en lo concerniente al diseño de la base de datos fraseológica, recomienda el modelo relacional ya que «ofrece las características necesarias para un correcto modelado que garantice niveles adecuados en cuanto a información redundante, y reutilización y recuperación de información» (Corpas Pastor, 2003, p. 197). Este será, por tanto, el modelo que

emplearemos en el diseño de Glossomatic, que analizaremos de forma pormenorizada en el siguiente apartado.

4.2.2.2 Glossomatic

El glosario Glossomatic, diseñado en una base de datos relacional en Microsoft Access 2016¹⁷, sigue un doble enfoque onomasiológico y semasiológico, debido a la propia naturaleza del mismo: dado que Glossomatic está concebido para la creación de equivalentes fraseológicos *ad hoc* en aquellos casos de manipulación fraseológica en el texto origen, es imprescindible que no solo se pueda consultar la información necesaria sobre una determinada UF y sus equivalencias en otras lenguas, sino que también sea posible obtener UF a partir de una serie de parámetros (*registro, campo semántico, características fonoestilísticas, etc.*).

Así, para cada entrada, además del equivalente en inglés y alemán, se ofrece información microestructural relevante tal como *tipo de somatismo* (adjetival, verbal, nominal, preposicional o verbal), *nociones*, *colocados*, *registro* (formal, neutral, informal o vulgar), *polaridad* (positiva, neutra o negativa) y *diferencias interlingües*. En el campo de *noción*, se incluyen los distintos conceptos asociados a una UF (*enfado, engaño, amor, etc.*), lo que permite una fácil recuperación de UF que comparten un mismo campo semántico. Por ejemplo, bajo la noción *pensamiento* encontramos UF como *rondar por la cabeza* y sus correspondencias en inglés (*run through someone's head*) y en alemán (*jemandem im Kopf [herumspuken/herumgeistern/herumgehen]*). Para aquellas UF que tengan un rango colocacional diferente en las distintas lenguas abordadas, se ha creado la casilla *colocados*. Por ejemplo, la UF en español *rondar por la cabeza* puede

¹⁷ Microsoft Access 2016 es un sistema de gestión de bases de datos diseñado por Microsoft y perteneciente al paquete de aplicaciones de Microsoft Office.

colocar con conceptos como *idea* o *proyecto* (entre otros) pero no con *música*, a diferencia de *to run through someone's head* que sí puede colocar con los tres. Estas y otras posibles divergencias entre las UF quedan resumidas en la casilla *differences*, para una rápida visualización por parte del traductor. Asimismo, se ha empleado el asterisco «*» para todas aquellas correspondencias que sean una UF pero no un somatismo y el doble asterisco «**» en caso de haber empleado un término neutro o perífrasis explicativa ante la ausencia de una correspondencia fraseológica.

Además de estas características, el glosario también permite la búsqueda de UF que contengan una determinada combinación de letras, en caso de que el traductor pretenda crear un concreto efecto fonoestilístico en el texto meta. Así, el sistema de búsqueda permite el uso de los comodines más usuales tales como «*», que encuentra 0 o más caracteres; «?», para un número específico de caracteres; «[]», para detectar los caracteres introducidos entre paréntesis; «¡», para excluir los caracteres entre paréntesis; «-», para buscar un determinado rango de caracteres, y «#», para recuperar un carácter numérico.

De esta manera, en Artículo 11, examinaremos en profundidad como estas diferentes funcionalidades de Glossomatic pueden ser de utilidad para la creación de equivalencias fraseológicas *ad hoc* ante distintos casos de manipulación fraseológica en el texto origen para los que no existe una correspondencia primaria que pueda crear un efecto análogo en el texto meta. Asimismo, a este pormenorizado estudio de la estructura interna y usabilidad de Glossomatic, le seguirá la presentación de una propuesta didáctica en la que los traductores en formación aprenderán una serie de estrategias y pasos para ser implementados con el glosario a fin de superar el desafío de la manipulación fraseológica.

Artículo 11

Corpas Pastor, G., Hidalgo Ternero, C. M., y Bautista Zambrana, M. R. (2020). Teaching idioms for translation purposes: a trilingual corpus-based glossary applied to Phraseodidactics (ES/EN/DE). En F. Mena Martínez y C. Strohschen (Eds.), *Teaching Phraseology in the XXI Century: New Challenges* (pp. 75-93). Peter Lang

Abstract

Phraseology plays a pivotal role in the development of translation competence as well as in translation quality assessment. Thus far, however, there remains a paucity of research on how to best teach idioms for translation purposes. Against such a background, this study aims to shed some light on the multiple applications of phraseodidactics to translation training. We will follow a corpus-based methodology and, for the sake of the argument, the focus will be on somatisms in Spanish, English and German. The overall structure of this paper takes the form of four sections. Section one begins by laying out the theoretical dimensions of phraseology and its convergence with translation. In section two we examine the main components of a corpus-based glossary of somatisms, named Glossomatic, and how it can be employed to establish ad hoc phraseological equivalences in those cases (analysed in section three) where the manipulation of idioms and the absence of one-to-one phraseological correspondences may pose some problems to translation. In this regard, given the importance of accurately conveying the pragmatic, semantic and discursive load of an idiom into a TT and, concomitantly, conveying the manipulation depicted in the ST, section four presents a teaching proposal in which students are prompted with a set of strategies and steps to be implemented with the aid of the glossary in order to solve these issues. Overall, the insights gained from this research will prove useful not only in developing trainees' phraseological competence but also in giving centre stage to phraseodidactics in Translation Studies.

Keywords: phraseodidactics, ad hoc phraseological equivalences, idiom manipulation, corpus linguistics, glossary

Resumen

En este capítulo, presentamos las posibilidades que ofrece el glosario de somatismos basado en corpus denominado Glossomatic. Diseñado en una base de datos relacional en Microsoft Access 2016¹⁸, sigue un doble enfoque onomasiológico y semasiológico, debido a la propia naturaleza del mismo: dado que Glossomatic está concebido para la creación de equivalentes fraseológicos *ad hoc* en aquellos casos de manipulación fraseológica en el texto origen, es imprescindible que no solo se pueda consultar la información necesaria sobre una determinada UF y sus equivalencias en otras lenguas, sino que también sea posible obtener UF a partir de una serie de parámetros (*registro, campo semántico, características fonoestilísticas*, etc.). Así, para cada entrada, además

¹⁸ Microsoft Access 2016 es un sistema de gestión de bases de datos diseñado por Microsoft y perteneciente al paquete de aplicaciones de Microsoft Office.

del equivalente en inglés y alemán, se ofrece información microestructural relevante tal como *tipo de somatismo* (adjetival, verbal, nominal, preposicional o verbal), *nociones*, *colocados*, *registro* (formal, neutral, informal o vulgar), *polaridad* (positiva, neutra o negativa) y *diferencias interlingües*. En el campo de *noción* (Berthenet, 2011) se incluyen los distintos conceptos asociados a una UF (*enfado, engaño, amor*, etc.), lo que permite una fácil recuperación de UF que comparten un mismo campo semántico. Por ejemplo, bajo la noción *pensamiento* encontramos UF como *rondar por la cabeza* y sus correspondencias en inglés (*run through someone's head*) y en alemán (*jemandem im Kopf [herumspuken/herumgeistern/herumgehen]*). Para aquellas UF que tengan un rango colocacional diferente en las distintas lenguas abordadas, se ha creado la casilla *colocados*. Por ejemplo, la UF en español *rondar por la cabeza* puede colocar con conceptos como *idea* o *proyecto* (entre otros) pero no con *música*, a diferencia de *to run through someone's head* que sí puede colocar con los tres. Estas y otras posibles divergencias entre las UF quedan resumidas en la casilla *differences*, para una rápida visualización por parte del traductor. Asimismo, se ha empleado el asterisco «*» para todas aquellas correspondencias que sean una UF pero no un somatismo y el doble asterisco «**» en caso de haber empleado un término neutro o perífrasis explicativa ante la ausencia de una correspondencia fraseológica.

Palabras clave: fraseodidáctica, equivalencias fraseológicas *ad hoc*, manipulación fraseológica, lingüística de corpus, glosario

4.2.3 *Análisis comparativo de distintos recursos y herramientas orientadas a la traducción de la manipulación fraseológica*

Este apartado final tiene por objetivo comparar la utilidad de distintos recursos y herramientas electrónicos ante el desafío de traducir somatismos manipulados en el TO. Con este fin, en Artículo 12, presentamos una propuesta didáctica en la que los traductores en formación aprendieron las posibilidades que ofrecen no solo las herramientas anteriormente descritas (corpus lingüísticos, glosarios o traductores automáticos) sino también otros recursos lexicográficos adicionales, como los motores de búsqueda y los *rimadores* (que describiremos más adelante), en la traducción de la manipulación fraseológica. Entre estos recursos lexicográficos adicionales se incluyen tanto diccionarios monolingües en español (como el diccionario de la Real Academia Española [DLE, n. d.]) y en inglés (Cambridge Dictionary [CD, 2021] y Léxico [OD, 2021]) como bilingües (Wordreference [WR, n. d.] y Collins [COL, 2021]), entre otros. Estos diccionarios, además de la consulta del significado de una UF y de sus posibles correspondencias primarias en otros idiomas, también permiten, por ejemplo, la obtención de UF con un determinado somatónimo, a fin de crear una manipulación concreta para el TO.

En lo que respecta a los corpus lingüísticos, los traductores en formación descubrieron las posibilidades que ofrece la herramienta de gestión de corpus Sketch Engine y los corpus que ella abarca, tanto paralelos (OPUS2) como monolingües en español (esTenTen18 y Timestamped JSI web corpus 2014-2021 Spanish) e inglés (enTenTen20 y Timestamped JSI web corpus 2014-2021 English). Dentro de Sketch Engine, se les enseñó distintas funcionalidades tales como *Concordance*, a fin de analizar los contextos de uso de una UF concreta; *Thesaurus*, para encontrar sinónimos o palabras similares, y *Word Sketch*, que, como examinamos en el apartado 4.2.1., permite analizar el comportamiento gramatical y colocacional de un determinado lema, en caso de querer un equivalente fraseológico *ad hoc* para una UFM manipulada.

Finalmente, en esta publicación analizaremos también otros recursos adicionales, como los motores de búsqueda, que, como veremos, pueden ser de especial interés a fin de buscar distintos hipónimos dentro de una categoría específica, cuya manipulación alcance un efecto paronomástico concreto en el TO, o los *rimadores*, a fin de crear una determinada rima en el TO.

A esta primera actividad formativa con las posibilidades que ofrecen los recursos documentales le siguió una segunda en la que los estudiantes tuvieron que enfrentarse a seis escenarios distintos con manipulación fraseológica en el TO. De forma concreta, tres de ellos incluían somatismos manipulados y tres contenían otro tipo de UF. La finalidad de esta división fue analizar si la selección tanto de la estrategia de traducción como de la herramienta documental empleada era diferente según el tipo de UF, así como qué escenarios presentaban menores dificultades para la creación de equivalentes fraseológicos *ad hoc* en el texto meta, lo que permitirá arrojar algo más de luz sobre el proceso de traducción ante el desafío de la manipulación fraseológica.

Artículo 12

Hidalgo Ternero C. M., y Corpas Pastor, G. (2021a). Herramientas y recursos electrónicos para la traducción de la manipulación fraseológica: un estudio de caso centrado en el estudiante. *CLINA, An Interdisciplinary Journal of Translation, Interpreting and Intercultural Communication*, 71-94. <https://doi.org/10.14201/clina2020627194>

Abstract

This article presents a case study carried out with students of the subject Traducción General «BA-AB» (II) - Inglés-Español / Español-Inglés, taught in the first year of the Bachelor's Degree in Translation and Interpreting, at the University of Malaga. In this regard, at a first stage, students were trained on how to exploit the possibilities offered by different e-tools and resources (language corpora, lexicographic resources, or the web, inter alia) for the creation of textual equivalences in those cases where the manipulation of idioms and the absence of one-to-one phraseological correspondences may pose problems to translation. To this end, an introductory seminar on phraseological manipulation and translation was followed by a hands-on session, where trainee translators were presented with some scenarios including manipulated idioms in the source text (ST). Overall, the insights gained from analysing the results obtained will allow us to determine to what extent the different tools can help students walk the tightrope of translating phraseological manipulation.

Keywords: case study, phraseological manipulation, e-tools for translation, English-Spanish

Resumen

En el presente artículo se analiza un estudio de caso llevado a cabo con estudiantes de la asignatura Traducción General «BA-AB» (II) - Inglés-Español / Español-Inglés, impartida en el segundo semestre del segundo curso del Grado en Traducción e Interpretación de la Universidad de Málaga. En él, en una primera fase, se les enseñó a los estudiantes a cómo sacar el máximo partido de diferentes recursos y herramientas documentales electrónicos (corpus lingüísticos, recursos lexicográficos o la web, entre otros) para la creación de equivalencias textuales en aquellos casos en los que, fruto del anisomorfismo fraseológico interlingüe, la modificación creativa de unidades fraseológicas (UF) en el texto origen y la ausencia de correspondencias biunívocas presentan serias dificultades para el proceso traslaticio. De esta manera, a una primera actividad formativa sobre la traducción de usos creativos de unidades fraseológicas le sucede una sesión práctica en la que los alumnos tuvieron que enfrentarse a distintos casos de manipulación en el texto origen. Con el análisis de dichos resultados se podrá vislumbrar en qué medida los distintos recursos documentales ayudan a los traductores en formación a superar el desafío de la manipulación fraseológica.

Palabras clave: estudio de caso, manipulación fraseológica, herramientas documentales electrónicas aplicadas a la traducción, español-inglés.

CAPÍTULO V. DISCUSSION, CONCLUSION AND FUTURE LINES OF RESEARCH

This dissertation aimed at analysing and designing different electronic resources and tools for the translation of somatisms under the challenge of idiom variation and manipulation. This topic was mainly motivated by practical factors: many of the obstacles faced by translators and MT systems are directly caused by phraseology in the broad sense (Colson, 2008 & 2013), not only because of the ubiquity of phraseological units but also because of their intrinsic problematic features such as syntactic anomaly, non-compositionality, diasystematic variation, manipulation, ambiguity as well as the crosslinguistic phraseological anisomorphism, to name just a few. In this context, the methods applied in this research allowed us to examine and design different essential e-tools in the translation of PUs such as text corpora, NMT systems, lexicographic resources, and the web, inter alia. The main findings of this PhD thesis are summarised and discussed below, followed by suggestions for future lines of research.

5.1 Summary of contributions

The task of improving NMT systems in the translation of somatisms under the challenge of idiom variation and discontinuity was accomplished in two different stages. First, the thorough analysis of the two most robust NMT systems carried out in Paper 1 allowed us to observe that, regarding variation, DeepL and Google Translate could only deliver an overall high-quality performance in the translation of somatisms with high normalised frequencies such as *tomar el pelo* and *meter la pata*, and their nominal variants, *tomadura de pelo* y *metedura de pata*. However, NMT systems still faced serious difficulties with other lower-frequency somatisms (such as *meter la pata hasta el corvejón*, *meter la gamba*, and its nominal variants *metedura de gamba* and *metida de gamba*) and, in the case of Google Translate, other diatopically-marked variants (*tomada de pelo* and

metida de pata), which present higher frequency scores in widely spoken Latin-American national varieties. These results hence accentuate the necessity of further improving NMTs' performance not only for rare and out-of-vocabulary words but also to encompass diatopic variation in all its facets.

In this line, Paper 1's results also showed that phraseological discontinuity poses serious difficulties for NMT systems: more specifically it meant a decline of DeepL's performance by 13% and of Google Translate's by 15%, when compared to the continuous form of these idioms. In order to overcome this obstacle, the text preprocessing system *gApp* was designed for the automatic detection and conversion of discontinuous somatisms into their continuous form, with very promising results. In Paper 2, *gApp* proved to offer a considerable enhancement in the translation of somatisms in relative terms, i.e., when contrasting the discontinuous versus the continuous scenario, with an average improvement by 10% for DeepL and by 14.6% for Google Translate in the ES>EN directionality, nearly attaining the manual conversion (there was only a 2.7% difference between both types of conversion). Nevertheless, this enhancement could not still reach an optimal performance in absolute terms: while the continuous form after *gApp* could only obtain a final average accuracy rate of 90.7% for DeepL, in the case of Google Translate's global performance, it could only amount to 75.4%. In Paper 3, *gApp* presented an overall enhancement for DeepL by 13.5% in the ES>EN and by 9% in the ES>DE directionalities (with only a 0.5% difference with the manual conversion). In absolute terms, the conversion through *gApp* could only obtain an accuracy rate of 62.5% in ES>EN and of 52.5% in ES>DE. In Paper 4, *gApp* meant an amelioration of DeepL's performance by 18% for FR>EN, and by 16.5% for FR>ES, with analogous results to the manual conversion. However, in absolute terms, *gApp* could only achieve a score of 58% in both language directionalities. In Paper 5, this system could improve

ModernMT's performance, on average, by 7.5% and DeepL and Google Translate's by 6.5% in the ES>EN/DE/FR/IT/PT directionalities. Regarding the NMTs' final performance after gApp, ModernMT could achieve an accuracy rate of 54.5%, DeepL of 75.2% and Google Translate of 70.1%. In Paper 6, gApp could improve VIP's performance by 21.5% (final NMT's accuracy rate after *gApp*: 67%), DeepL's by 13.5% (final accuracy rate: 85.5%) and Google Translate's by 8.5% (final accuracy rate: 77.5%) in the ES>EN directionality. Finally, in Paper 7, we could observe how gApp was able to ameliorate, on average, ModernMT's performance by 12.8% (final NMT's accuracy rate after *gApp*: 39.5%), Google Translate's by 14.3% (final accuracy rate: 71.8%) and DeepL's by 2.8% (final accuracy rate: 71.3%) in the IT>EN/DE directionalities. Regarding the types of conversion, in Papers 5-7 *gApp* once again showed a quasi-analogous performance to the manual conversion ($\pm 1\%$). Against such a background, a summary of *gApp*'s results is presented in Table 2.

	Language directionalities	NMTs analysed	NMTs' accuracy before gApp	NMTs' accuracy after gApp	Improvement after gApp	Manual conversion
Paper 2	ES>EN	DeepL	80.7%	90.7%	10%	+3.2%
		Google Translate	60.7%	75.4%	14.6%	+2.1%
Paper 3	ES>EN	DeepL	49%	62.5%	13.5%	+0.5%
	ES>DE		43.5%	52.5%	9%	+0.5%
Paper 4	FR>EN	DeepL	40%	58%	18%	=
	FR>ES		41.5%	58%	16.5%	=
Paper 5	ES>EN	ModernMT	50%	60%	10%	=
	ES>DE		23.3%	33.3%	10%	=
	ES>FR		49.3%	60%	10.7%	=
	ES>IT		56.7%	60.7%	4%	=
	ES>PT		56%	58.7%	2.7%	=
	ES>EN	DeepL	70.7%	81.3%	10.7%	+0.7%
	ES>DE		59.3%	66.7%	7.3%	+0.7%
	ES>FR		69.3%	74%	4.7%	+0.7%
	ES>IT		76%	80%	4%	+0.7%
	ES>PT		68%	74%	6%	+0.7%

	ES>EN	Google Translate	66%	75.3%	9.3%	=
	ES>DE		35.3%	43.3%	8%	=
	ES>FR		65.3%	73.3%	8%	=
	ES>IT		78.7%	79.3%	0.7%	=
	ES>PT		72.7%	79.3%	6.7%	=
Paper 6	ES>EN	VIP	45.5%	67%	21.5%	-0.5%
	ES>EN	DeepL	77%	85.5%	8.5%	=
	ES>EN	Google Translate	64%	77.5%	13.5%	-1%
Paper 7	IT>EN	ModernMT	25.5%	42%	16.5%	+1%
	IT>DE		28%	37%	9%	=
	IT>EN	Google Translate	64.5%	82.5%	18%	+0.5%
	IT>DE		50.5%	61%	10.5%	-1%
	IT>EN	DeepL	75%	75%	0%	-0.5%
	IT>DE		62%	67.5%	5.5%	=
Total			58.3%	70.9%	12.5%	+0.6%

Table 2. Summary of gApp's results

Regarding the use of corpora for the translation of somatisms, in Paper 8 it was possible to observe the possibilities (and limitations) of these tools in the correct identification and interpretation of *ambiguous* phraseological units, i.e., those cases of PUs with different meanings due to a slight alteration in the form (*de narices* vs. *de las narices*, *levantar cabeza* vs. *levantar la cabeza*, *a pie* vs. *al pie*) or a dissimilar context (*a pelo* either meaning ‘without a saddle’, ‘naked’ or ‘without protection’, depending on the context). In this regard, students were prompted with a teaching proposal where they could learn a set of heuristic strategies to refine their searches in the consulted corpora (CORPES XXI, Linguee, and Sketch Engine) and to select adequate equivalences after an appropriate interpretation of the concordances obtained. In this context, the results yielded by this study allowed to observe that Sketch Engine was the corpus management system offering a better performance. In Paper 9, we analysed the query interface of different corpus management systems (CREA, in its traditional and annotated version, CORPES XXI, and Sketch Engine) for the retrieval process of idioms (*tener entre manos*, *traer entre manos* and *llevar entre manos*, as well as *ir al pelo* and *venir al pelo*) in all their formal manifestations (lexical variation, grammatical

inflection, discontinuity...). In the light of these results, it was possible to determine that Sketch Engine was once again the system offering a better performance both in quantitative and qualitative terms in the task of providing a larger (and better) amount of results with the least possible searches. In Paper 10, we examined the possibilities of Sketch Engine's functionality, Word Sketch, in the creation of *ad hoc* phraseological equivalences for those cases where the manipulation of idioms in the source text and the crosslinguistic phraseological anisomorphism may pose serious difficulties for the translation task. In spite of this *phraseological obstacle*, in the translation proposals presented by trainee translators it was possible to observe that most of them were able to provide not only a manipulated idiom but more specifically a manipulated somatism for the TT (97% in translation scenario 1, 74% in scenario 2, 51% in scenario 3, and 71% in scenario 4). This is mainly due to the pervasiveness of body images in the phraseological repertoire of both languages which, hence, considerably increases the possibilities of creating a manipulated somatism in the target text. Nevertheless, not all somatonyms have the same recursiveness, which thus determines the use of Word Sketch in the query process, as it can be observed in the different translation proposals: while 91% had recourse to Word Sketch for translation scenario 1 (with productive somatonyms in Spanish such as *cabeza* ['head']); only 51% of respondents in translation scenario 3, and 57% in translation scenario 4 employed this feature (with less productive somatonyms in Spanish such as *espalda* ['back']). Another interesting finding is that in those cases where the manipulated somatism in the ST (*to go on a head* in scenario 2) has a primary correspondence in Spanish (*ir en cabeza*) whose manipulation can purvey an analogous double reading in the TT, exclusively 8 trainee translators (23%) had recourse to Word Sketch in contrast with the 17 respondents (49%) who employed the simple query in the parallel corpus OPUS2 so as to provide a phraseological equivalent for the TT. These results suggest that

trainee translators only employ Word Sketch as a information tool when the manipulated idiom in the ST does not possess a primary correspondence in the phraseological repertoire of the target language which may be subject to a similar manipulation.

In this quest for translating manipulated somatisms in the ST, the trilingual corpus-based glossary Glossomatic was presented in Paper 11. In this regard, after the main components of the glossary were summarised, we examined the multiple applications of Glossomatic for the creation of *ad hoc* phraseological equivalences in those cases where the manipulation of idioms and the absence of one-to-one phraseological correspondences may pose some problems to translation. Subsequently, a teaching proposal was presented in which students were prompted with a set of heuristic strategies and steps to be implemented with Glossomatic in the different translation scenarios in the EN>ES and the EN>DE language directionalities. Further resources for the translation of idiom manipulation were presented in Paper 12. In this regard, we carried out a case study with trainee translators, where, at a first stage, they were trained on how to exploit the possibilities offered by different e-tools (language corpora, lexicographic resources, or the web, inter alia) for the creation of textual equivalences for manipulated idioms in the ST. In the light of the results yielded by this study, it was possible to observe that, regarding the translation strategy, those cases in which trainee translators mainly opted for manipulated idioms in the TT corresponded to two main types of scenarios. In the first scenario, it was possible to find those cases in which the primary correspondences of a given PU in the ST shared an analogous denotative and connotative meaning and a similar metaphorical base to the TT, such as *tener cabeza* and *to have a good head on one's shoulders*. The second scenario corresponds to those cases in which trainee translators could offer a manipulated idiom in the TT even if there was no primary correspondence in the TT that could be subject to an analogous manipulation because of

sharing a similar lexical base. In this case, translators could overcome this obstacle by resorting to the considerable number of conceptual metaphors and metonymies underlying the phraseological repertoires of both languages, which makes it easier to detect idioms in the target language whose manipulation may attain an analogous effect to the one in the ST.

Regarding the e-tools, the results reported in Paper 12 allowed to observe a predominant use of lexicographic resources and textual corpora, with analogous global scores: 34.6% and 32.4%, respectively. With regard to the other tools, there is a markedly lower usage trend in the different scenarios, with the exception of two cases. In case 1, the need to find toponyms or ethnonyms whose manipulation could achieve a similar paronomastic effect to the one in the source text led a large majority of trainee translators (79.3%) to consult search engines. In case 2, the second most used resource was rhymers, employed by 8 out of the 32 students who made use of resources for case 2 (i.e., 25%). Finally, the results show that machine translation was only employed on one occasion (0.7%), making it the least frequently used tool in the translation of these cases of phraseological manipulation in the source text.

The cases with manipulated somatisms in the target text show even more revealing results in relation to the e-tools. In these scenarios, among trainee translators who have made use of e-tools, it is possible to observe an almost exclusive employment of lexicographic resources and textual corpora: 51.9% and 44.4%, respectively, in case 4; 60% and 30% in case 5; and 41.7% and 58.3% in case 6. The only other resource employed in these cases was search engines, in a very residual way: 3.7% in case 4, 10% in case 5 and no use in case 6. This implies that the predominant use of one or other of the resources is closely linked to the type of phraseological manipulation of the source text. In cases requiring generic searches (such as for different toponyms or ethnonyms) or a rhyme in the target text, trainee translators will preferably use search engines and rhymers,

respectively. On the other hand, if they need to create an ad hoc phraseological equivalent with a specific base (e.g., *head*), they will mainly resort to lexicographic resources and textual corpora.

5.2 Future lines of research

Regarding the different possibilities offered by the analysed e-tools and resources in the translation of somatisms, there are still some areas we would like to explore further. For instance, with respect to *gApp* (Papers 2-7), the promising results with semi-fixed VNC and VPP somatisms in the ES>EN/DE/FR/IT/PT, FR>EN/ES and IT>EN/DE directionalities invite to expand its detection lexicon and conversion mechanism so as to examine to what extent it can also improve NMTs' performance for other MWE categories affected by discontinuity (such as verb-particle constructions [VPCs]), other MWE typologies (chromatisms, zoologisms, or naturisms, among others), additional language directionalities, as well as other NMT systems. Furthermore, we also intend to analyse the scalation of this model to other NLP tasks where the pre-processing of discontinuous MWEs may play an important role (information retrieval, automatic text summarisation, sentiment analysis, inter alia), in order to enhance MWE-aware NLP systems.

With respect to corpora and lexicographic resources in their role of essential e-tools in the translation of somatisms, the promising results obtained with Sketch Engine (Papers 8 and 9), the limitations of lexicographic resources in the treatment of phraseology (Paper 9), and the possibilities offered by Glossomatic (Paper 11) encourage us to design a hybrid dictionary of somatisms, which, besides providing useful lexicographic information (definition, register, notions, correspondences in other languages...), will also provide translators with direct access to quality corpora in order to be able to examine a somatism (and its variants) in contexts of real use. In this way, as stated in Paper 9, we will develop an eclectic approach that will combine the best of both

worlds: the conciseness and clarity of dictionaries with the comprehensiveness and textual richness of corpora.

Finally, in subsequent studies we would also like to delve into the interaction of Construction Grammar and phraseology applied to the study of somatisms. In this regard, we would like to carry out a corpus analysis of productive phraseological variants (Mellado Blanco & Gutiérrez Rubio, 2020) such as *estar hasta la polla/el coño/los cojones/los huevos/el nabo/el moño/la coronilla/las narices/el gorro...* ('to be fed up'), *tocar la polla/los cojones/el coño/los huevos/el nabo/las narices/las pelotas/la moral...* ('to pester someone'), or *y una polla/una mierda/un cojón/una leche...* ('no way'), in order to examine which somatonyms usually occur in the paradigm constituting the slot, and what euphemisms (such as *la moral* in the idiom *me toca la moral*) are normally employed to substitute taboo somatonyms related to sexual activities (*la polla/el coño/los cojones*, etc.), among other research questions.

BIBLIOGRAFÍA

- Anastasiou, D. (2010). *Idiom Treatment Experiments in Machine Translation*. Cambridge Scholars Publishing.
- Bahdanau, D., Cho, K., y Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Baran À Nkoum, P. (2015). *Estudio contrastivo español-francés de las locuciones verbales somáticas relativas a la cabeza* [Tesis doctoral, Universidad Complutense de Madrid] <https://eprints.ucm.es/id/eprint/33799/>
- Bentivogli, L., Bisazza, A., Cettolo, M., y Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.
- Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., ... Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation (WMT16). En O. Bojar, C. Buck, R. Chatterjee, C. Federmann, L. Guillou, B. Haddow, M. Huck, A. Jimeno Yepes, A. Névél, Mariana Neves, P. Pecina, M. Popel, P. Koehn, C. Monz, M. Negri, M. Post, L. Specia, K. Verspoor, J. Tiedemann, M. Turchi (Eds.), *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers* (pp. 131–198). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W16-2301>
- Booij, G. (2002). Constructional idioms, morphology, and the Dutch lexicon. *Journal of Germanic linguistics*, 14(4), 301–329.
- Burger, H., Buhofer, A., y Sialm, A. (1982). *Handbuch der Phraseologie*. Walter de Gruyter.
- CD – Cambridge University Press (2021). *Cambridge Dictionary*. <https://dictionary.cambridge.org/>
- Casacuberta Nolla, F., y Peris-Abril, Á. (2017). Traducció automàtica neuronal. *Tradumàtica*, 15, 66-74. <https://doi.org/10.5565/rev/tradumatica.203>
- Casares, J. (1950). *Introducción a la lexicografía española*. CSIC.
- Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., y Way, A. (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1), 109–120.

- Colson, J.-P. (2008). Cross-linguistic phraseological studies: An overview. En S. Granger y F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 191–206). John Benjamins. <https://doi.org/10.1075/z.139.19col>
- Colson, J.-P. (2013). Pratique traduisante et idiomatité: l'importance des structures semi-figées. En P. Mogorrón Huerta, D. Gallego Hernández, P. Maseau y M. Tolosa Igualada (Eds.), *Fraseología, opacidad y traducción* (pp. 207–218). Peter Lang.
- Colson, J.-P. (2015). *The contribution of corpus-based phraseology to translation studies: from experiments to theory* [Conferencia invitada]. Europhras 2015, Málaga, España. <http://hdl.handle.net/2078.1/165297>
- Colson, J.-P. (2019). Multi-word Units in Machine Translation: why the Tip of the Iceberg Remains Problematic—and a Tentative Corpus-driven Solution. En G. Corpas Pastor, R. Mitkov, M. Kuilovskaya, M. A. Losey León (Eds.), *Proceedings of the Third International Conference EUROPHRAS 2019* (pp. 145–156). Tradulex.
- Constant, M., Eryigit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., y Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, 43(4), 1–92.
- CDSD – Corriere Della Sera/Dizionari (2021). *Dizionario dei modi di dire*. <https://dizionari.corriere.it/dizionario-modi-di-dire/>
- Corpas Pastor, G. (1996). *Manual de fraseología española*. Gredos.
- Corpas Pastor, G. (Ed.). (2000). *Las lenguas de Europa, estudios de fraseología, fraseografía y traducción*. Comares.
- Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Peter Lang.
- Corpas Pastor, G. (2013). Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. En I. Olza y E. Manero (Eds.), *Fraseopragmática* (pp. 335–373). Frank & Timme.
- Corpas Pastor, G. (2015). Translating English Verbal Collocations into Spanish: on Distribution and other Relevant Differences related to Diatopic Variation. *Linguisticae Investigationes*, 38(2), 229–262. <https://doi.org/10.1075/li.38.2.03cor>
- Corpas Pastor, G. (2018). Laughing one's head off in Spanish Subtitles: a corpus-based study on diatopic variation and its consequences for translation. En P. Mogorrón Huerta y A.

- Martínez Albaladejo (Eds.), *Fraseología, Diatopía y Traducción* (pp. 32-71). John Benjamins.
- Corpas Pastor, G. (2021). Constructional idioms of 'insanity' in English and Spanish: a corpus-based study. *Lingua*. <https://doi.org/10.1016/j.lingua.2020.103013>
- Corpas Pastor, G. y Colson, J. P. (Eds.). (2020). *Computational Phraseology*. John Benjamins Publishing Company.
- Corpas Pastor, G.; Bautista Zambrana, M. R. e Hidalgo Ternero, C. M. (eds.) (2021). *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus*. Comares. ISBN: 9788413690254.
- Corpas Pastor, G., Hidalgo Ternero, C. M., y Bautista Zambrana, M. R. (2020). Teaching idioms for translation purposes: a trilingual corpus-based glossary applied to Phraseodidactics (ES/EN/DE). En F. Mena Martínez y C. Strohschen (Eds.), *Teaching Phraseology in the XXI Century: New Challenges* (pp. 75-93). Peter Lang.
- CORPES XXI – RAE (2018). *Corpus del Español del Siglo XXI (CORPES XXI)*.
<http://www.rae.es/recursos/banco-de-datos/corpes-xxi>
- CREA – RAE (2018). *Corpus de Referencia del Español Actual (CREA)*.
<https://www.rae.es/banco-de-datos/crea>
- Davies, M. (2018). *Corpus del Español: NOW*. <http://www.corpusdelespanol.org/>
- Derczynski, L., Ritter, A., Clark, S., y Bontcheva, K. (2013). Twitter part-of-speech tagging for all: overcoming sparse and noisy data. En R. Mitkov, G. Angelova y K. Bontcheva (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (pp. 198–206). INCOMA Ltd. <http://www.aclweb.org/anthology/R13-1026>
- Dobrovol'skij, D. (2011a). Phraseologie und Konstruktionsgrammatik. En A. Lasch y A. Ziem (Eds.), *Konstruktionsgrammatik III: Aktuelle Fragen und Lösungsansätze* (pp. 111-130). Stauffenburg Verlag.
- Dobrovol'skij, D. (2011b). Cross-linguistic equivalence of idioms: does it really exist. En A. Pamies y D. Dobrovol'skij (Eds.), *Linguo-cultural competence and phraseological motivation* (pp. 7–24). Schneider.
- Dobrovol'skij, D. (2013). German-Russian idioms online: on a new corpus-based dictionary. En V. Selegej (Eds.), *Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialogue"*. RGGU.

- Dobrovol'skij, D. (2016). Fraseología y Gramática de Construcciones (M. J. Sánchez Sánchez, trad.). *Language design: journal of theoretical and experimental linguistics*, 18, 71-106.
- DPLP – Priberam Informática, S.A. (2021). *Dicionário Priberam da Língua Portuguesa*. <https://dicionario.priberam.org>
- DLE – Real Academia Española (n. d.). *Diccionario de la Lengua Española*. <http://www.rae.es>
- DUDEEN – Dudenredaktion (Bibliographisches Institut) (2006). *Duden: Die deutsche Rechtschreibung*. Dudenverlag.
- Durán-Muñoz, I., y Corpas Pastor, G. (2020). Corpus-based multilingual lexicographic resources for translators: an overview. En M. J. Domínguez Vázquez, M. Mirazo Balsa y C. Válcárcel Riveiro (Eds.), *Studies on Multilingual Lexicography (Lexicographica, Series Maior)* (pp. 159-178). De Gruyter. <https://doi.org/10.1515/9783110607659-009>.
- EOLD – Oxford University Press (2018). *English Oxford 'Living' Dictionaries*. Oxford Dictionaries. <http://www.oxforddictionaries.com/>
- Erman, B., y Warren, B. (2000). The idiom principle and the open-choice principle. *Text*, 20(1), 29–62. <https://doi.org/10.1515/text.1.2000.20.1.29>
- ELIS – European Language Industry Survey (2018). *2018 Language Industry Survey – Expectations and Concerns of the European Language Industry*. https://ec.europa.eu/info/sites/default/files/2017_language_industry_survey_report_en.pdf
- ELIS – European Language Industry Survey (2020). *2020 Language Industry Survey – 2020 before & after COVID-19*. https://ec.europa.eu/info/sites/default/files/2020_language_industry_survey_report.pdf
- ELIS – European Language Industry Survey (2021). *2021 Language Industry Survey*. https://ec.europa.eu/info/sites/default/files/about_the_european_commission/service_standards_and_principles/documents/elis_2021_european_language_industry_survey.pdf
- Fadaee, M., Bisazza, A., y Monz, C. (2018). Examining the tip of the iceberg: A data set for idiom translation. *arXiv preprint arXiv:1802.04681*.
- Fantinuoli, C., y Zanettin, F. (Eds.). (2015). *New directions in corpus-based translation studies*. Language Science Press.
- Farø, K. J. (2006). Dogmatismus, Skeptizismus, Nihilismus und Pragmatismus bei der Idiomübersetzung: Grundfragen zu einer idiomtranslatorischen Theorie. En A. Häcki Buhofer y H. Burger (Eds.), *Phraseology in Motion I* (pp. 188-202). Baltmannsweiler.

- FD – Farlex (n.d.) in the hollow of (one's) hand. En *Farlex Dictionary of Idioms*. Recuperado 13 de febrero de 2021, de [https://idioms.thefreedictionary.com/in+the+hollow+of+\(one%27s\)+hand](https://idioms.thefreedictionary.com/in+the+hollow+of+(one%27s)+hand)
- Fleischer, W. (1983). Phraseologie. En W. Fleischer et al. (Eds.), *Kleine Enzyklopädie. Deutsche Sprache* (pp. 107-126). Bibliographisches Institut.
- Fleischer, W. (1997). *Phraseologie der deutschen Gegenwartssprache, 2. Durchgesehene und ergänzte Auflage*. Niemeyer.
- Foufi, V., Nerima, L., y Wehrli, E. (2019). Multilingual parsing and MWE detection. En Y. Parmentier y J. Waszczuk (Eds.), *Representation and parsing of multiword expressions: Current trends* (pp. 217–237). Language Science Press.
- Gablasova, D., Brezina, V., y Mcenery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning*, 67(S1), 155-179.
- García-Page, M. (1996a). *Problemas en el empleo de fraseología española por hablantes extranjeros: la violación de las restricciones*. En M. Rueda, J. Le Men y F. J. Grande (Eds.), *Actuales tendencias en la enseñanza del español como lengua extranjera II: Del pasado al futuro, Actas de VIII Congreso Internacional de ASELE* (pp. 363–369). Universidad de Alcalá.
- García-Page, M. (1996b). Sobre las variantes fraseológicas en español. *Revista canadiense de estudios hispánicos*, 20(3), 477–490.
- García-Page, M. (1999). Variantes morfológicas y unidades fraseológicas. *Paremia*, 8, 225–230.
- García-Page, M. (2007). Esquemas sintácticos de formación de locuciones adverbiales. *Moenia: Revista lucense de lingüística and literatura*, 13, 121-144.
- García-Page, M. (2008). *Introducción a la fraseología española*. Anthropos.
- Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. *Collocations and idioms: Linguistic, lexicographic, and computational aspects*, 23. Continuum Press.
- Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.
- Goldberg, A. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

- González Rey, M. I. (2012). De la didáctica de la fraseología a la fraseodidáctica. *Paremia*, 21, 67–84.
- González Rey, M. I. (2017). Le dictionnaire phraséodidactique: sa place dans la didactique de la phraséologie. *Studii de lingvistică*, 7, 27–44.
- González Rey, M. I. (Ed.). (2015). *Phraseology, Phraseodidactics and Construction Grammar(s)*. *Journal of Social Sciences*, 11(3). <https://thescipub.com/issue-jss/11/3>
- Gréciano, G. (1987). Idiom und sprachspielerische Textkonstitution. En J. Korhonen (Ed.), *Beiträge zur allgemeinen und germanistischen Phraseologieforschung* (pp. 193–206). Oulu.
- Gross, M. (1984). Une classification des phrases figées du français. *Revue Québécoise de linguistique* 11(2), 237–271.
- Gui, T., Zhang, Q., Huang, H., Peng, M., y Huang, X. (2017). Part-of-speech tagging for twitter with adversarial neural networks. En M. Palmer, R. Hwa y S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2411–2420). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1256>.
- Gutiérrez Pérez, R. (2010). *Estudio cognitivo-contrastivo de las metáforas del cuerpo. Análisis empírico del corazón como dominio fuente en inglés, francés, español, alemán e italiano*. Peter Lang.
- Hallsteinsdóttir, E. (2011a). Aktuelle Forschungsfragen der deutschsprachigen Phraseodidaktik. *Linguistik online*, 47, 3–31.
- Hallsteinsdóttir, E. (2011b). Phraseological competence and the translation of phrasemes. En A. Pamiés-Bertrán, L. Luque-Nadal y J.-M. Pazos (Eds.), *Multi-Lingual Phraseography: Second Language Learning and Translation Applications* (pp. 279–288). Schneider Verlag.
- Hallsteinsdóttir, E., Winzer-Kiontke, B., y Laskowski, M. (Eds.). (2011). *Phraseodidaktik/Phraseodidactics*. *Linguistik online*, 47(3).
- COL – HarperCollins Publishers (2021). *Collins dictionary*. <https://www.collinsdictionary.com/>
- Heid, U. (2008). Computational Phraseology: An Overview. En S. Granger y F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 337–360). John Benjamins.
- Hidalgo Ternero, C. M. (2018). [Reseña del libro *Corpus-based Approaches to Translation and Interpreting: From Theory to Applications*, de G. Corpas Pastor y M. Seghiri (coord.)]. *Hikma: Revista de Traducción*, 17, 235–237 <https://doi.org/10.21071/hikma.v17i0.11168>

- Hidalgo-Ternero, C. M. (2019a). Translating manipulated idioms (EN>ES) in the Word Sketch scenario. En G. Corpas Pastor y R. Mitkov (Eds.), *Computational and Corpus-Based Phraseology* (pp. 173-186). Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-30135-4_13
- Hidalgo-Ternero, C. M. (2019b) [Reseña del libro *Trends in e-tools and resources for translators and interpreters*, de G. Corpas Pastor e I. Durán Muñoz (Eds.)]. *TRANS: Revista de Traductología*, 23, 342–345. <https://doi.org/10.24310/TRANS.2019.v0i23.6415>
- Hidalgo-Ternero, C. M. (25-27 de septiembre de 2019a). Translating manipulated idioms (EN>ES) in the Word Sketch scenario [Sesión de conferencia]. *Congreso Internacional ‘Computational and Corpus-based Phraseology’ (EUROPHRAS 2019)*, Málaga, España. <http://www.lexytrad.es/europhras2019/wp-content/uploads/2019/09/Europhras-2019-programme.pdf>
- Hidalgo-Ternero, C. M. (25-27 de septiembre de 2019b). Google Translate vs. DeepL: analysing machine translation performance under the challenge of phraseological variation [Sesión de conferencia]. *Congreso Internacional ‘Computational and Corpus-based Phraseology’ (EUROPHRAS 2019)*, Málaga, España. <http://www.lexytrad.es/europhras2019/wp-content/uploads/2019/09/Europhras-2019-programme.pdf>
- Hidalgo-Ternero, C. M. (2020a). Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation. En P. Mogorrón Huerta (Ed.), *Análisis multidisciplinar del fenómeno de la variación en traducción e interpretación / Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting. MonTI Special Issue 6*, 154-177. <https://doi.org/10.6035/MonTI.2020.ne6.5>
- Hidalgo-Ternero, C. M. (2020b). [Reseña del libro *Computational Phraseology*, de G. Corpas Pastor, Gloria y J.-P. Colson (Eds.)]. *Yearbook of Phraseology*, 11, 203–205. <https://doi.org/10.1515/phras-2020-0011>
- Hidalgo-Ternero C. M. (2021). El algoritmo ReGap para la mejora de la traducción automática neuronal de expresiones pluriverbales discontinuas (FR>EN/ES). En G. Corpas Pastor, M. R. Bautista Zambrana y C. M. Hidalgo-Ternero (Eds.), *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus* (pp. 253-270). Comares

- Hidalgo Ternero, C. M., y Corpas Pastor, G. (2020a). Estrategias heurísticas con corpus para la enseñanza de la fraseología orientada a la traducción. En M. Seghiri (Ed.), *La lingüística de corpus aplicada al desarrollo de la competencia tecnológica en los estudios de traducción e interpretación y la enseñanza de segundas lenguas* (pp. 181-204). Peter Lang. <https://doi.org/10.3726/b14734>
- Hidalgo-Ternero, C. M., y Corpas Pastor, G. (2020b). Bridging the ‘gApp’: improving neural machine translation systems for multiword expression detection. *Yearbook of Phraseology*, 11(1), 61–80. <https://doi.org/10.1515/phras-2020-0005>
- Hidalgo Ternero C. M., y Corpas Pastor, G. (2020c). Herramientas y recursos electrónicos para la traducción de la manipulación fraseológica: un estudio de caso centrado en el estudiante. *CLINA, An Interdisciplinary Journal of Translation, Interpreting and Intercultural Communication*, 71-94. <https://doi.org/10.3726/b14734>
- Hidalgo-Ternero, C. M., y Corpas Pastor, G. (2021a). La variación fraseológica: análisis del rendimiento de los corpus monolingües como recursos de traducción. *Études romanes de Brno*, 359-379. <https://doi.org/10.5817/ERB2021-1-19>
- Hidalgo-Ternero C. M., y Corpas Pastor, G. (2022a/en prensa). *ReGap: a text preprocessing algorithm to enhance MWE-aware neural machine translation systems*. En J. Monti, G. Corpas Pastor y R. Mitkov (Eds.), *Recent Advances in MWU in Machine Translation and Translation technology*. John Benjamins Publishing Company.
- Hidalgo-Ternero C. M., y Corpas Pastor, G. (2022b/en prensa). Qué se traerá *gApp* entre manos... O cómo mejorar la traducción automática neuronal de variantes somáticas (ES>EN/DE/FR/IT/PT). En M. Seghiri y M. Pérez Carrasco (Eds.), *Aproximación a la traducción especializada*. Peter Lang.
- Hidalgo-Ternero C. M., y Corpas Pastor, G. (2022c/en prensa). A la cabeza de la traducción automática neuronal asistida por *gApp*: somatismos en VIP, DeepL y Google Translate. En G. Corpas Pastor y M. Seghiri (Eds.), *Aplicaciones didácticas de las tecnologías de la interpretación*. Comares.
- Hidalgo-Ternero C. M., Lista, F. y Corpas Pastor, G. (2022/en revisión). *gApp-assisted NMT: how to improve the neural machine translation of discontinuous multiword expressions (IT>EN/DE)*. Language Resources and Evaluation.

- Honnibal, M., y Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.
- Huang, P. S., Wang, C., Huang, S., Zhou, D., y Deng, L. (2017). Towards neural phrase-based machine translation. *arXiv preprint arXiv:1706.05565*.
- Hurtado Albir, A. (Ed.). (1999). *Enseñar a traducir: Metodología en la formación de traductores e intérpretes*. Edelsa Grupo Discalia.
- Jackendoff, R. S. (1995). The boundaries of the lexicon. En M. Everaert, E. van der Linden, A. Schenk, y R. Schroeder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 133–165). Lawrence Erlbaum Associates.
- WR – Kellogg, M. (n. d.). *Wordreference*. <http://www.wordreference.com/>
- Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovvár, V., Michelfeit, J., Rychlý, P. y Suchomel, V. (2003). *The Sketch Engine*. <http://www.sketchengine.co.uk>
- Koller, W. (2007). Probleme der Übersetzung von Phrasemen. En H. Burger, D. Dobrovolskij, P. Kühn y N. Norrick (Eds.), *Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung/Phraseology. An International Handbook of Contemporary Research* (pp. 605-613). Walter de Gruyter.
- Ladmiral, J. (2015). La traductologie: entre phraseologie et didactique. En P. Mogorrón Huerta y F. Navarro Domínguez (Eds.), *Fraseología, Didáctica y Traducción* (pp. 227–240). Peter Lang.
- Lakoff, G., y Johnson, M. (1980). *Metáforas de la vida cotidiana*. Cátedra.
- LAR – Larousse (2018). *Dictionnaires Larousse français monolingue et bilingues en ligne*. <http://www.larousse.fr/dictionnaires/francais>
- LING – Linguee GmbH (2008). *Linguee*. <https://www.linguee.com/>
- Martí Sánchez, M. (2015). La búsqueda de sentido en la desautomatización fraseológica. En P. Mogorrón Huerta y F. Navarro Domínguez (Eds.), *Fraseología, Didáctica y Traducción* (pp. 117-135). Peter Lang.
- MCERL – Ministerio de Educación, Cultura y Deporte (2002). *Marco Común Europeo De Referencia Para Las Lenguas: Aprendizaje, Enseñanza, Evaluación*. Anaya
- Mel’čuk, I. (1995). Phrasemes in language and phraseology in linguistics. En M. Everaert, E. van der Linden, A. Schenk, y R. Schroeder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 167–232). Lawrence Erlbaum Associates.

- Mellado Blanco, C. (2004). *Fraseologismos somáticos del alemán*. Peter Lang.
- Mellado Blanco, C. (2008). *Colocaciones y fraseología en los diccionarios*. Peter Lang.
- Mellado Blanco, C. (2013). Tipología de la motivación fraseológica en un corpus onomasiológico alemán-español. En P. Mogorrón Huerta, D. Gallego Hernández, P. Masseau y M. Tolosa Igualada (Eds.), *Fraseología, Opacidad y Traducción* (pp. 39-65). Peter Lang..
- Mellado Blanco, C. (2015). Parámetros específicos de equivalencia en las unidades fraseológicas (con ejemplos del español y el alemán). *Revista de Filología de la Universidad de La Laguna*, 33, 153–174.
- Mellado Blanco, C. (2020). Esquemas fraseológicos y construcciones fraseológicas en el continuo léxico-gramática. En E. Tabares, C. Sinner y E. T. Montoro (Eds.), *Clases y categorías en la fraseología de la lengua española* (pp. 13–36). Peter Lang.
- Mellado Blanco, C. y Gutiérrez Rubio, E. (2020). Nuevas aportaciones de la Gramática de Construcciones a los estudios de fraseología en las lenguas románicas. *Romanica Olomucensia*, 32(1).
- Mena Martínez, F. (2003). En torno al concepto de desautomatización fraseológica: aspectos básicos. *Tonos, Revista Electrónica de Estudios Filológicos* 5. <http://www.tonosdigital.es/ojs/index.php/tonos/article/view/516>
- Mendívil-Giró, J. L. (1990). On the Formal Structure of Idioms. *Lynx. Studies in Contrastive Linguistics*, 2, 97-110.
- Mendívil-Giró, J. L. (2009). Palabras con estructura externa. En E. De Miguel (Ed.), *Panorama de la lexicología* (pp. 83–113). Ariel.
- Meunier, F. y Granger, S. (2008). *Phraseology in foreign language learning and teaching*. John Benjamins.
- Mogorrón Huerta, P. (2008). Traduction et compréhension des locutions verbales. *Meta: journal des traducteurs/Meta: Translators' Journal*, 53(2), 378–406.
- Mogorrón Huerta, P. (2010). Analyse du figement et de ses possibles variations dans les constructions verbales espagnoles, *Linguisticae Investigationes*, 33(1), 86–151.
- Mogorrón Huerta, P. (2012). Explotación informática de una base de datos multilingüe de unidades fraseológicas. En M. I. González Rey (Ed.), *Unidades fraseológicas y TIC* (pp. 63–80). Instituto Cervantes.
- DUE – Moliner, M. (2010). *Diccionario de uso del español 3ª edición*. Gredos.

- Monti, J., Arcan, M., y Sangati, F. (2020). Translation asymmetries of multiword expressions in machine translation. En G. Corpas Pastor y J.-P. Colson (Eds.), *Computational Phraseology* (pp. 23-42). John Benjamins.
- Monti, J., Seretan, V., Corpas Pastor, G. y Mitkov, R. (2018). Multiword units in machine translation and technology. En R. Mitkov, J. Monti, G. Corpas Pastor y V. Seretan (Eds.), *Multiword Units in Translation and Translation Technology* (pp. 1-37). John Benjamins.
- Montoro del Arco, E. T. (2003). La manipulación humorística de las expresiones fijas como medio para la enseñanza de la fraseología del español. En J. A. Moya Corral y M. I. Montoya Ramírez (Eds.), *Variación lingüística y enseñanza de la lengua española: Actas de las VIII Jornadas sobre la Enseñanza de la Lengua Española* (pp. 225–246). Centro de Formación Continua de la Universidad de Granada.
- Montoro del Arco, E. T. (2005). Hacia una sistematización de la variabilidad fraseológica. En M. A. Pastor Millán (coord.), *Estudios lingüísticos en recuerdo del profesor Juan Martínez Marín* (pp. 125–152). Universidad de Granada.
- Montoro del Arco, E. T. (2008). El concepto de locución con casillas vacías. En C. Mellado Blanco (Eds.), *Colocaciones y fraseología en los diccionarios* (pp. 131–146). Peter Lang.
- Neunerdt, M., Trevisan, B., Reyer, M. y Mathar, R. (2013). Part-Of-Speech Tagging for Social Media Texts. En I. Gurevych, C. Biemann y T. Zesch (Eds.), *Language Processing and Knowledge in the Web. Lecture Notes in Computer Science 8105* (pp. 139–150). Springer. <https://doi.org/10.13140/2.1.2905.7284>
- Olímpio de Oliveira Silva, M. E. (2007). *Fraseografía teórica y práctica*. Peter Lang.
- Olza Moreno, I. (2011). *Corporalidad y lenguaje: La fraseología somática metalingüística del español*. Peter Lang.
- Omazić, M. (2008). Processing of idioms and idiom modifications: A view from cognitive linguistics. En S. Granger y F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 67–79). John Benjamins.
- ONTSI. (2018). *Estudio de Caracterización del Sector de las Tecnologías del Lenguaje en España 2018*. ONSI (Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información)
- OD – Oxford University Press (2021). *Oxford Dictionary*. <https://www.lexico.com>
- Palm, Ch. (1995). *Phraseologie. Eine Einführung*. Gunter Narr Verlag.

- Pedersen, V. (1997). Description and Criticism: Some Approaches to the English Translation of Hans Christian Andersen. En A. Trosborg (Ed.), *Text, Typology and Translation*. John Benjamins.
- Penadés Martínez, I. (2014). Fixación estrutural e desautomatización das locucións. *Cadernos de Fraseoloxía Galega*, 16, 273–301.
- Penadés Martínez, I. (2019). *Diccionario de locuciones idiomáticas del español actual*.
www.diccionariodilea.es
- Philip, G. (2008). Reassessing the Canon: ‘Fixed’ Phrases in General Reference Corpora. En S. Granger y F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective*. (pp. 95-108). John Benjamins Publishing Company.
- Pintado Gutiérrez, L. (2012). Fundamentos de la traducción pedagógica: traducción, pedagogía y comunicación. *Sendebarr*, 23.
- Popović, M. (2017). Comparing language related issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics*, 108(1), 209–220.
- Ramisch, C., y Villavicencio, A. (2018). Computational treatment of multiword expressions. En R. Mitkov (Ed.), *Oxford Handbook on Computational Linguistics* (2^a ed).
<https://doi.org/10.1093/oxfordhb/9780199573691.013.56>
- Rayyan, M. S. (2014). *Fraseología y lingüística informatizada. Elaboración de una base de datos electrónica contrastada arabe-español / español-arabe de fraseologismos basados en partes del cuerpo* [Tesis doctoral, Universidad de Granada].
- Recio Ariza, M. Á. (2021). Equivalences of phraseological units in corpora. En P. Mogorrón Huerta (Ed.), *Análisis multidisciplinar del fenómeno de la variación en traducción e interpretación / Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting*. *MonTI Special Issue* 6, 386–415.
<https://doi.org/10.6035/MonTI.2020.ne6.12>
- Rikters, M. y Bojar, O. (2017). Paying attention to multi-word expressions in neural machine translation. *arXiv preprint arXiv:1710.06313*.
- Roberts, R. (1998). Phraseology and Translation. En P. Fernández Nistal y J. M. Bravo Gonzalo (Eds.), *La traducción: orientaciones lingüísticas y culturales* (pp. 61–77). Servicio de Publicaciones de la Universidad de Valladolid.

- Rocha, C. M. C. (2014). *A elaboração de um repertório semibilíngue de somatismos fraseológicos do português brasileiro para aprendizes argentinos* [Tesis doctoral, Universidade Estadual Paulista Júlio de Mesquita Filho].
- Rohanian, O., Taslimipoor, S., Kouchaki, S., An Ha, L., y Mitkov, R. (2019). Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions. En J. Burstein, C. Doran, T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1*. (pp. 2692–2698). Association for Computational Linguistics.
- Ruiz Gurillo, L. (1997). *Aspectos de fraseología teórica española*. Universitat de València.
- Sánchez Ramos, M. D. M. (2019). Corpus paralelos y traducción especializada: ejemplificación de diseño, compilación y alineación de un corpus paralelo bilingüe (inglés-español) para la traducción jurídica. *Lebende Sprachen*, 64(2), 269–285. <https://doi.org/10.1515/les-2019-0015>
- Santoyo, J., y Rabadán, R. (1991). Basic Spanish Terminology for Translation Studies: a Proposal. *Meta: journal des traducteurs/Meta: Translators' Journal*, 36(1), 318–322.
- DFDEA – Seco, M., Andrés, O. y Ramos, G. (2017). *Diccionario fraseológico documentado del español actual, locuciones y modismos españoles (2ª edición)*. Aguilar
- Sennrich, R., Haddow, B., y Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.
- Sennrich, R., Haddow, B., y Birch, A. (2016). Edinburgh neural machine translation systems for WMT 16. *arXiv preprint arXiv:1606.02891*.
- Shterionov, D., Superbo, R., Nagle, P., Casanellas, L. O., O'Dowd, T., y Way, A. (2018). Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32, 217–235. <https://doi.org/10.1007/s10590-018-9220-z>
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.
- Stasimioti, M., Sosoni, V., Kermanidis, K. L., y Mouratidis, D. (2020). Machine Translation Quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs. En A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberoef, M. Nurminen, L. Marg y M. L. Forcada (Eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 441–450). European Association for Machine Translation

- Stefanowitsch, A. (2011). Konstruktionsgrammatik und Grammatiktheorie. En A. Lasch y A. Ziem (Eds.), *Konstruktionsgrammatik III: Aktuelle Fragen und Lösungsansätze* (pp. 13–27). Stauffenburg.
- Sułkowska, M. (2018). Phraséologie appliquée et ses nouvelles branches: phraséodidactique et phraséotraduction. *Romanica Cracoviensia*, 3, 159–170. doi:10.4467/20843917RC.18.017.9589
- Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-making*. Cambridge University Press.
- Tang, Y., Meng, F., Lu, Z., Li, H., y Yu, P. L. (2016). Neural machine translation with external phrase memory. *arXiv preprint arXiv:1606.01792*.
- Taylor, J. R. (2016). Cognitive linguistics. En K. Allan (Ed.), *The Routledge Handbook of Linguistics* (pp. 455–469). Routledge.
- Teliya V. N. (1975). Die Phraseologie. En B. A. Serebrennikow (Ed.), *Allgemeine Sprachwissenschaft: Die innere Struktur der Sprache* (H. Zikmund y G. Feudel, trad.) (pp. 374–429). Akademie-Verlag.
- Timofeeva, L. (2009). La desautomatización fraseológica: un recurso para crear y divertir. En J. L. Jiménez Ruiz y L. Timofeeva (Eds.), *Estudios de Lingüística: Investigaciones lingüísticas del siglo XXI* (pp. 249–271). Universidad de Alicante.
- TREC – Treccani (n.d.). *Vocabolario - Treccani*. <https://www.treccani.it/vocabolario/>
- TURENG – Tureng Dictionary and Translation Ltd. (2018). *Termbank*. <https://termbank.com>
- Vinay, J.-P. y Darbelnet, J. (1995). *Comparative Stylistics of French and English, a Methodology for Translation* (J. C. Sager y M.-J. Hamel, trad.). John Benjamins Publishing Company.
- Wang, X., Tu, Z., Xiong, D., y Zhang, M. (2017). Translating phrases in neural machine translation. *arXiv preprint arXiv:1708.01980*.
- Wotjak, G. (1986). Zur Bedeutung Ausgewählter Verbaler Phraseologismen Des Deutschen. *Zeitschrift Für Germanistik*, 7(2), 183–200.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., y Dean, J. (2016). Google's

- Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.
- Yamada, M. (2019). The impact of Google Neural Machine Translation on Post-editing by student translators. *The Journal of Specialised Translation*, 31, 87-106.
- Zaninello, A., y Birch, A. (2020). Multiword expression aware neural machine translation. En N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. 3816–3825.
- Zaretskaya, A. (2015). The use of machine translation among professional translators. En G. Corpas Pastor, A. Zaretskaya y M. Seghiri (Eds.), *Proceedings of the EXPERT Scientific and Technological Workshop* (pp. 1-12). Tradulex.
- Zaretskaya, A., Corpas Pastor, G. y Seghiri, M. (2015). Translators' requirements for translation technologies: a user survey. En G. Corpas Pastor, M. Seghiri, R. Gutiérrez Florido y M. Urbano Mendaña (Eds.), *Nuevos horizontes en los Estudios de Traducción e Interpretación (Trabajos completos)/New Horizons in Translation and Interpreting Studies (Full papers)* (pp. 247-254). Tradulex.
- Ziem, A. (2018). Construction Grammar meets Phraseology: eine Standortbestimmung. *Linguistik Online*, 90(3). <https://doi.org/10.13092/lo.90.4316>
- Zuluaga, A. (1975). La fijación fraseológica. *Thesaurus*, 30(2), pp. 223-247.
- Zuluaga, A. (1980). *Introducción al estudio de las expresiones fijas*. Peter Lang.
- Zuluaga, A. (1992). Spanisch: Phraseologie. En G. Holtus, M. Metzeltin y C. Schmitt (Eds.), *Lexikon der Romanistischen Linguistik* (pp. 125–131). De Gruyter.
- Zuluaga, A. (2001). Análisis y traducción de las UF desautomatizadas. *PhiN*, 16, 67–83. <http://www.fu-berlin.de/phin/phin16/p16t5.htm>