

EUROPHRAS 2019



Computational and Corpus-based Phraseology

**Proceedings of the Third International Conference
EUROPHRAS 2019**

(short papers, posters and MUMTTT workshop contributions)

September 25-27, 2019
Malaga, Spain

ORGANISERS

EUROPHRAS
EUROPÄISCHE GESELLSCHAFT FÜR PHRASEOLOGIE

RGCL
Research Group in Computational Linguistics

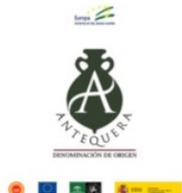
LEXYTRAD
Grupo de Investigación
(Cód. HUM-106 – Junta de Andalucía)



SPONSORS

SKETCH ENGINE

EUROPHRAS
EUROPÄISCHE GESELLSCHAFT FÜR PHRASEOLOGIE



Málaga Convention Bureau

PUBLISHER

tradulex
multilingual communication

ISBN 978-2-9701095-6-3



978-2-9701095-6-3

2019. Editions Tradulex, Geneva

©European Association for Phraseology EUROPHRAS

©University of Wolverhampton (Research Group in Computational Linguistics)

©University of Malaga (Research Group “Lexicography and Translation”)

©Association for Computational Linguistics (Bulgaria)

This document is downloadable from www.tradulex.com and
<http://rgcl.wlv.ac.uk/europhras2019/>

Editors of the Proceedings

Gloria Corpas Pastor
Ruslan Mitkov
Maria Kunilovskaya
María Araceli Losey León

Organisers:

EUROPHRAS 2019 is jointly organised by the European Association for Phraseology EUROPHRAS, the University of Malaga (Research Group in Lexicography and Translation), the University of Wolverhampton (Research Group in Computational Linguistics) and the Association for Computational Linguistics – Bulgaria.

Conference Co-Chairs:

Gloria Corpas Pastor, University of Malaga, Spain
Ruslan Mitkov, University of Wolverhampton, UK

Programme Committee:

Mariangela Albano, University Dokuz Eylül of Izmir
Verginica Barbu Mititelu, Romanian Academy Research Institute for Artificial Intelligence
Farouk Bouhadiba, University of Oran 2
Nicoletta Calzolari, Institute for Computational Linguistics
María Luisa Carrió Pastor, Polytechnic University of Valencia
Sheila Castilho, Dublin City University
Cristina Castillo Rodríguez, University of Malaga
Ken Church, Baidu
Jean-Pierre Colson, Université Catholique de Louvain
Anna Čermáková, Charles University
María Sagrario del Río Zamudio, University of Udine
Dmitrij Dobrovolskij, Russian Language Institute
Peter Ďurčo, University of St. Cyril and Methodius
Jesse Egbert, Northern Arizona University
Natalia Filatkina, University of Trier
Thierry Fontenelle, Translation Centre for the Bodies of the European Union
José Enrique Gargallo, University of Barcelona
Sylviane Granger, Université Catholique de Louvain
Kleanthes Grohmann, University of Cyprus
Miloš Jakubiček, Lexical Computing
Simon Krek, University of Ljubljana
Natalie Kübler, Paris Diderot University
Alessandro Lenci, University of Pisa
Elvira Manero, University of Murcia
Carmen Mellado Blanco, University of Santiago de Compostela
Flor Mena Martínez, University of Murcia
Pedro Mogorrón Huerta, University of Alicante
Johanna Monti, “L’Orientale” University of Naples

Sara Moze, University of Wolverhampton
Michael Oakes, University of Wolverhampton
Inés Olza, University of Navarra
Petya Osenova, Sofia University
Stéphane Patin, Paris Diderot University
Alain Polguère, University of Lorraine
Encarnación Postigo Pinazo, University of Malaga
Carlos Ramisch, Laboratoire d'Informatique Fondamentale de Marseille
Rozane Rebechi, Federal University Rio Grande do Sul
M^a Ángeles Recio Ariza, University of Salamanca
Irene Renau, The Pontifical Catholic University of Chile
Omid Rohanian, University of Wolverhampton
Ute Römer, Georgia State University
Leonor Ruiz Gurillo, University of Alicante
Agata Savary, François Rabelais University
Miriam Seghiri Domínguez, University of Malaga
Julia Sevilla Muñoz, Complutense University of Madrid
Kathrin Steyer, Institute of German Language
Joanna Szerszunowicz, University of Bialystok
Shiva Taslimipoor, University of Wolverhampton
Yukio Tono, Tokyo University of Foreign Studies
Cornelia Tschichold, Swansea University
Agnès Tutin, University of Stendhal
Aline Villavicencio, Federal University of Rio Grande do Sul and University of Essex
Tom Wasow, Stanford University
Eric Wehrli, University of Geneva
Juan Jesús Zaro Vera, University of Malaga
Michael Zock, French National Centre for Scientific Research

Additional Reviewers:

Le An Ha
Rocío Caro Quintana
Souhila Djabri
Emma Franklin
Maria Kunilovskaya
María Araceli Losey León
Encarnación Núñez
Maria Stasimioti

Invited Speakers:

Aline Villavicencio, University of Sheffield; Federal University of Rio Grande do Sul, Brazil
Miloš Jakubiček, Lexical Computing and Masaryk University, Czech Republic
Natalie Kübler, Paris Diderot University, France
Sylviane Granger, Université Catholique de Louvain, Belgium

SketchEngine Tutorial Speaker:

Miloš Jakubiček, Lexical Computing and Masaryk University, Czech Republic

Organising Committee:

University of Malaga

María Rosario Bautista Zambrana
Isabel Durán Muñoz
Javier Alejandro Fernández Sola
Mahmoud Gaber
Rut Gutiérrez Florido
Carlos Manuel Hidalgo Ternero
Francisco Javier Lima Florido
Gema Lobillo Mora
María Araceli Losey León
Desiré Martos García
Luis Carlos Marín Navarro
Juan Pascual Martínez Fernández
Míriam Pérez Carrasco
Fernando Sánchez Rodas
Anastasia Taramigou

University of Wolverhampton

Rocío Caro Quintana
Sandra Elfiky
Suman Hira
Maria Kunilovskaya
Sara Moze
Alistair Plum
Tharindu Ranasinghe Hettiarachchige
Shiva Taslimipoor

Association for Computational Linguistics (Bulgaria)

Nikolai Nikolov
Ivelina Nikolova

Table of Contents

Short papers and posters

<i>Phraseology in Learner Language: The Case of French Idioms and Collocations Translated by Italian-speaking Adult Learners</i>	
Mariangela Albano and Rosa Leandra Badalamenti	1
<i>¿Nadan en la Abundancia los Jóvenes Españoles y Alemanes? Un Estudio Empírico Sobre la Competencia Fraseológica</i>	
Isabel Andugar Andreu	11
<i>Variaciones Fraseológicas en la Terminología Médico-Farmacéutica y su Aplicación en las Traducciones EN>ES y DE>ES</i>	
Francisco Bautista	19
<i>Phrase Frames en un Corpus Oral de Alemán como Lengua Extranjera para el Turismo</i>	
María Rosario Bautista Zambrana	31
<i>Desarrollo de la Fraseología Especializada en Brasil</i>	
Cleci Regina Bevilacqua	40
<i>Orthography in Practice: Corpus-based Verification of Writing Ktetics in MWU's in Croatian</i>	
Goranka Blagus Bartolec and Ivana Matas Ivanković	46
<i>A Didactic Sequence for Phrasemes in L2 French</i>	
Maria Francesca Bonadonna and Silvia Domenica Zollo	53
<i>Procesos de Reconocimiento e Interpretación de Unidades Fraseológicas Metafóricas y Factores Influyentes</i>	
Silvia Cataldo	61
<i>Processing European Portuguese Verbal Idioms: From the Lexicon-Grammar to a Rule-based Parser</i>	
Ana Galvão, Jorge Baptista and Nuno Mamede	70
<i>Constructive Linguistics for Computational Phraseology: the Esperanto Case</i>	
Federico Gobbo	78
<i>Corpus-based Empirical Research on Resurgent Collocation beyond Existing Grammatical Rules: Make Angry/Mad as an Example</i>	
Ai Inoue	86
<i>Vivid Phrasal Idioms and the Green New Deal: Teaching Idioms to EAP Students Via Authentic Contexts</i>	
Melissa Larsen-Walker	90
<i>Extracción Terminológica Basada en Corpus Para la Traducción de Fichas Técnicas de Impresoras 3D</i>	
Ángela Luque Giraldez and Míriam Seghiri	99

<i>Corpus Analysis of Complex Names with Common Nouns in Croatian</i>	
Ivana Matas Ivanković and Goranka Blagus Bartolec	106
<i>Fixed Phrases in Language of International Law: A Problem of Translating Latin Formulaic Expressions into Farsi</i>	
Seyed Mohammad Hossein Mirzadeh	114
<i>On the Impact of (Il)literacy on L2 Italian Acquisition of Unaccompanied Foreign Minors</i>	
Castrenze Nigrelli	118
<i>Improving Textual Competence in a Second Language Initial Literacy Classroom</i>	
Castrenze Nigrelli	126
<i>Multiword Terms and Machine Translation</i>	
Serge Potemkin	133
<i>Towards a Cross-linguistic Study of Phraseology across Specialized Genres</i>	
Ana Roldan-Riejos and Lukasz Grabowski	140
Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2019) workshop contributions	
<i>Multi-word Units in Machine Translation: why the Tip of the Iceberg Remains Problematic – and a Tentative Corpus-driven Solution</i>	
Jean-Pierre Colson	145
<i>Automatic Term Extraction from Turkish to English Medical Corpus</i>	
Gokhan Dogru	157
<i>Lexicographic Criteria for Selecting Multiword Units for MT Lexicons</i>	
Jack Halpern	167
<i>Multiword Expressions Under the Microscope</i>	
Aline Villavicencio	181

Phraseology in Learner Language: The Case of French Idioms and Collocations Translated by Italian-speaking Adult Learners

Mariangela Albano¹[0000-0001-7157-1482] and Rosa Leandra Badalamenti²

¹ Dokuz Eylül University & University Sorbonne Nouvelle, Paris
albanomariangela@gmail.com

² University of Palermo
rosaleandra.bk@gmail.com

Abstract. This paper focuses on the treatment of French idioms and collocations by Italian-speaking adult learners of French as a foreign language. Since allophones do not have access to the figurative dimension that characterizes the frozen expressions of a language, mastery of them is very difficult and is usually only reached at an advanced stage of acquisition. The purpose of our study is to provide a contribution to the analysis of the semantic processing that takes place at the cognitive level in their interpretation. These processes will be approached from two perspectives of analysis: one pertaining to the acquisitional linguistics, aiming to examine the approaches and interpretative strategies put in place by foreign learners when confronted with an idiomatic expression, and the other pertaining to cognitive linguistics and aimed at the analysis of cognitive operations that create semantic networks of a metaphorical, metonymic and analogical nature.

Keywords: Idiom, Collocation, Acquisition of foreign languages, Analogy, Metaphor, Metonymy

1 Introduction

In this study, we analyze the interpretation strategies of 15 frozen expressions (henceforth FE)¹ by advanced Italian-speaking learners of French as a foreign language (henceforth FFL). To analyze the processes of foreign language appropriation, different interdependent elements must be taken into account, such as the socio-cultural context in which the appropriation takes place, the learner's previous knowledge, his

1 The analyzed FE are: 1) poser un lapin, 2) se regarder le nombril, 3) yeux de biche, 4) tomber dans les pommes, 5) avoir la grosse tête, 6) filer à l'anglaise, 7) être un ours mal léché, 8) mettre de l'eau dans son vin, 9) mettre son grain de sel, 10) se mettre sur son trente et un, 11) casser sa pipe, 12) peigner la girafe, 13) parler polémique ; 14) temps de cochon, 15) avoir du blé.

level of metalinguistic awareness, or cognitive skills, developed during the learning path.

As far as FEs are concerned, the learner faces a double semantic opacity: the first is inherent to the approach of a foreign language (hence FL), since the learner has to deal with words structures and expressions that are unknown to him/her; the second is that FE requires the ability to make a transition between a literal level and a figurative level.

The attitude of the learner, his level of tolerance of linguistic ambiguity, his ability to fill the gaps are essential factors in the treatment of IE, treatment that we analyze here. The task of the learners was to interpret and translate texts that contained one or more word-for-word non-translatable FE into Italian.

Our approach is situated between cognitive linguistics and acquisitional linguistics: on the one hand, the cognitive approach can be useful for analyzing the cognitive projections activated during the processes of comprehension, translation and motivation of FE; on the other hand, the acquisition approach leads us to understand the processes of identification and isolation and, thanks to the analysis of the linguistic input processing operations implemented by the learners, can lead to possible didactic applications in the teaching of FEs in the FFL class².

1.1 Lexicalization, Projection, and Analogy

The study of lexicalization leads us to take into consideration a set of expressions in which morphosyntactic and lexical variations are limited. The FEs represents a unit that reveals a memory and psychological constant still perceived by the speakers of a language (Cacciari, 2001). If we base ourselves on the studies that concern the FEs, we can define the level of lexicalization based on seven criteria: polylexicality (G. Gross, 1996); non-compositionality (Wray, 2002); grammatical or syntactic block (Hudson, 1998; Gross, 1996), semantic block (Ibid.: 154), conventionality for which the set of words gives perception of unity (Nunberg et al., 1994: 493), block of synonymic paradigms (Shapira, 1999: 8-9; Gross, 1996) and semantic opacity (Gross, 1996).

In this study we have chosen to propose to the learners idioms (i.e. *se regarder le nombril*), and this for their high level of figurativity (Hudson, 1998; Moon, 1998: 19-25; Norrick, 1985: 72; Brinton and Traugott, 2005; Svensson, 2004), some collocations (Burger, 2007; Benson, 1985), and some collocations in which only a segment possesses the linguistic status of idiom (i.e. *yeux DE BICHE*; *être un OURS MAL LÉCHÉ*).

The criteria for ranking FEs are not the only area on which phraseology studies have been concentrated: there has been increasing interest in, for example, cognitive

2 Mariangela Albano wrote: 1.Introduction, 1.1Lexicalization, projection and analogy, 2.3Activation of conceptual connections: metaphor, metonymy and analogy, 3.Conclusions; Rosa Leandra Badalamenti wrote: 1.2 FE in the acquisition of foreign languages, 1.3Methodology, 2.Analysis of data, 2.1Reference to context, 2.2. Analogies with L1 or other known languages, 3.Conclusions.

analyzes. In this sense, the semantics of idioms aims to search for the motivation of IEs, to study the function of mental images in the use of these, or to explore the cultural phenomena that motivate them (Burger, 2007: 790; Dobrovolskij and Piirainen, 2005).

To understand the cognitive aspects of IEs, the semantics of idioms use the theory of conceptual metaphor (Lakoff and Johnson, 1980) to analyze the cognitive processes underlying IE treatment. With regard to the metaphorical projection, the cognitive process takes place between a source domain and a target domain; as far as the metonymic projection is concerned, the process takes place within the same conceptual domain. In this study, we approach cognitive projection to seek to understand the translation and interpretation processes of learner-initiated IEs. It is also necessary to analyze the data from the analogy, a cognitive tool that allows learners to «group together in the same class or category» perceived «entities, consciously or not, as similar» (Monneret, 2003).

1.2 FEs in the Acquisition of Foreign Languages

According to Hudson, phraseology responds to principles of economy. Speakers repeat the expressions they have already heard instead of creating new ones each time. Indeed, since these are stored as a single lexical unit, they require much less processing effort at the time of production and their use increases the speed of speech and reduces the frequency of breaks.

However, a good knowledge of the FE would, according to Wood (2006) and Pawley & Syder (2000: 195), the verbal fluidity «it is the store of memorized constructions and expressions, more than anything, that is the key to nativelike fluency». In addition, it seems that the use of FEs would allow non-native speakers (henceforth NNS) to express themselves in a more idiomatic and therefore more natural way «formulaic sequences used by native speakers are not easy for learners to identify and master, and [...] their absence greatly contributes to learners not sounding idiomatic» (Wray, 2002: 176).

Figurative language is rooted in everyday life and makes use of important factors in the process of appropriation of a FL, such as imagination and affectivity, and thus makes it possible to develop language expressivity (Ruiz Quemoun, 2007). In particular, idioms reflect the culture shared by a linguistic community and are therefore fundamental to getting closer to the culture of the target language.

In the acquisition of L1, the appropriation processes generally start from a holistic approach and go towards an analytical approach. In adult learners of a FL or L2, the opposite is true: the adult learner starts from an analytical approach, gradually evolving towards memory storage and holistic treatment of FE (Wray, 2002).

This has already been noted by Gibbs (1986), who emphasizes that native speakers (henceforth NS) do not activate the literal meaning of FEs unless the figurative meaning is irrelevant. On the contrary, NNSs try to activate a figurative meaning only when the literal meaning is perceived as irrelevant.

In the case of FL learning, another important factor that comes into play is the learning context. This is fundamental in that it determines the amount and type of

input to which learners are subjected, as well as the types of interactions in which they participate. In the classroom, learners tend to adopt analytical approaches, especially if they have been subjected to a metalinguistic teaching of the grammar of the FL.

As we have seen, FEs have an ontological and cultural motivation. For foreign learners, the cultural sphere is not directly accessible. From this perspective, learning for a FL must also be distinguished from second-language learning, since in the second case learners have more opportunities to contact the target language. In FL learning, idioms can only be learned in the classroom, and the role of teachers and methods adopted becomes paramount. FEs are therefore an essential component of communication, acquisition and idiomaticity in FL. This is why it is useful to question the factors that may favor, or on the contrary, disadvantage, the development of learners.

1.3 Methodology

For our study, we chose 10 advanced learners of FFL in a university context. Students of our study have all studied French for at least 8 years, and they all have a master's degree in foreign languages. They have all been exposed to formal teachings of French grammar, and therefore tend to adopt analytical approaches to texts. We selected 15 FEs from dictionaries and data in short texts of different kinds (magazine extracts, daily newspapers, online forum, novels, and so on). We selected FEs which are not directly translatable into Italian. We asked the learners to translate or interpret the texts orally. Oral productions have been recorded and transcribed.

First, we focused on the isolation processes of FE within the text. We observed that, in most cases, learners were able to identify them even when they had never met them before. Subsequently, we asked about strategies and approaches taken by students to interpret FEs in a consistent way. We also considered the possible use of L1 in the treatment of FEs. In some cases, learners used their mother tongue or other known languages to try to interpret figurative expressions, but we also observed that, in most cases, they were able to avoid interference, demonstrating a high level of awareness of the cultural motivation of the chosen FEs. In this perspective, we questioned the processes of metaphorical or metonymic interpretation, in which learners abandoned their cultural sphere to engage in an ontological sphere, trying to reactivate symbolic processes from the literal meaning of the components of the FE.

2 Data Analysis

Interpretative approaches are based on an analytical approach that allows students to isolate FE within the text. Given the semantic opacity of the lexical unit, in most cases the isolation is based on processes of semantic relevance: in the given context, the lexical unit could not keep its literal and compositional meaning that was not relevant. Depending on the degree of accessibility of the context and the relevance of the FE speakers have put together or combined different interpretative strategies. If the con-

text provides different or contradictory interpretative possibilities, the foreign speaker also uses other interpretative approaches, such as analogies with L1 and /or other known languages or even metonymic, analogical and metaphorical processes.

2.1 Context Reference

In our study, FEs were proposed within texts of different typologies. By context, we try to show the representations activated by the readers during the reading of the text.

NSs confronted with a figurative use of language, as in the case of unconventional metonymies, activate more interpretative hypotheses based on principles of relevance (Sperber and Wilson, 1986).

In one of the proposed texts, learners were asked to interpret the idiom *poser un lapin*, engl. stand somebody up, which appeared in a comics. The degree of semantic opacity and irrelevance to the context proved very high, and in most cases, after questioning the possible metaphorical values to be given to the expression, the students ended up with interpret from the image.

In another example, with the idiom *se regarder le nombril*, engl. to look on own navel, we see that the reader first assesses relevance with the context, before proceeding to a metaphorical interpretation of the FE.

2.2 Analogies with L1 or Other Known Languages

For foreign learners, L1 represents an inevitable interpretative tool. In the context of our study, we will not consider interference as an unconscious transfer of features from one language to another, but we will focus on the use of known languages as a conscious strategy in interpretation of the text.

In collocation n. 3 *yeux de biche*, the interpretation based on analogies with L1 is superior to contextual information and we could see two interpretative phases based on analogies with Italian.

At first, the learner selects the referent from a phonetic analogy between the French word *biche* [bif] and the Italian word *biscia* [bifa]. Subsequently, the interpretative choice is semantically motivated by metaphorical processes aiming at analogically transferring the characteristics of the animal *biscia* (engl.: grass snake) to the character of the comics, thus giving fr. *yeux de salope*, engl. slut eyes. However, the selected referent being irrelevant to the context, the learner subsequently uses another analogy with the L1 and in particular between the French collocation *yeux de biche*, engl. dee eyes, it. *occhi da cerbiatta* having the same meaning. We thus observe the transfer from a phonetic analogy to a semantic analogy allowing to advance an interpretative hypothesis more relevant to the context.

L1 has a prominent role in the interpretation of IEs, but learners do not just tap into it. In example n. 4 *tomber dans les pommes*, a student uses the Sicilian dialect. In particular, this student uses a Sicilian expression that has a certain degree of formal similarity with the French IE: *tomber dans les pommes*, engl. to fall in the apples and *cariri comu a piru*, engl. to fall like a peer share the same verbal base (*tomber* and *cariri*) and in both expressions the argument of the verb is a fruit, hence the use of the

Sicilian expression which is however deceptive compared to the signified of the French IE. In this example, we see how the reader does not limit himself to drawing on his prior knowledge, but also tries to give, through analogical methods, a semantic motivation to his interpretative choice.

2.3 Activation of Conceptual Connections: Metaphor, Metonymy and Analogy

It should be emphasized here that during the process of interpretation of idioms and collocations, students reactivate the semantics of these expressions through cognitive operations. The expression *ne pas avoir la grosse tête*, engl. not having the big head is interpreted correctly in the majority of the analyzed cases. However, some of the students translate the expression as being stupid. In fact, the inferences made by the learners show a stereotypical cultural evaluation according to which the footballers are stupid.

This evaluation is due, first, to a metonymy since the head is not simply a part of the body but represents the center of intelligence according to the report container-content. Then, they subconsciously resort to conceptual metaphors *more is up*, the head is a container and the greater the amount of an entity increases in a container and the more positively the container is marked (Lakoff and Johnson, 1980), and they say that the bigger the head, the more intelligent the person involved is.

An interesting case is represented by the interpretation of FE *se regarder le nombril*, engl. looking at our own navel. First, the students translate IE with *stare with the mani in mano* (lit.: stay with hands in hand, and do nothing) using an analogy with a FE in L1. Then they come closer to an interpretative hypothesis of a metaphorical nature explaining the motivation of the FE through an embodied perspective. Indeed, students imagine the position of a person who is looking at the navel and they highlight the effects that such a position may have on the actions of the subject, such as the impossibility of moving, the obligation to remain bent over oneself to look at the navel, the obligation to remain extended and maintain a horizontal position. They find, therefore, that such a position prevents movement and forces us to do nothing.

The interpretation of the FE *mettre son grain de sel* (engl. put his grain of salt) is translated by some learners by *metterci del suo* (fig. give his/her opinion). Learners come to this translation by reflecting on the effects of salt in an injury and claim that salt burns. This allows them to reach, by analogical methods, the Italian expression *mettere fuoco* (engl. set fire) and connect analogically the effects of salt and fire. Students focus, then, on the effects of fire in a fire and they project metaphorically the idea that every element that burst into a foreign situation creates an alteration and that, in the same way, each person who gives his opinion in a foreign situation creates an alteration.

3 Conclusion

Examples taken into account allows us to observe the interpretation processes developed by students to analyze idioms and collocations. We have marked that the activation of conceptual connections in the processing of FE does not take place from a single cognitive tool, but that students tend to juxtapose cognitive operations of a metaphorical, metonymic and analogical nature with the use of cognitive tools, context and analogies with known languages.

In some cases, students have focused their attention on the specificity of the conceptual relationships between the source domain and the target domain of FE and they obtained new inferences, showing us how primary conceptual metaphors work and operate in the dynamics of imaginative nature. This study has thus shown that students get own motivation of an idiom or a collocation by a path where the hypotheses and the remarks made during the treatment help them to «deep the notion of image from itself» (Monneret, 2004: 105).

In Foreign Language learning, learners deal with a cultural and connotative lag when trying to interpret a FE. As in the natural contexts of communication, they are forced to interpret the FE to interpret the entire message.

In addition, texts present different levels of semantic opacity since the foreign learner is already confronted with an input whose degree of ambiguity is high.

In our examples, the semantic opacity of FEs is usually solved by the use of context. If the latter does not lead to coherent interpretations, learners try to use their language skills in L1 or other known languages or to reactivate conceptual, metaphorical or metonymic processes.

Learners are well aware that frozen units are culturally motivated. They try to activate all of their linguistic, cultural and historical knowledge of the target language in order to translate the FE, and in some cases they push their analyzes to an ontological level.

Kövecses and Szabo (1996) argue that the presence of underlying cognitive metaphors in the mind is not sufficient to activate the use of FEs in foreign learners. According to Marquez (2007: 28), it is necessary to show learners that a great deal of phraseological expression is metaphorically motivated.

Explanation of the underlying metaphorical projections would be useful for the acquisition of equivalences among the conceptual frames that activate the FE.

Our study confirms that the analysis and the explanation of the underlying metaphors can indeed prove to be an asset in the appropriation of idioms and collocations. The effort made to negotiate their meaning, in relation to the context but also to the units that compose them, seems to be able to facilitate their storage in memory and processing.

References

1. Benson, M.: Collocations and idioms. In R. Ilson (Ed.), *Dictionaries, Lexicography and Language Learning VIII* (pp. 61-68). Oxford : Pergamon (1985).
2. Brinton, L. J. & Traugott E. C.: *Lexicalization and Language Change*. Cambridge : CUP (2005).
3. Burger, H. (Ed.): *Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung*. Berlin : Walter de Gruyter (2007).
4. Cacciari, C.: *Psicologia del linguaggio*. Bologna : Il Mulino (2001).
5. Dobrovól'skij, D.: Idiome aus kognitiver Sicht. In K. Steyer (Ed.), *Wortverbindungen - mehr oder weniger fest* (pp. 117-143). Berlin : Walter de Gruyter (2004).
6. Dobrovól'skij, D. & Piirainen, E.: *Figurative language: Cross-cultural and cross-linguistic perspectives*. Amsterdam : Elsevier (2005).
7. Fontenelle, Th.: What on earth are collocations?. *English today: the international review of the English language*, 10. 4 (40), 42-48 (1994).
8. Gibbs, R. W. Jr.: Skating on thin ice: Literal meaning and understanding idioms in conversation. *Discourse Processes*, 7, 17-30 (1986).
9. González, R. (Ed.) : *Les expressions figées en didactique des langues étrangères*. Fernelmont : E.M.E. (2007).
10. Gross, G. : *Les expressions figées en français; noms composés et autres locutions*. Paris : Éditions Ophrys (1996).
11. Hudson, J.: *Perspectives on fixedness: applied and theoretical*. Lund : Lund University Press (1998).
12. Kleiber, G.: Les proverbes: des dénominations d'un type 'très très spécial'. *Langue française*, 123, 52-69 (1999).
13. Kövecses, Z. & Szabó P. : *Idioms: a view from cognitive linguistics*. *Applied Linguistics*, 17, 326-355 (1996).
14. Lakoff, G. & Johnson, M.: *Metaphors we live by*. Chicago : The University of Chicago Press (1980).
15. Martin, R. : Sur les facteurs du figement lexical. In M. Martins-Baltar (Ed.), *La locution entre langue et usages* (pp. 291-305). Fontenay Saint Cloud : ENS Éditions (1997).
16. Mel'čuk, I. : La phraséologie et son rôle dans l'enseignement/apprentissage d'une langue étrangère. *Étude de Linguistique Appliquée*, 92, 82-113 (1993).
17. Monneret, Ph. : Le sens du signifiant. Implications linguistiques et cognitives de la motivation. Paris : Champion (2003).
18. Monneret, Ph. : *Essais de linguistique analogique*. Dijon : A.B.E.L.L. (2004).
19. Moon, R.: *Fixed expressions and idioms in English, a corpus-based approach*. Oxford : Clarendon Press (1998).
20. Norrick, N. R.: *How proverbs mean: semantic studies in English proverbs*. Berlin : Mouton (1985).
21. Nunberg, G.; Sag, I. A. & Wasow, T. (Eds): *Idioms*. *Language*, 70, 3, 491-538. Washington DC: Linguistic Society of America (1994).

22. Pawley, A. & Syder, F. H.: The One-Clause-at-a-Time Hypothesis. In H. Riggenbach (Ed.), *Perspectives on Fluency* (pp. 163-199). University of Michigan Press (2000).
23. Schapira, Ch. : *Les stéréotypes en français : proverbes et autres formules*. Paris : Éditions Ophrys (1999).
24. Sperber, D. & Wilson, D.: *La pertinence: communication et cognition*. Paris : Éditions de Minuit (1986).
25. Svensson, M. H. : *Critères de figement. L'identification des expressions figées en français contemporain*. Umeå : Tommy Sund (2004).
26. Wood, D.: *Uses and Functions of Formulaic Sequences in Second Language Speech: An Exploration of the Foundations of Fluency*. *The Canadian Modern Language Review*, 63, 1, 13-33 (2006).
27. Wray, A.: *Formulaic Language and the Lexicon*. Cambridge : Cambridge University Press (2002).

Dictionaries

28. Imbs, P. (Ed.): *Trésor de la langue française. Dictionnaire de la langue du XIX et du XX siècle (1789-1960). Tome troisième*. Paris : Éditions du Centre National de la Recherche Scientifique (1974).
29. Imbs, P. (Ed.) : *Trésor de la langue française. Dictionnaire de la langue du XIX et du XX siècle (1789-1960). Tome septième*. Paris : Éditions du Centre National de la Recherche Scientifique (1979).
30. Gorcy, G. (Ed.) : *Trésor de la langue française. Dictionnaire de la langue du XIX et du XX siècle (1789-1960). Tome seizième*. Paris : Gallimard (1994).
31. Pottier, B. (Ed.) : *Trésor de la langue française. Dictionnaire de la langue du XIX et du XX siècle (1789-1960). Tome dixième*. Paris : Éditions du Centre National de la Recherche Scientifique (1983).
32. Pottier, B. (Ed.) : *Trésor de la langue française. Dictionnaire de la langue du XIX et du XX siècle (1789-1960). Tome onzième*. Paris : Gallimard (1985).
33. Robert, P. (Ed.) *Le nouveau petit Robert de la langue française 2007: dictionnaire alphabétique et analogique de la langue française*. Paris : Le Robert (2007).

Sitography

34. BiblioBabil: *Mon Ipad a cassé sa pipe*. <http://bibliobabil.com/2011/05/16/mon-ipad-a-casse-sa-pipe/>, last accessed 2012/03/15 (2011).
35. Commune de Bellonne : *L'ancêtre du G.P.S. à Bellonne*. <http://www.bellonne.fr/pageLibre000103d8.html>, last accessed 2012/03/15 (2012).
36. Futura Forum : *Que se passe-t-il quand on tombe dans les pommes ?* <http://forums.futura-sciences.com/sante-medecine-generale/156930-se-passe-t-on-tombe-pommes.html>, last accessed 2012/03/15 (2004).

37. L'île de Crète : Voyage en images. Forum des voyageurs. Re : Athena Palace.
<http://www.ile-de-crete.com/forums/read.php?1,11681,12095>, last accessed 2012/03/15 (2006).
38. Le Figaro.fr. : Abidal : «Au Barça, on ne se regarde pas le nombril».
<http://www.lefigaro.fr/sport/2011/03/07/02001-20110307ARTFIG00665-abidal-au-barca-on-ne-se-regarde-pas-le-nombril.php>, last accessed 2012/03/15 (2011).
39. Le Figaro.fr : Les républicains se cherchent un champion contre Obama.
<http://www.lefigaro.fr/international/2011/05/06/01003-20110506ARTFIG00615-les-republicains-se-cherchent-un-champion-contre-obama.php>, last accessed 2012/03/15 (2011).
40. Le Figaro.fr. : François Hollande en tête-à-tête avec la reine Elizabeth.
<http://www.lefigaro.fr/international/2012/07/10/01003-20120710ARTFIG00591-hollande-a-discute-en-tete-a-tete-avec-la-reine-elizabeth.php?page=&pagination=17>, last accessed 2012/07/25 (2012).
41. Plurielles.fr. : Beauté homme : quels soins basiques pour ses premiers pas ?
<http://www.plurielles.fr/beaute/soins/beaute-homme-quels-soins-basiques-pour-ses-premiers-pas-7219330-402.html>, last accessed 2012/06/15 (2012).

¿Nadan en la Abundancia **los Jóvenes Españoles y Alemanes?** **Un Estudio Empírico Sobre la Competencia** **Fraseológica**

Isabel Andúgar Andreu

Universitat Jaume I, Castelló de la Plana 12071, España
isandugar@gmail.com

Abstract. La presente comunicación se enmarca en el ámbito de la fraseología contrastiva español-alemán. Partiendo de la noción de equivalencia entre *unidades fraseológicas* (UF), surge la necesidad de disponer de datos reales de competencia de los jóvenes de ambas comunidades lingüísticas. Para ello se realiza un estudio empírico, utilizando un corpus de expresiones fraseológicas equivalentes como *Coger el toro por los cuernos* ≈ *den Stier bei den Hörnern packen*. El proceso de investigación consta de una revisión previa de diferentes aspectos sobre la disciplina y del concepto de equivalencia, y posteriormente se sientan las bases metodológicas de la investigación empírica. Esta investigación se desarrolla en diferentes fases: el planteamiento de hipótesis, la selección del corpus de UF, la redacción del cuestionario, la selección de informantes, la recogida de datos y análisis de los mismos. Hay que destacar que, en la presente investigación se ha confeccionado un corpus de UF *ad hoc*, conocido y utilizado por los hablantes nativos de ambas lenguas. Una vez completada la investigación, se procede al análisis cuantitativo y cualitativo de los datos mediante herramientas estadísticas, del que se obtienen datos esclarecedores sobre las diferencias, pero también inesperadas similitudes entre ambas poblaciones. El desarrollo de este tipo de investigaciones es un instrumento muy valioso para establecer posibles mínimos fraseológicos en la didáctica de las lenguas, tanto de las maternas como de las lenguas extranjeras, en la traducción, y por supuesto en la elaboración de herramientas como diccionarios.

Keywords: Fraseología contrastiva español-alemán, Unidad fraseológica (UF), Competencia, Corpus, Mínimo fraseológico

1 Introducción

1.1 Planteamiento de la investigación

La presente comunicación ofrece un análisis de la competencia fraseológica de los jóvenes de las dos comunidades de habla, la española y la alemana, en su lengua materna.

Es generalmente asumido que la competencia lingüística es la disponibilidad activa o pasiva de determinadas estructuras que poseen los hablantes de una lengua: es decir, no solo qué unidades lingüísticas son capaces de usar (*competencia activa*), sino también aquellas que conocen, aunque no suelen utilizar (*competencia pasiva*).¹ En el caso que nos ocupa, el objetivo es realizar un estudio empírico que nos arroje datos especialmente cuantitativos, pero no sólo, acerca del conocimiento de expresiones fraseológicas que poseen los jóvenes.

Las expresiones fraseológicas en cualquiera de sus formas (*colocaciones, locuciones, refranes*, etc.) son un componente importante de las lenguas, no sólo por su frecuencia de aparición, sino también por su complejidad, pues a diferencia de lo que ocurre con las *palabras*, la consulta de las *unidades fraseológicas* (UF) en los diccionarios no resulta tarea fácil. En definitiva, si la correcta utilización de la fraseología no es sencilla para los hablantes nativos de una lengua, es mucho más difícil para quienes intentan dominar el sistema fraseológico de una segunda lengua, pues no sólo se trata de esquemas lingüísticos diferentes, sino también de referentes culturales distintos.

Si bien es frecuente encontrar estudios lingüísticos que investigan la competencia lingüística en cuanto a la disponibilidad léxica en general, tanto en español como en alemán, es decir, el caudal de vocabulario que activa o pasivamente conoce un grupo de hablantes, no podemos decir lo mismo de la competencia o disponibilidad fraseológica en general, que cuenta con un número de trabajos especializados sensiblemente menor.²

Tal vez esto se deba a que la misma investigación fraseológica es relativamente joven comparada con la lingüística en general, pues debemos remontarnos apenas a las últimas décadas para presenciar un desarrollo importante de los estudios fraseológicos, no solo desde una perspectiva teórica, sino también aplicada, de manera que actualmente podemos considerar la situación de la disciplina bastante consolidada.

Aun así, la competencia fraseológica ha sido poco estudiada empíricamente, si exceptuamos los refranes. También es escasa la investigación de la competencia que relaciona los sistemas fraseológicos de lenguas diferentes.³ Precisamente esta ausencia de trabajos al respecto ha motivado la presente investigación, en la que se intenta dilucidar qué competencia tienen los jóvenes en su lengua materna a partir de un corpus fraseológico equivalente. Nos interesa especialmente averiguar si: a) la competencia es similar en ambas lenguas contrastadas, y b) algunos factores sociológicos pueden condicionar esa competencia.

1 Yaguello (1983: 79-80) se refiere a la *competencia pasiva* como la «aptitud para reconocer», y la *competencia activa* a la «capacidad de reproducir».

2 No obstante, se han realizado algunos estudios en lo que respecta a refranes, por ejemplo, las investigaciones realizadas por Sevilla para el español (Sevilla, 2010).

3 Los estudios contrastivos de fraseología abordan generalmente aspectos sobre la lengua literaria, el análisis de corpus y sus traducciones (López Roig, 2002), o las equivalencias entre las lenguas (Tarnowska, 2004).

1.2 Metodología

La investigación aúna las vertientes lingüística y sociolingüística, así pues su base metodológica está determinada por preceptos de las dos disciplinas, como no podía ser de otra manera. Las fases de la investigación se desarrollan atendiendo a la secuencia siguiente: la elaboración del corpus de UF equivalente, la redacción del cuestionario, la selección de informantes y la recogida de datos.

Elaboración del corpus equivalente. Además de cumplir con el criterio previo de equivalencia en la comparación interlingüística español-alemán, se considera que las UF integradas en el corpus deben cumplir otros requisitos, como *representatividad*, pues deben ser UF conocidas ampliamente por los hablantes nativos jóvenes en cada una de las muestras, y *adecuación*, que hace referencia puesto que la forma externa de la UF no venía dada con antelación en un texto fijado, era necesario una verificación de la misma para garantizar el rigor de la investigación. La representatividad de las UF se decidió a partir de estudios previos, en los que se verificó un grado de conocimiento superior al 50 %.⁴ Además, la adecuación de la forma elegida se realiza a partir del cotejo de diferentes fuentes fraseográficas, comprobando que la variante elegida por nosotros fuera la forma más usual o ‘canónica’, según denominación de Conca y Guia (2014).

El corpus de UF es el siguiente:

- UF1 poner los cuernos ≈ Hörner aufsetzen*
- UF2 poner la mano en el fuego ≈ die Hand ins Feuer legen*
- UF3 como el perro y el gato ≈ wie Hund und Katze*
- UF4 no tener pelos en la lengua ≈ kein Blatt vor den Mund haben*
- UF5 ponerse los pelos de punta ≈ jemandem stehen die Haare zu Berge*
- UF6 descubrir la pólvora ≈ das Pulver erfunden haben*
- UF7 tirar la casa por la ventana ≈ das Geld zum Fenster hinauswerfen*
- UF8 sonar a chino ≈ Spanisch vorkommen*
- UF9 ser todo oídos ≈ ganz Ohr sein*
- UF10 hacerse la boca agua ≈ das Wasser im Mund zusammenlaufen*
- UF11 llamar a las cosas por su nombre ≈ die Dinge bei ihrem Namen nennen*
- UF12 echar leña al fuego ≈ Öl ins Feuer giessen*
- UF13 pagar con la misma moneda ≈ mit gleicher Münze heimzahlen*
- UF14 no tener pies ni cabeza ≈ weder Hand noch Fuss haben*
- UF15 matar dos pájaros de un tiro ≈ zwei Fliegen mit einer Klappe schlagen*
- UF16 faltar un tornillo ≈ nicht alle Tassen im Schrank haben*
- UF17 amor a primera vista ≈ Liebe auf den ersten Blick*

⁴ Tarnovska (2004) ha considerado también este porcentaje como un buen nivel de competencia paremiológica. Al igual que ocurre en nuestro trabajo, Tarnovska desarrolló su investigación empírica por medio de encuestas, a fin de delimitar un refranero básico español (unas 250 paremias). Este trabajo lo completa con el análisis contrastivo de otras unidades fraseológicas equivalentes en ruso y ucraniano. La finalidad de este trabajo consiste en acercar las paremias al estudiante de español como lengua extranjera, estableciendo un mínimo paremiológico español.

UF18 tener algo en la punta de la lengua≈ etwas auf der Zunge liegen
 UF19 nadar en la abundancia≈ im Geld schwimmen
 UF20 perder hasta la camisa≈ bis aufs Hemd ausziehen
 UF21 meter la cuchara≈ seinen Senf dazu geben
 UF22 estar en el séptimo cielo≈ im siebten Himmel sein
 UF23 ser la oveja negra de la familia≈ das schwarze Schaf der Familie sein
 UF24 ser un lobo con la piel de cordero≈ ein Wolf im Schafspelz sein
 UF25 tener sangre azul≈ blaues Blut haben
 UF26 no ser carne ni pescado≈ weder Fisch noch Fleisch
 UF27 nadar a contracorriente≈ gegen den Strom schwimmen
 UF28 saber dónde le aprieta el zapato a alguien≈ wissen, wo einen der Schuh drückt
 UF29 tender un puente de plata≈ eine goldene Brücke bauen
 UF30 perder el hilo≈ die Faden verlieren
 UF31 jugar con las cartas boca arriba≈ mit offenen Karten spielen
 UF32 llevar los pantalones≈ die Hosen anhaben
 UF33 hacer la corte a alguien≈ jemandem den Hof machen
 UF34 llevar una venda en los ojos≈ ein Brett vor dem Kopf haben
 UF35 ser el cuento de la lechera≈ eine Milchmädchenrechnung sein
 UF36 estar en boca de todos≈ in aller Munde sein
 UF37 estar hasta las narices≈ die Nase voll haben
 UF38 construir sobre arena≈ auf Sand bauen
 UF39 poner la zancadilla a alguien≈ jemandem ein Bein stellen
 UF40 ponerle los nervios de punta a alguien≈ jemandem auf die Nerven gehen
 UF41 hacer una montaña de un grano de arena≈ aus einer Mücke einen Elefanten machen
 UF42 tener gato encerrado≈ die Katze im Sack kaufen
 UF43 levantarse con el pie izquierdo≈ mit dem linken Fuß zuerst aufstehen
 UF44 sacar las castañas del fuego a alguien≈ jemandem die Kastanien aus dem Feuer holen
 UF45 coger el toro por los cuernos≈ den Stier bei den Hörnern packen
 UF46 ser pobre como una rata≈ arm wie ein Kirchenmaus sein
 UF47 echar tierra en los ojos a alguien≈ jemandem Sand in die Augen streuen
 UF48 encontrar un pelo en la sopa≈ ein Haar in der Suppe finden
 UF49 meter las narices en todo≈ die Nase in alles stecken
 UF50 enseñarle los dientes a alguien≈ jemandem die Zähne zeigen

Redacción del cuestionario. Los datos de la investigación son recogidos en un cuestionario elaborado *ad hoc*, con dos partes diferenciadas, una lingüística, en la que se presentan cincuenta expresiones equivalentes en ambas lenguas y se investiga la competencia que poseen los hablantes, y otra sociológica, en la que se obtienen datos de los sujetos por medio de un breve cuestionario sociológico.

Selección de informantes. La selección de los informantes no se hace de forma manipulada respecto a ninguno de los factores sociales, aunque sí es necesario

establecer premisas que determinen un punto de partida básico, como es la edad. Para este fin, se han elegido niveles educativos similares en ambas comunidades lingüísticas de manera que al menos estuviera asegurado el rango similar de edades. En ambos casos son niveles de la enseñanza pre-universitaria, entre los 14 y los 17 años. Las muestras se tomaron en Castellón y Hamburgo.

La distribución de los informantes por variables edad y sexo se muestran en las tablas siguientes.

Tabla 1. Edad de los informantes

EDAD (años)	14	15	16	17	TOTAL INF
ESPAÑOL	14	15	15	1	45
ALEMÁN	16	16	5	8	45

Tabla 2. Sexo de los informantes

SEXO	MUJERES	HOMBRES	TOTAL INF
ESPAÑOL	21	24	45
ALEMÁN	25	20	45

Recogida de datos. En el proceso de recogida de datos participaron profesores de lenguas y cada una de las partes del cuestionario tuvo un tiempo limitado.

Posteriormente, se procede al análisis con la ayuda de herramientas estadísticas que facilitan la interpretación de los resultados obtenidos, tanto cuantitativa como cualitativamente. Entre otras posibilidades, el programa SPSS (*Statistical Package for Social Sciences*), versión 22, se erige como una buena elección, pues se trata de un programa estadístico muy utilizado en investigaciones relacionadas con las ciencias sociales, como la psicología o la medicina, en las que se analizan las estadísticas en relación a datos sociológicos como el sexo, la edad, etc. Este tipo de datos también aparecen en este estudio, por ello la conveniencia de su utilización.

Las variables que se manejan en la investigación corresponden a *variables independientes* y *variables dependientes*. Las variables independientes son aquellas que vienen dadas en los sujetos y, a su vez, pueden condicionar la variable dependiente.

Las variables independientes utilizadas en la investigación son idioma (español, alemán), edad (14, 15, 16, 17) y sexo (mujer, hombre). La variable dependiente, aquella sobre la cual queremos ver la incidencia de las otras, es la competencia fraseológica.

Las operaciones principales aplicadas para el análisis de los datos fueron de varios tipos. Por un lado, se utilizaron *medidas de tendencia central*, es decir, medidas que agrupan los datos obtenidos, y por lo tanto son valores que resultan representativos de todos los valores que toma la variable. Entre estas podemos destacar la *media*, que supone la media aritmética de todos los valores.

Para poder observar además si los valores que ofrecen los análisis son significativos se utilizan también diferentes métodos, de los cuales hemos empleado, entre otros, los estadísticos conocidos como T-Student. Puesto que en el entorno lingüístico lo primordial es facilitar una interpretación visual, se ofrecen a continuación los datos.

2 Datos Relevantes

En la investigación se plantea una hipótesis previa: hay diferencias en la competencia fraseológica global de los jóvenes. A partir del análisis de los datos se comprobará si se valida o refuta. Pasamos a exponerlos a continuación.

2.1 La Competencia Fraseológica Global de los Jóvenes

Las medias de competencia obtenidas en ambas muestras se observan a continuación (Tabla 3):

Tabla 3. Nivel de competencia global de las muestras española y alemana

	Estadísticos de grupo				
	IDIOMA	N	Media	Desviación típ.	Error típ. de la media
COMPETENCIA	ESPAÑOL	45	30,11	7,493	1,117
FRASEOLÓGICA	ALEMÁN	45	28,02	6,538	,975
GLOBAL					

De las cincuenta UF analizadas en el cuestionario, los estudiantes españoles muestran una media de competencia de 30,11 UF, mientras que los alemanes tienen una competencia del 28,02 UF. Aunque se observa diferencia de competencia, a priori se desconoce si esta puede ser significativa. Para comprobarlo, se realiza una prueba T-de Student (Tabla 4). Se trata de una prueba de contraste estadístico que considera las diferencias de medias entre dos muestras que son independientes entre sí como en este caso.

Tabla 4. Nivel de competencia global: Prueba T (1)

	Prueba de muestras independientes				
		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias	
		F	Sig.	T	
TOTAL	Se han asumido varianzas iguales	,377	,541	-1,409	
COMPETENCIA	No se han asumido varianzas iguales			-1,409	

Y a continuación se observa si es significativa (Tabla 5):

Tabla 5. Nivel de competencia global: Prueba T (2)

Prueba de muestras independientes				
Prueba T para la igualdad de medias				
		Gl	Sig. (bilateral)	Diferencia de medias
TOTAL	Se han asumido varianzas iguales	88	,162	-2,089
COMPETENCIA FRASEOLÓGICA	No se han asumido varianzas iguales	86,414	,162	-2,089

Los análisis indican que, al menos con la muestra disponible en nuestro estudio, no hay diferencias entre los resultados de ambas poblaciones: $T=-1,40$ (Tabla 4), ya que ofrece un valor sig. = ,162 (Tabla 5), mayor de ,05. Esto significa que, aunque las medias son distintas, esa diferencia de medias no resulta significativa.

Puesto que ambas comunidades son distintas, tanto lingüística como culturalmente, cabía esperar diferencias significativas entre la competencia mostrada por los alumnos españoles y alemanes. Sin embargo, los jóvenes españoles y alemanes presentan una competencia muy similar en su lengua nativa.

3 Conclusiones

Del mismo modo que un estudio sobre disponibilidad léxica ofrece datos interesantes sobre el uso del idioma por parte de la población que lo sustenta, entendemos que este tipo de estudio en el campo fraseológico puede ser también de gran ayuda, pues no solo arrojará luz sobre el uso real de fraseología en las comunidades lingüísticas respectivas y se comprobará si las expresiones más usadas en una comunidad coinciden en la otra, sino que, además, verificará si esa competencia es socialmente homogénea o por el contrario depende de ciertos factores sociales.

El objetivo perseguido con este tipo de investigaciones no acaba con la descripción de la situación, sino que la información lingüística obtenida puede ser utilizada en posteriores estudios y permite a su vez establecer *mínimos fraseológicos*, tan útiles para trabajar en situaciones de interacción lingüística como son la docencia de lenguas extranjeras o la traducción, y también para elaborar las herramientas necesarias de estas dos disciplinas lingüísticas: diccionarios fraseológicos que atiendan a criterios de uso efectivo, indispensables cuando se trata de contrastar estas expresiones.

Desde esta pequeña contribución esperamos sentar las bases para realizar estudios más amplios a fin de disponer de datos más representativos sobre muestras de población mayores.

Referencias

1. Yaguello, M.: *Alicia en el país del lenguaje*, Mascarón, Madrid (1983).
2. Sevilla Muñoz, J.: «La competencia paremiológica en la generación española de más de 65 años», *Phraseologie global - areal – regional*, Jarmo Korhonen, Wolfgang Mieder, Elisabeth Piirainen, Rosa Piñel (eds.). Helsinki: Universitat Helsinki, pp. 151-158 (2010).
3. López Roig, C.: *Aspectos de Fraseología (alemán-español) en el sistema y en el texto*, Peter Lang, Frankfurt a. M. (2002).
4. Tarnovska, O.: *Refranero básico español con correspondencias en ruso y ucraniano*, Logos, Kiev (2004).
5. Conca, M. y J. Guia: *La fraseología. Principis, mètodes i aplicacions*, Bromera, Valencia (2014).

Variaciones Fraseológicas en la Terminología Médico-Farmacéutica y su Aplicación en las Traducciones EN>ES y DE>ES

Francisco Bautista Becerro

Universidad de Salamanca
Facultad de Traducción y Documentación
C/Francisco de Vitoria, 6, C.P.: 37008, Salamanca, España
fran_bautista@usal.es

Abstract. El lenguaje de tipo médico-farmacéutico se caracteriza, entre otras cosas, por su extensa terminología específica y especializada. En los textos de este tipo rigen la precisión y la exactitud y para ello se busca una univocidad terminológica que no siempre es factible. Por ello, es muy frecuente encontrar unidades fraseológicas con valor de término e incluso perífrasis compuestas por varias palabras. Si bien este fenómeno es habitual en todos los campos de especialidad y, por supuesto, también en el lenguaje coloquial, no deja de ser relevante y en numerosas ocasiones es algo que se debe tener en cuenta. El objetivo de esta publicación es analizar determinadas situaciones en las que términos del ámbito médico-farmacéutico tienen una estructura fraseológica diferente en inglés o alemán que en español y algunos tipos de casos que nos podemos encontrar. No solo se estudian estas variaciones y las diferencias entre idiomas, sino también las consecuencias de una mala traducción y las estrategias para evitarlo.

Keywords: Traducción, Lenguaje médico-farmacéutico, Traducción médico-farmacéutica

1 Lenguaje Médico-farmacéutico y su Traducción

Lo primero que queremos hacer en esta publicación es proporcionar una definición de lenguaje *médico-farmacéutico*. Son muchos los expertos que recuerdan lo difícil (por no decir imposible) que resulta delimitar un lenguaje especializado, en este caso el científico (y, por ende, el farmacéutico), y diferenciarlo de otros lenguajes especializados y del común (Gutiérrez Rodilla, 2005: 19; Cabré, 1999: 189 en Cabré *et al.*, 2001: 173). Si bien es cierto que *a priori* no parece factible establecer unos límites definidos para el «lenguaje científico», ni para ninguno de los subtipos de lenguaje que abarca (aquí incluiremos el «lenguaje médico-farmacéutico» o simplemente «lenguaje farmacéutico», y aprovechamos para remarcar que utilizaremos ambas denominaciones, en ocasiones indistintamente), a efectos de esta publicación intentaremos establecer unos límites, por muy difusos que puedan resultar en ocasiones, y definiremos una serie de características globales. En este caso, cuando hablemos de *lenguaje médico-farmacéutico*, nos referiremos a todo lenguaje presente en textos de

ámbitos médicos y farmacéuticos: es decir, textos que traten sobre medicamentos (así como su uso, mecanismo de acción, investigación, interacciones...), enfermedades que afecten al ser humano o diferentes condiciones fisiopatológicas del ser humano y que puedan estar presentes en instituciones sanitarias, como hospitales, clínicas, centros de salud, oficinas de farmacia, laboratorios farmacéuticos, etc. Asimismo, y también en la línea defendida por Gutiérrez Rodilla (2005:22), incluiremos todos los tipos de comunicación, ya se dé esta entre especialistas o entre especialistas y un público general, y por todos los canales (tanto oral como escrito). Concluimos que se trata de un tipo de lenguaje con un elevado grado de especialización, con todo lo que ello conlleva y para su traducción serán necesarias unas «habilidades y destrezas particulares» (Corpas, 2004: 138), así como la formación necesaria (Calonge, 2001:104) y «conocer los elementos metodológicos y recursos para resolver los problemas de terminología» (Cabré, 2004 en Corpas, 2004: 138).

1.1 Características del Lenguaje Médico-farmacéutico

Una vez marcados los límites de lo que entendemos por lenguaje médico-farmacéutico, expondremos las características principales de estos tipos de texto. Para ello, de nuevo nos remitiremos a Gutiérrez Rodilla (2005:9), que recuerda que «una de las características máspreciadas de la ciencia es el rigor» y que defiende las siguientes características (*ibidem*: 22-30):

- **Precisión:** Gutiérrez Rodilla (*ibidem*: 22) la califica como «la cualidad máspreciada del discurso científico» y, aunque esta puede referirse al uso de aclaraciones o incisos explicativos que deshagan las ambigüedades, la relaciona, sobre todo, con «la precisión de los términos» (*ibidem*: 23). Otro requisito planteado por Gutiérrez Rodilla (*ibidem*) es que los términos sean monosémicos y no cuenten con sinónimos, por muy quimérico que esto pueda parecer. Esto resulta especialmente relevante a efectos de nuestra publicación, que estudia los términos científicos como unidades fraseológicas con las que conseguir la mayor precisión y, con ella, el mayor rigor posible en los textos médico-farmacéuticos.
- **Neutralidad:** definida por Gutiérrez Rodilla (Gutiérrez Rodilla, 2005:23) como «la carencia de valores y connotaciones afectivas [y] subjetivas». Aunque se trate de una característica deseable y destacable en el lenguaje científico, no abundaremos mucho en ella, ya que no tiene las repercusiones fraseológicas que sí encontrábamos en la precisión como característica del lenguaje farmacéutico.
- **Economía:** el mensaje científico «debe expresarse con el menor número posible de unidades» (*ibidem*: 25) y busca la concisión ligada al menor número de palabras posible, cuyo máximo exponente es «la sustitución de frases enteras por un solo término». Esta característica sí es relevante para nuestro estudio, en el que analizamos los casos en que una unidad fraseológica formada por un solo término en inglés o alemán debe traducirse por una mucho mayor en español y viceversa. De hecho, aunque lo ideal pudiera ser tender a

la simplificación, es inevitable (y muy frecuente) recurrir a un número elevado de palabras (*ibidem*).

- Los recursos empleados: en el mensaje científico es habitual encontrar «dibujos, esquemas, planos, fórmulas, diagramas, cuadros, modelos, etc» (*ibidem*: 25), mecanismos que no pertenecen a «ningún tipo de mensaje» y en los que tampoco profundizamos en esta publicación por su falta de relación con la fraseología.
- El vocabulario científico: Gutiérrez Rodilla (*ibidem*: 28) defiende que existe un cierto acuerdo entre los distintos autores en que el vocabulario del lenguaje científico se sirve de su «aspecto más distintivo, menos diferenciador». Además, añade algunas de las características del vocabulario científico: está compuesto por «adjetivos, verbos y, fundamentalmente, sustantivos» (*ibidem*), que en una gran proporción «se construyen mediante la combinación de formantes griegos y latinos» (como también enfatiza Calonge [2001:104]) y que «muestra una velocidad de crecimiento muy grande» (Gutiérrez Rodilla, 2005: 28). Esto hace que sea prácticamente imposible calcular el número de términos del lenguaje científico en general y del farmacéutico en particular, y aquí entra en juego una vez más la importancia de unidades fraseológicas de distinto tipo que nos aporten esa precisión y exactitud que se busca y se necesita en todo texto del ámbito médico-farmacéutico y que sería imposible conseguir de otra manera. Esto es algo que todo traductor especializado debe tener en cuenta: como defiende Corpas (2004: 140), la terminología es uno de los pilares de la traducción del ámbito biosanitario (junto con la documentación), en línea con Cabré (2004: 2): «la terminología es absolutamente imprescindible para el ejercicio de la traducción especializada»

Después de haber explicado en qué consisten estas principales características, podemos aventurar si es posible alcanzar una terminología «perfecta» en el lenguaje médico-farmacéutico e incluso una respuesta: no. Es cierto que la terminología del lenguaje médico es muy extensa (la propia Gutiérrez Rodilla [2005:10; 2014:54] nos da un dato revelador: en los tres primeros años de carrera, los estudiantes de Medicina aprenden alrededor de 15 000 palabras nuevas, «cifra muy superior a la del vocabulario de un curso básico de cualquier idioma extranjero»), y recuerda que los estudiantes que acceden a cualquiera de las titulaciones del ámbito científico y técnico deben enfrentarse no solo al aprendizaje de los conocimientos específicos propios, sino también al lenguaje en que esos conocimientos se expresan (Gutiérrez y Quijada, 2014:53). Dicho lo cual, no podemos olvidar que la terminología seguirá aumentando y se seguirán formando nuevos términos y unidades fraseológicas con los que hacer referencia a todos los fenómenos químicos, biológicos y fisiológicos, así como para todas las enfermedades, compuestos químicos, etc.

1.2 Características Fraseológicas del Lenguaje Médico-farmacéutico

Como hemos comentado en el apartado anterior, las unidades fraseológicas son necesarias en el lenguaje científico en general —y en el médico-farmacéutico en par-

ticular— para conseguir la precisión, exactitud y univocidad necesarias en estos textos. Para conseguir las, como ya hemos visto, es muy habitual recurrir a unidades fraseológicas y perífrasis, que, si bien son unidades compuestas por dos palabras o más, consiguen cumplir dicho propósito. En esta publicación seguimos la definición de unidades fraseológicas propuesta, por ejemplo, por Penadés (2000:11): «una combinación fija de palabras que, en numerosas ocasiones, tiene un significado que no se desprende del significado de sus elementos constituyentes» y por Martínez Marín (2000) de que poseen «contenido semántico como otros elementos lingüísticos». Además, al estar hablando de un campo semántico especializado, también nos referiremos a ellas como unidades fraseológicas especializadas (UFE), unidades de significación especializada (USE) con las siguientes propiedades (Bevilacqua, 2001:114): incluyen una unidad terminológica (UT) simple o sintagmática, incluyen un elemento con valor eventivo a partir del cual se organiza semánticamente el discurso, tienen un cierto grado de fijación determinado por la relación semántica establecida entre sus elementos, más que por las relaciones sintácticas y tienen una frecuencia relevante en el conjunto de texto en que aparecen (en este caso, los textos médico-farmacéuticos). Dicho de otra forma, (Bevilacqua, 2004: 29-30) algunos de sus requisitos son su carácter sintagmático, su estabilidad sintáctica y semántica o el uso en un ámbito específico.

Las UFE son necesarias en ámbitos especializados (Sevilla, 2015:237) y a menudo se estudian «de la misma manera que la terminología» (*ibidem*). En este trabajo hemos adoptado un enfoque similar y, a continuación, exponemos una serie de ejemplos de unidades fraseológicas de distinto tipo en inglés y alemán respecto del español, las diferencias entre ellos y su traducción. Siguiendo las definiciones expuestas en el párrafo anterior, aportaremos una serie de UFE con distinta estructura fraseológica en los idiomas en cuestión.

2 Variaciones Fraseológicas en la Traducción EN>ES de Textos de Carácter Médico-farmacéutico

La traducción médico-farmacéutica EN>ES es un mercado que mueve millones de euros en el mundo¹ y en la que se busca y se precisa un rigor absoluto. Para ello, es necesario adaptar el texto original al español, lo que en numerosas ocasiones exige recurrir a unidades fraseológicas diferentes en uno y otro idioma. Los distintos casos que pretendemos analizar en esta publicación son los siguientes (por motivos de espacio, solo podremos exponer un número limitado de ejemplos en cada caso).

2.1 Verbos

El primer apartado se lo dedicamos a los verbos y, en concreto, a los celeberrimos *phrasal verbs*, un auténtico quebradero de cabeza para los estudiantes de inglés, pero también una estructura muy interesante desde el punto de vista fraseológico. Estos

¹ <https://www.merca2.es/traducion-e-interpretacion-una-industria-en-crecimiento/>

verbos, compuestos de dos palabras (verbo + preposición) y con un significado muy concreto, no existen como tal en español y su traducción será, en la práctica totalidad de los casos, un verbo monoléxico o una perífrasis verbal. Si bien esta condición no es específica ni exclusiva de los textos farmacéuticos, y aunque ya hemos visto que la categoría gramatical predominante en el lenguaje científico es el sustantivo, y los *phrasal verbs* suelen tener equivalentes no preposicionales de registro más alto, sí que queremos aportar una lista con algunos de los más frecuentes en este campo:

- *-get over*: recuperarse (de una enfermedad);
- *-pass away*: fallecer;
- *-pass out*: desmayarse, perder el sentido;
- *-pick up*: contraer (una enfermedad);
- *-rule out*: en su *Libro rojo*, A. Navarro la define como «descartar en el sentido de excluir o rechazar una posibilidad diagnóstica o de otro tipo». Además de ser un verbo muy utilizado, por ejemplo, en artículos científicos, la palabra *rule* nos ofrece la base para muchas otras unidades fraseológicas perífrásticas: *palm of the hand rule*, *rule of nines* (regla de los nueve), *work-to-rule* (huelga de celo), *Markownikoff rule* (regla de Markóvnikov), etc.
- *-throw up*: vomitar.

2.2 Sustantivos

Este apartado lo dedicamos a estructuras fraseológicas con categoría gramatical sustantiva en inglés y en español. Como hemos comentado previamente, los sustantivos son especialmente relevantes en textos médico-farmacéuticos y son el tipo de palabra más habitual. A continuación, enumeramos una serie de ejemplos:

- *drug-drug interactions*: esta estructura formada por tres sustantivos, los dos primeros de los cuales tienen función adjetiva, se traduce preferentemente como *interacciones farmacológicas* por encima de los calcos *interacciones fármaco-fármaco*, *interacciones droga-droga* o *interacciones entre fármacos*, que serían los equivalentes en español con la misma estructura fraseológica. En contraposición, sí hay otro tipo de interacción (*drug-body interaction*) cuya traducción sigue esta última estructura: *interacción entre fármaco y organismo*, que en cualquier caso nos interesa desde el punto de vista fraseológico por el cambio que representa respecto del original, pues es necesaria la adición de una construcción preposicional con función adjetiva.
- *cost-effectiveness*: en muchos casos, el inglés permite una mayor flexibilidad gramatical en cuanto al orden de las palabras y su función gramatical. En español, por suerte, tenemos un término que recoge el mismo significado con una sola palabra: *rentabilidad*, que, como declara Fernando A. Navarro², «es mucho más concisa que el calco habitual *relación coste-efectividad*». Es un

² En la entrada del *Libro rojo* relativa a *cost-effectiveness* (*Diccionario de dudas y dificultades de traducción del inglés médico* (3.ª edición) Versión 3.13; marzo de 2019)

término muy extendido en farmacoeconomía, donde también es habitual verlo traducido como *eficiencia*.

- *low back pain* (también *lower back pain*, *lower back ache*, *low back ache* o incluso *lumbar pain*): si lo tradujéramos textualmente, tendríamos una UFE de ocho palabras (dolor en la parte baja de la espalda), mientras que también contamos en español con una palabra para designarlo: lumbalgia (o lumbago), con lo que conseguimos la economía que Gutiérrez Rodilla defendía en los textos médicos sin perder la precisión.
- *stroke*: al contrario que en el ejemplo anterior, esta palabra tan sencilla en inglés tiene una traducción mucho más larga en español: *accidente cerebrovascular*, que afortunadamente se puede abreviar con el acrónimo ACV.

2.3 Adjetivos

Una vez hemos visto una serie de ejemplos de verbos y sustantivos con relevancia fraseológica en la redacción y traducción de textos farmacéuticos, también querríamos nombrar una serie de adjetivos:

- *compassionate*: término habitual en ciertos contextos médicos, Fernando A. Navarro propone como traducción *compasivo*, pero también el uso de la unidad fraseológica «por motivos humanitarios». Además, cita el ejemplo de *compassionate leave*, cuya traducción sería *permiso por motivos familiares* (UFE formada por un sustantivo y una construcción preposicional).
- *impaired*: el adjetivo *impaired* se usa con mucha frecuencia en textos médico-farmacéuticos en inglés. Tanto es así que en español se puede traducir de numerosas maneras, incluyendo otros adjetivos (alterado, inválido...) y, lo que nos interesa en nuestro estudio, estructuras fraseológicas diferentes. Es el caso de *renally impaired patients* (pacientes con insuficiencia renal), *impaired growth* (retraso del crecimiento), *impaired hearing* (hipoacusia, deficiencia auditiva), *impaired immune system* (inmunodeficiencia), *impaired metabolism* (trastorno metabólico), etc.

Unidades fraseológicas formadas por adjetivo (o función adjetival) + sustantivo.

Aunque en algunos de los ejemplos anteriores hemos vistos cómo las unidades fraseológicas en español cambian al añadir un adjetivo ante un sustantivo (y viceversa), hay otros términos formados por más de una palabra y con la estructura adjetivo + sustantivo cuya traducción al español resulta interesante desde el punto de vista fraseológico por las diferencias estructurales que encontramos. A continuación planteamos algunos ejemplos de especial interés o relevancia.

- *Cardiac death*: la traducción literal *muerte cardiaca* no funciona en español porque, literalmente significaría «muerte del corazón». La traducción correcta, por tanto, sería «muerte de causa cardiaca» o «muerte de origen cardiaco».

Fernando A. Navarro explica esto en su *Libro rojo*³ y cita otros dos ejemplos análogos pero con distintas estructuras fraseológicas de traducción: *cardiovascular death* (muerte por causas cardiovasculares) y *cell death* (muerte celular).

- *Clinical trial*: aunque la primera opción de traducción y la mundialmente aceptada es «estudio clínico» o «ensayo clínico», hay muchos tipos de ensayos clínicos distintos, que nos ofrecen alternativas de traducción tan variopintas como interesantes
 - *crossover clinical trial* (estudio [o ensayo] clínico con grupos cruzados): *crossover* en un sustantivo con función adjetival que, al no tener equivalente en español, se traduce con una construcción preposicional con valor adjetivo.
 - *phase I clinical trial* (estudio [o ensayo] clínico de fase I): una vez más, el sustantivo *phase I* adquiere función adjetival al anteponerse a *clinical trial*, y una vez más es necesaria una construcción preposicional en español.
- *graft-versus-host disease*: esta unidad fraseológica tiene una estructura muy interesante debido a que la función adjetiva la cumple un sintagma nominal. Su traducción resulta igualmente interesante, pues aunque el término más asentado es «enfermedad del injerto contra el anfitrión» (EICA) y, principalmente, «enfermedad injerto contra huésped», su mecanismo y etiología llevan a Fernando A. Navarro a proponer una alternativa más sencilla y con una estructura fraseológica diferente: *rechazo inverso*⁴.
- *Investigational new drug (IND)*: este término es muy utilizado en artículos científicos en los que hablan de nuevos *productos en fase de investigación clínica*, que sería su traducción exacta y que, como vemos, constituye una estructura fraseológica totalmente distinta al no ser válido en español un adjetivo parecido a «investigacional».
- *Rate*: aunque en inglés puede funcionar también como verbo, nos centraremos en algunas construcciones en las que tiene función sustantiva y va acompañada de otras palabras con función adjetiva:
 - *new case rate*: al contrario de lo que suele ser habitual, en español contamos con una construcción más sencilla (dos sustantivos en vez de tres): *tasa de incidencia*.
 - *birth death rate*: esta construcción, formada por tres sustantivos (los dos primeros con función adjetival) se traduce en español con una construcción preposicional formada por un sustantivo antecedido por la preposición *de*: *tasa [o índice] de mortalidad*.

³ En la entrada del *Libro rojo* relativa a *cardiac death* (*Diccionario de dudas y dificultades de traducción del inglés médico* (3.^a edición) Versión 3.13; marzo de 2019)

⁴ En la entrada del *Libro rojo* relativa a *graft-versus-host disease* (*Diccionario de dudas y dificultades de traducción del inglés médico* (3.^a edición) Versión 3.13; marzo de 2019)

- *erythrocyte sedimentation rate*: una vez más, los dos sustantivos con función adjetival se pueden traducir en español con una construcción preposicional. Del mismo modo, cabe destacar que una de las alternativas implica el uso de un sustantivo prefijado: *velocidad de sedimentación globular*, *velocidad de eritrosedimentación*.
- *pulse rate*: la amplia polisemia de *rate* en inglés hace que su uso en español a veces resulte superfluo. Es el caso de *pulse rate*, cuya traducción en español puede ser *número de pulsaciones*, pero también simplemente *pulso*. Algo similar ocurre con *absorption rate constant* (constante de absorción) y *basal metabolic rate* (metabolismo basal). Mantener el *rate* en español, por tanto, daría lugar a textos más largos y podría llegar a dificultar la comprensión, por lo que se podría considerar un error de traducción.
- *immune*: se puede traducir como *inmunitario*, *inmunológico* o *inmunizante* (en función del contexto, y evitando en general el calco «inmune»), pero en esta publicación nos interesan los casos en que en español se usa el prefijo ‘inmuno-’ como traducción del adjetivo inglés *immune*, p. ej., con *immune body* (anticuerpo) o *immune cell* (inmunocito). Una vez más, una traducción literal (¿cuerpo inmune?) constituiría un fallo grave de traducción y denotaría falta de conocimientos o documentación por parte del traductor.

3 Variaciones Fraseológicas en la Traducción DE>ES de Textos de Carácter Médico-farmacéutico

Existen muchas diferencias entre la estructura gramatical del alemán y la del español que hacen que a la hora de traducir un texto (de cualquier campo) del alemán al español sea necesario recurrir a unidades fraseológicas de diferente tipo. A continuación, veremos algunas de las más frecuentes y expondremos ejemplos propios del ámbito médico-farmacéutico.

3.1 Verbos Separables

En línea con los *phrasal verbs* ingleses, el alemán tiene sus verbos separables, caracterizados por incluir una partícula-prefijo que en las frases enunciativas e interrogativas de indicativo se coloca en última posición, algo que evidentemente no ocurre en castellano. Algunos de los verbos de este tipo frecuentes en textos médico-farmacéuticos son los siguientes:

- anpassen: adaptar
- ansprechen: corresponder
- beitragen: contribuir
- darstellen: representar
- feststellen: asegurar

- fortfahren: continuar
- nachgehen: seguir, perseguir
- vorsehen: prever

3.2 Sustantivos

Este apartado se lo dedicaremos a un tipo específico de sustantivos: los famosos *Komposita* alemanes. La composición es el procedimiento morfológico por el que las lenguas pueden unir lexemas para crear nuevas palabras. Es un mecanismo altamente productivo tanto en español como en alemán (Recio y Torijano, 2018:387) y, lo que nos interesa, muy presente en el lenguaje médico-farmacéutico, en el que la precisión a menudo obliga a ello. Algunos ejemplos son los que mencionamos a continuación:

- *Abdominalsyndrom*: buen ejemplo de palabra formada por dos raíces (*Abdominal-* y *-syndrom*) y, además, de palabra que puede dar lugar a confusión, pues su traducción correcta sería *peritonismo* o *seudoperitonitis* (y no «síndrome abdominal»).
- *kastrationsresistente Prostatakarzinom*: una enfermedad muy concreta, que en español sería *cáncer de próstata resistente a la castración*. Mientras que en alemán tenemos una UFE formada por dos palabras, la primera de las cuales está compuesta de un sustantivo y un adjetivo y la segunda por dos sustantivos unidos, en castellano el resultado es nada más y nada menos que una UFE de siete palabras.
- *Transkriptionsfaktoren*: un ejemplo sencillo de cómo un sintagma nominal en español (factores de transcripción) en alemán se expresa con una sola palabra.
- *[Polymerase]-Kettenreaktion*: reacción en cadena [de la polimerasa]. Una vez más vemos cómo las unidades fraseológicas son totalmente distintas en español, al no contar con el recurso de la composición.
- *Röntgenstrahlen*: literalmente, «rayos de Röntgen», en español se denominan habitualmente «rayos X». Es un término interesante por todas las palabras derivadas de él: *Röntgentherapie* (radioterapia), *Röntgenfilm* (película de rayos X), *Röntgenbild* (radiografía), *Röntgenologie* (radiología), *Röntgenarzt* (radiólogo), *Röntgendermatitis* (radiodermatitis), *röntgen* (radiografiar), *Röntgendiagnostik* (radiodiagnóstico), etc. (A. Navarro, 1997: 79).

Curiosidades traductológicas y fraseológicas DE>ES. La facilidad que tiene el alemán para el proceso de la composición da lugar a palabras en las que se consigue una gran precisión, pero también a otras que podemos considerar curiosas por su estructura, origen o significado. Además de las enumeradas anteriormente, queremos aportar los siguientes ejemplos:

- *Hausapotheke*: literalmente significa «farmacia casera», pero en realidad hace referencia al *botiquín*.
- *Höhensonne*: aunque literalmente significaría «sol de altitud», en realidad es una lámpara de rayos ultravioleta.

- *Krankengymnastik*: literalmente significa «gimnasia para enfermos», pero su traducción correcta sería *fisioterapia*. Hay que tener especial cuidado, pues también existe en alemán la palabra *Physiotherapie*, con un significado más general y que designa todos los tratamientos por agentes físicos (A. Navarro, 1997:78).
- *Meerschweinchen*: literalmente significa «cerdito de mar», pero hace referencia al que tal vez sea el animal más emblemático en investigación: la cobaya.

3.3 Adjetivos

- *Alzheimersche Krankheit*: aunque no es difícil adivinar que se trata de la enfermedad de Alzheimer (o «alzhéimer», como es utilizada habitualmente), es destacable ver cómo se ha adjetivado el epónimo (la enfermedad se llama así en honor al neurólogo Alois Alzheimer)⁵. En la misma línea, tenemos la *Addisonische Krankheit* (enfermedad de Addison), la *Bornholmer Krankheit* (pleurodinia epidémica [de Bornhorn]) o la *Basedowsche Krankheit* (enfermedad de Graves-Basedow, también conocida como hipertiroidismo, bocio exoftálmico hipertiroides).
- *englische Krankheit*: aunque literalmente significa «enfermedad inglesa», su traducción correcta es una unidad fraseológica mucho más sencilla: raquitismo.

4 Conclusiones

Una vez hemos aportado una posible definición para «lenguaje farmacéutico» y hemos estudiado algunas de las estrategias de traducción de términos concretos en los que el resultado es una unidad fraseológica distinta, podemos extraer una serie de conclusiones. Por ejemplo, es importante tener en cuenta las diferentes estructuras gramaticales del inglés y el alemán y no ceñirse a las originales a la hora de traducir al español. Si bien esto es fundamental en todas las traducciones, y no es en ningún caso un aspecto que sea exclusivo del lenguaje farmacéutico, en este tipo de textos puede ser crucial debido a los contextos en que los encontraremos (hospitales, centros de investigación...) y la importancia de su contenido. Los ejemplos enumerados son habituales en este tipo de textos y hay que saber cómo afrontar su traducción o, puesto que hablamos de estructuras sintácticas diferentes, *transposición*. Por ejemplo, en inglés la estructura adjetivo + sustantivo a menudo obliga a recurrir en español a construcciones preposicionales, y ocurre lo propio cuando en alemán tenemos *Komposita* o palabras compuestas, aunque tampoco podemos olvidar los casos en que sucede lo contrario. Por todo ello, y recalando una vez más la importancia de la precisión en los textos de carácter médico-farmacéutico, consideramos fundamental seguir investigando las unidades fraseológicas para ayudar a aportar dicha precisión, deseada en

⁵ Más información en la entrada de «Alzheimer» del *Libro Rojo*.

todo texto de este tipo. Del mismo modo, destacamos una vez más la importancia de un uso correcto de la terminología, a menudo constituida por unidades fraseológicas de varias palabras o de un número de palabras distinto entre la lengua origen y la lengua meta, como hemos visto repetidamente en este trabajo.

Bibliografía

1. Bevilacqua, C.: «Unidades fraseológicas especializadas (UFE): elementos para su identificación y descripción». En: *La terminología científico-técnica*, pp. 113-141. Ed. por María Teresa Cabré y Judit Feliu. Institut Universitari de Lingüística Aplicada, Barcelona (2001).
2. Bevilacqua, C.: 2004. *Unidades fraseológicas especializadas eventivas: descripción y reglas de formación en el ámbito de la energía solar*. Universidad Pompeu Fabra, Barcelona (2004).
3. Cabré, M. T., Domènech, M., Morel, J. y Rodríguez, C.: *La terminología científico-técnica*. Institut Universitari de Lingüística Aplicada, Barcelona (2001).
4. Cabré Castellví, M. T.: «La terminología en la traducción especializada». En: *Manual de documentación y terminología para la traducción especializada*, pp. 89-122. Ed. por Gonzalo García, Consuelo; García Yebra, Valentín. Madrid: Arco/Libros. (2004).
5. Calonge Prieto, M.: «La complejidad del lenguaje de los textos médicos y la terminología especializada. Nociones para el estudiante de traducción médica (inglés-español)». En *Panorama actual del estudio y la enseñanza de discursos especializados*, pp. Ed. por María-José Varela Salinas. Peter Lang, Berna (2009).
6. Corpas Pastor, G.: «La traducción de textos médicos especializados a través de recursos electrónicos y corpus virtuales». En: *Las palabras del traductor. Actas del II Congreso Internacional «El español, lengua de traducción»*, vol. 20, págs. 137-164. Ed. por Luis González/Pollux Hernández, Bruselas (2004).
7. Gutiérrez Rodilla, B. M.: *El lenguaje de las ciencias*. Gredos, Madrid (2005).
8. Gutiérrez Rodilla, B. M. y Quijada Diez, C.: «El lenguaje médico en los planes de estudios de las titulaciones biosanitarias». En: *La importancia del lenguaje en el entorno biosanitario*, pp. 53-62. Ed. por Bertha M. Gutiérrez Rodilla y Fernando A. Navarro. Barcelona. Fundación Dr. Antonio Esteve (2014).
9. Martínez Marín, J.: «El significado de las unidades fraseológicas en los diccionarios monolingües del español: el caso de las locuciones». En: *Las lenguas de Europa: estudios de fraseología, fraseografía y traducción*, ed. por Gloria Corpas Pastor. Editorial Comares, Granada (2000).
10. Navarro, F. A.: «Palabras alemanas de traducción engañosa en medicina». En *Monografías Dr. Antonio Esteve. Traducción y lenguaje en medicina*, pp. 69-82. Ed. por Fernando A. Navarro. Fundación Dr. Antonio Esteve, Barcelona (1997).
11. Penadés Martínez, I.: *La hiponimia en las unidades fraseológicas*. Servicio de publicaciones Universidad de Cádiz, Cádiz (2000).
12. Sevilla Muñoz, M.: «Las unidades fraseológicas del discurso científico-técnico y su traducción (inglés-español)». En *Enfoques actuales para la traducción fraseológica y paremiológica: ámbitos, recursos y modalidades*, pp. 239-256. Ed. por Germán Conde Tarrío, Pedro Mogorrón Huerta, Manuel Martí Sánchez y David Prieto García-Seco. Centro Virtual Cervantes (Instituto Cervantes), Biblioteca fraseológica y paremiológica (2015).
13. Torijano, J. A. y Recio Ariza, M. A.: «La problemática de los *Komposita* en la fraseología». En *Lenguas en contacto, ayer y hoy. Traducción y variación desde una perspectiva*

filológica, pp. 383-404. Ed. por Santiago del Rey Quesada, Florencio del Barrio de la Rosa y Jaime González Gómez. Peter Lang, Berlín (2018).

Recursos online

14. Navarro, Fernando A. 2019. *Diccionario de dudas y dificultades de traducción del inglés médico* (3.ª edición) Versión 3.13; marzo de 2019
15. *Traducción e Interpretación: una industria en crecimiento* (merca2). Última fecha de acceso: 31 de agosto de 2019.
16. <https://www.merca2.es/traduccion-e-interpretacion-una-industria-en-crecimiento/>

***Phrase Frames* en un Corpus Oral de Alemán como Lengua Extranjera para el Turismo**

María Rosario Bautista Zambrana¹

¹ Universidad de Málaga
mrbautista@uma.es

Abstract. En este estudio partimos de una concepción amplia de la fraseología y nos proponemos analizar las estructuras fijas o *phrase frames* presentes en un corpus de textos orales de dos libros de texto de alemán turístico de nivel A2, con el fin, en primer lugar, de cuantificar su uso y contrastarlo con los datos de un corpus formado por los textos orales de dos libros de texto de alemán general como lengua extranjera, también de nivel A2. En segundo lugar, nos proponemos analizar 20 estructuras fijas, tanto desde un punto de vista cuantitativo como cualitativo, discerniendo, entre otros aspectos, las funciones comunicativas predominantes. Nuestros resultados apuntan a que el discurso oral del ámbito turístico, al menos tal como figura en los libros de texto, está más convencionalizado, y que podría tratarse de un discurso especializado, a pesar de su aparente similitud con el discurso oral general.

Keywords: *Phrase frames*, Corpus, Alemán como lengua extranjera, Turismo.

1 Introducción

Este artículo¹ pretende ahondar en la importancia del aprendizaje de unidades *prefabricadas* en el proceso de aprendizaje de una lengua extranjera, concretamente de la lengua alemana en el ámbito del turismo. Partimos de la idea de que gran parte de la lengua que usamos es de naturaleza fraseológica, tal como establece el *idiom principle* de Sinclair (1991: 110): “the principle of idiom is that a language user has available to him a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments.”

Numerosos investigadores, a partir de estudios basados en corpus, han apoyado esta tesis; es el caso, por ejemplo, de Bolinger (1976), Pawley y Syder (1983), Wray (2002), Granger y Meunier (2008) o Morley (2019), entre otros. Esta constatación tiene consecuencias muy importantes para la enseñanza y aprendizaje de lenguas extranjeras, incluso en los niveles iniciales. Según O’Keeffe et al. (2007: 46), la enseñanza de vocabulario en los niveles básicos resulta insuficiente si no se presta

¹ El presente trabajo se enmarca en el seno de los proyectos ‘VIP: sistema integrado Voz-texto para IntérPretes’ (ref. FFI2016-75831-P, MINECO) e INTERPRETA 2.0 (PIE 17-015, UMA).

debida atención a los *chunks* (combinaciones de palabras)² más usuales, ya que muchos de ellos se usan con igual o mayor frecuencia que ciertas palabras individuales comunes. Granger y Meunier (2008: 247) ahondan en esta idea: ya que, según han demostrado estudios psicolingüísticos, la lengua se adquiere, almacena y procesa por medio de *chunks*, es de recibo que la fraseología ocupe un lugar primordial en la enseñanza de segundas lenguas.

Sobre la base de estas premisas, el presente artículo parte de una concepción amplia de la fraseología (véase Granger y Paquot, 2008) y pretende analizar, desde un punto de vista cuantitativo y cualitativo, cuál es el tratamiento que reciben las unidades pluriverbales en los libros de texto de alemán turístico, con especial referencia al componente oral, tanto desde un punto de vista receptivo como productivo. De forma más concreta, nos centraremos en analizar la presencia de estructuras fijas en un corpus formado por los textos orales de dos libros de texto de alemán turístico de nivel A2. Este tipo de unidad pluriverbal viene definido en el texto del Marco Común Europeo de Referencia para las Lenguas (Consejo de Europa, 2002; en adelante MCER), documento que ha servido de base para la unificación de directrices sobre el aprendizaje y la enseñanza de lenguas dentro del contexto europeo.

Las *estructuras fijas* se abordan en el MCER dentro de la llamada competencia léxica, que es definida como el conocimiento del vocabulario de una lengua — formado por elementos léxicos y elementos gramaticales— y la capacidad para utilizarlo. Dentro de los elementos léxicos encontramos las *expresiones hechas*, que son aquellas compuestas de varias palabras que se utilizan y se aprenden como un todo. Entre ellas se encuentran las *fórmulas fijas* y las *estructuras fijas*. Estas últimas son expresiones “aprendidas y utilizadas como conjuntos no analizados, en los que se insertan palabras o frases para formar oraciones con sentido” (MCER, 2002: 108). Como ejemplos encontramos al. *Könnte ich bitte ... haben?*, o esp. *Por favor, ¿sería tan amable de + infinitivo?* Estas estructuras fijas son equivalentes a lo que Römer (2009: 150) denomina *phrase frame*: “sets of n-grams which are identical except for one word, e.g. *at the end of*, *at the beginning of*, and *at the turn of* would all be part of the p[hrase]-frame *at the * of*.” Consideramos que estas secuencias que presentan algún elemento variable son de extraordinaria relevancia en el aprendizaje de una lengua, no solo por su frecuencia de aparición (como veremos en este artículo), sino también por su productividad. Römer (2010) y Fuster (2014) han realizado estudios

² Hay muchos términos para denominar a las combinaciones frecuentes de palabras: *unidades fraseológicas* (término usado sobre todo por el enfoque más estricto de la fraseología), *secuencias formulaicas* (Wray, 2002), *chunks* (De Cock, 2000), *expresiones hechas* (MCER, 2002), *phrasal patterns* (como tipo de *construcción*; Goldberg, 2006), *expresiones fijas*, *unidades pluriverbales*, *combinaciones usuales de palabras*, etc. En este trabajo hablaremos de *expresiones hechas* o *unidades pluriverbales*, en el sentido que les asigna el MCER (2002): expresiones compuestas de varias palabras que se utilizan y se aprenden como un todo.

muy destacados sobre *phrase frames* que han servido de base para una parte de los análisis presentados en este trabajo.³

En relación con las estructuras fijas, serán de relevancia para nuestro trabajo las funciones comunicativas que el MCER establece dentro de la competencia funcional que debe adquirir el alumno (MCER, 2002: 122-123):⁴ ofrecer y buscar información factual (1.1), expresar y descubrir actitudes (1.2), persuasión (1.3), vida social (1.4), estructuración del discurso (1.5) y corrección de la comunicación (1.6).

En este estudio nos proponemos, pues, analizar las estructuras fijas presentes en dos libros de texto de alemán turístico, con el fin, en primer lugar, de cuantificar su uso y contrastarlo con los datos de un corpus formado por los textos orales de dos libros de texto generales de alemán como lengua extranjera; en segundo lugar, nos proponemos analizar 20 estructuras fijas, tanto desde un punto de vista cuantitativo como cualitativo, discerniendo, entre otros aspectos, las funciones comunicativas predominantes. Nuestros resultados serán de carácter preliminar, debido al tamaño del corpus y el número limitado de estructuras estudiadas. El artículo se encuentra organizado de la siguiente forma: en el Apartado 2 presentamos la metodología seguida para llevar a cabo el trabajo; en el Apartado 3 exponemos los resultados del análisis cuantitativo y cualitativo; finalmente, el Apartado 4 expone una breve discusión de los resultados, así como las conclusiones preliminares de este estudio.

2 Metodología

Hemos seguido una metodología cuantitativa y cualitativa. A fin de realizar los análisis lingüísticos que nos hemos propuesto, hemos compilado en primer lugar un corpus de textos orales a partir de dos libros de texto de enseñanza de alemán turístico de nivel A2: *Herzlich willkommen Neu* (Cohen y Lemcke, 2001) y *Menschen im Beruf Tourismus A2* (Schümann et al., 2015). Hemos seleccionado las transcripciones de los diálogos y monólogos relativos a situaciones profesionales, así como las listas de expresiones usuales (*Redemittel*), orientadas igualmente a la expresión oral de situaciones profesionales.⁵ Hemos descartado aquellos textos orales que no tuvieran relación con situaciones profesionales (diálogos entre turistas, por ejemplo) y los ejercicios de fonética; se han eliminado en los textos seleccionados todos los enunciados explicativos, títulos y referencias a los hablantes al principio de cada intervención. Este corpus, denominado *Corpus 1*, contiene 153 textos, con un total de 15.996 *tokens* y 1.895 *types* (*type-token ratio* 12%).

³ Utilizaremos en este trabajo las denominaciones *phrase frames* y *estructuras fijas* de forma indistinta.

⁴ Se trata, concretamente, de *microfunciones*, definidas como “categorías para el uso funcional de enunciados aislados (generalmente breves), normalmente como turnos de palabra de una interacción” (MCER, 2002: 122).

⁵ Las situaciones profesionales de estos libros de texto abarcan la hostelería, la atención al cliente en recepciones de hotel y la información turística, y tienen lugar generalmente entre un especialista y un consumidor.

A fin de contrastar los datos extraídos del Corpus 1, hemos compilado un corpus general de alemán, con textos orales de manuales de enseñanza de alemán como lengua extranjera de nivel A2; llamaremos a este corpus *Corpus 2*. De la misma forma que en el caso anterior, se han compilado las transcripciones de diálogos y monólogos, descartando aquellos textos orales relacionados con ejercicios de fonética. Los textos proceden de los libros *DaF kompakt A2* (Sander et al., 2011) y *Aussichten A2.1* (Ros-El Hosni et al., 2010). Este corpus contiene 150 textos, que cuentan con un total de 18.665 *tokens* y 3.011 *types* (*type-token ratio* 16%).

El estudio de ambos corpus ha sido en principio de tipo *corpus-driven*, en el sentido de que hemos extraído los *phrase frames* de forma automática sin tener en cuenta ninguna categoría lingüística previa ni ninguna lista predefinida de expresiones.

Hemos llevado a cabo el análisis cuantitativo por medio de *kfNgram* (Fletcher, 2007), con el que hemos extraído todas las estructuras fijas o *phrase frames* de 2 a 6 palabras del Corpus 1 y del Corpus 2, respectivamente. Después realizamos un estudio cualitativo de los *phrase frames* de 4 palabras del Corpus 1, para lo cual extrajimos todos aquellos que aparecían con una frecuencia normalizada superior o igual a 30 veces por 100.000 palabras; esto, en el Corpus 1, supuso que la frecuencia de corte para extraer *phrase frames* fuese de 5 *tokens*. De estos tomamos 20 ejemplos, tratando de seleccionar aquellas estructuras que constituyeran una unidad comunicativa mínima, es decir, que contuvieran un sintagma, aunque incompleto, o que incluyeran al menos sujeto y verbo (aun cuando uno de estos fuera el constituyente variable). Luego asignamos a cada *phrase frame* su función comunicativa y comparamos su frecuencia de aparición con aquella de la misma estructura fija en el corpus general. Para esto extrajimos igualmente todas las estructuras fijas del Corpus 2 con una frecuencia normalizada de 30 *tokens* por 100.000 palabras, lo que corresponde a una frecuencia de corte de 6 *tokens*.

3 Resultados

Ofrecemos en primer lugar los resultados de la extracción de *phrase frames* con *n* 3 a 6, tanto en el Corpus 1 como en el Corpus 2, realizada con *kfNgram*. Hemos reflejado en la siguiente tabla, como explicamos en el apartado anterior, aquellos *phrase frames* que cuentan con al menos una frecuencia de 5 *tokens* en el Corpus 1, y con al menos 6 *tokens* en el Corpus 2. Hemos indicado el número total de *phrase frames* encontrados, así como la cifra total de variantes detectadas. Se ha indicado igualmente cuál es el cociente entre el número de variantes y el número de *phrase frames*.

Tabla 1. Frecuencia absoluta de *phrase frames* en Corpus 1 y 2.

Corpus 1 (frec. absoluta)	Var./PF	Corpus 2 (frec. absoluta)	Var./PF
n=3		n=3	
1.155 PF / 7.380 var.	6,39	774 PF / 6.381 var.	8,2
n=4		n=4	

300 PF / 1.379 var. n=5	4,6	73 PF / 349 var. n=5	4,8
68 PF / 290 var. n=6	4,3	18 PF / 59 var. n=6	3,3
30 PF / 101 var.	3,4	9 PF / 31 var.	3,4

Como observamos, resulta evidente que en el corpus de textos orales turísticos hay un número mayor de estructuras fijas, frente al corpus general, tanto en lo que se refiere a los *phrase frames* de 3 palabras, como a los de 4, 5 y 6 palabras. Resulta llamativo que el número de estructuras fijas baje considerablemente al pasar de 3 a 4 palabras, tanto en el Corpus 1 como en el Corpus 2. Es preciso también resaltar que el número medio de variantes por *phrase frame* va descendiendo a medida que aumenta el número de *n*.

A continuación, hemos extraído 20 ejemplos de entre las estructuras fijas que aparecían 5 veces o más en el Corpus 1 (véase Tabla 2). Para cada uno de ellos, hemos indicado su rango (orden en el que aparece en la lista de *phrase frames* extraídos), frecuencia (número de *tokens*), número de variantes, *variant/p-frame ratio* (VPR) y función comunicativa (según el código asignado por el MCER, véase arriba). El *variant/p-frame ratio* (VPR) es un cociente propuesto por Römer (2010), que expresa la relación entre el número de variantes de cada *phrase frame* con respecto a los *tokens* que existen con ese patrón (cociente de variantes entre *tokens*, multiplicado por cien). Un resultado bajo indica que hay pocas variantes, mientras que un ratio alto significa que una estructura fija cuenta con numerosas variantes.

Tabla 2. Datos de *phrase frames* seleccionados.

Rango	Phrase frame	Frecuencia	Variantes	VPR	Función comunicativa
9	darf ich Ihnen *	18	7	38%	1.2
12	kann ich Ihnen *	16	6	37,5%	1.2
15	haben Sie schon *	14	3	21%	1.1
19	können Sie mir *	11	9	81%	1.2/1.3
21	ich hätte gern *	11	9	81%	1.2
23	dann nehme ich *	11	9	81%	1.2
30	ich bringe Ihnen *	10	5	50%	1.3
37	haben Sie noch *	10	6	60%	1.1
43	wir möchten gern *	9	6	66,66%	1.2
49	wie viel kostet *	9	3	33,33%	1.1
50	das Essen war *	8	6	75%	1.2
57	hier ist * Rechnung	8	2	25%	1.1
64	das hier ist *	8	7	87,5%	1.1
80	ich bringe * gleich	7	2	28,57%	1.3

86	das macht * Euro	7	2	28,57%	1.1
104	hier ist die *	7	6	85,71%	1.1
128	mit dem * fahren	6	2	33,33%	1.1
131	ich hoffe, dass *	6	3	50%	1.2
144	wie kommt man *	6	4	66,66%	1.1
154	ich wünsche Ihnen *	6	3	50%	1.2

Como se puede observar, la mayoría de estructuras fijas siguen el patrón A B C *, mientras que solo cuatro de ellas presentan el patrón A B * D. Según Römer (2010) y Fuster (2014), en estos casos no se trataría estrictamente de *phrase frames*; sin embargo, otros autores como Biber (2009) sí los tienen en cuenta, al tiempo que el propio MCER (2002) considera estructuras fijas todas aquellas que se aprenden y utilizan como conjuntos no analizados, sin tener en cuenta la posición del elemento variable. En el ejemplo del MCER (2002) *Por favor, ¿sería tan amable de + infinitivo?* podemos observar como el último constituyente es el variable, que además puede completarse por medio de más de una palabra.

Otro resultado destacable es que aproximadamente el 50% de los *phrase frames* seleccionados cumplen la función 1.1 (ofrecer y buscar información factual), mientras que el otro 50% presenta sobre todo la función 1.2 (expresar y descubrir actitudes).

En cuanto al cociente VPR, observamos una variación considerable entre los distintos *phrase frames*. La media de todos los resultados es de 53,99%, con una mediana de 50 y una desviación estándar de 22,67.

Si buscamos en el Corpus 2 los *phrase frames* seleccionados, encontramos que solo cinco de ellos se encuentran en las listas de estructuras fijas extraídas por *kfNgram*, teniendo en cuenta incluso aquellos *phrase frames* que solo cuentan con 2 *tokens*. En la Tabla 3 podemos comparar la frecuencia normalizada de aparición (por 100.000 palabras) de estas cinco estructuras:

Tabla 3. Comparación de frecuencias y rango de PF.

Phrase frame	Rango Corpus 1	Frec. norm. Corpus 1	Rango Corpus 2	Frec. norm. Corpus 2
haben Sie schon *	15	87,5	7.107	10,7
können Sie mir *	19	68,8	527	26,8
ich hätte gern *	21	68,8	3.171	10,7
hier ist die *	104	43,8	2.107	16,1
mit dem * fahren	128	37,5	2.839	10,7

4 Discusión y conclusiones

En este trabajo nos hemos propuesto compilar un corpus de textos orales de alemán como lengua extranjera en el ámbito específico del turismo —concretamente de textos

orales de comunicación profesional especialista-consumidor—, a fin de estudiar las estructuras fijas presentes en el mismo. Igualmente, hemos compilado un corpus de textos orales de dos libros de alemán general como lengua extranjera, que nos ha servido para establecer comparaciones.

Aunque el hecho de que ambos corpus estén formados por materiales del nivel A2 se puede considerar una limitación, creemos que de hecho es una ventaja haber contado con textos orientados a un mismo nivel de referencia, ya que esto nos permite realizar contrastes entre textos que (supuestamente) cuentan con estructuras gramaticales parecidas y están diseñados para adquirir las mismas competencias.⁶

Hemos constatado que hay más *phrase frames* (de 3, 4, 5 y 6 palabras) en el Corpus 1 que en el Corpus 2, lo que apunta a que el discurso oral turístico, al menos tal como se presenta en los libros de texto, está más convencionalizado que el discurso oral general. Los constituyentes libres de las estructuras fijas que hemos seleccionado pueden completarse con una palabra o con varias. Hemos comprobado, por otro lado, que el número de variantes no es uniforme, pues algunos *phrase frames* son muy productivos y admiten varias variantes, mientras que otros apenas pueden modificarse con dos constituyentes diferentes; el VPR medio es de 53,99% (mediana: 50; desviación estándar: 22,67). Observamos que ambos corpus presentan un cociente entre variantes y estructuras fijas descendente, de manera que a medida que aumenta el número de constituyentes de los *phrase frames*, desciende el número de variantes.

Las estructuras fijas seleccionadas (de cuatro palabras) presentan en su mayoría unas funciones comunicativas muy concretas: predominan, por un lado, las del tipo 1.1 (ofrecer y buscar información factual), sobre todo en forma de preguntas, peticiones de información y respuestas; y por otro, del tipo 1.2 (expresar y descubrir actitudes), con expresiones de volición (deseos), modalidad (permiso) y emociones.

Los *phrase frames* seleccionados, por otro lado, suelen presentar la distribución A B C *, con unos pocos del tipo A B * D. Aunque no hemos mostrado el resto de *phrase frames* detectados, estos siguen mayoritariamente el mismo patrón, de manera que predominan los del tipo A B C * o * B C D.

La comparación de las estructuras fijas encontradas en el Corpus 1 con aquellas presentes en el Corpus 2 arroja unos resultados significativos: de las 20 estudiadas, 15 no se encuentran en el Corpus 2; mientras que las cinco que sí encontramos presentan una frecuencia de aparición mucho menor. Esto apunta a que el discurso oral turístico sería un discurso especializado, a pesar de que pueda parecer que se presta a muchas similitudes con el discurso general. Fuster (2014) llega a una conclusión muy parecida al estudiar los *lexical bundles* y *phrase frames* de un corpus en inglés de páginas web hoteleras.

El posible (alto) grado de convencionalización de este tipo de discurso oral y su carácter de especialidad tienen consecuencias relevantes para su enseñanza y aprendizaje. Estos resultados, aunque preliminares, apuntan a que sería conveniente dar más peso a las estructuras fijas en la enseñanza del alemán turístico, y promover

⁶ Véase el descriptor global del nivel A2 del MCER (2002: 26): El alumno “(e)s capaz de comprender frases y expresiones de uso frecuente relacionadas con áreas de experiencia que le son especialmente relevantes.”

la práctica y repetición de distintos patrones y de sus variantes más frecuentes. Varios estudios han demostrado que las expresiones hechas o unidades pluriverbales (*formulaic language*) presentan ventajas de procesamiento, y que los aprendices son capaces de extrapolar información lingüística a partir de este tipo de unidades (véanse los estudios citados por Vandeweerd y Keijzer, 2018). Esperamos poder ampliar los resultados y discusión de este estudio con un corpus de mayor tamaño y con textos no solo provenientes de materiales didácticos.

Bibliografía

1. Biber, D.: A corpus-driven approach to formulaic language in English multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3), 275–311 (2009).
2. Bolinger, D.: Meaning and memory. *Forum Linguisticum* 1, 1–14 (1976).
3. Cohen, U., Lemcke, C.: *Herzlich willkommen Neu*. Langenscheidt, Berlín (2001).
4. Consejo de Europa / Ministerio de Educación, Cultura y Deporte: Marco Común Europeo de Referencia para las Lenguas: Aprendizaje, Enseñanza, Evaluación. Secretaría General Técnica del MECD-Subdirección General de Información y Publicaciones, y Grupo Anaya, Madrid (2002).
5. De Cock, S.: Repetitive phrasal chunkiness and advanced EFL speech and writing. In: Mair, C., Hundt, M. (eds.) *Corpus Linguistics and Linguistic Theory*. Papers from ICAME 20, pp. 51–68. Rodopi, Ámsterdam (2000).
6. Fletcher, W. H.: *kfNgram* [software], <http://www.kwicfinder.com/kfNgram/kfNgramHelp.html>, último acceso 10/05/2019 (2007).
7. Fuster Márquez, M. Lexical bundles and phrase frames in the language of hotel websites. *English Text Construction* 7(1), 84–121 (2014).
8. Goldberg, A.: *Constructions at Work*. Oxford University Press, Oxford (2006).
9. Granger, S., Meunier, F.: Phraseology in language learning and teaching: where to from here? In: Meunier, F., Granger, S. (eds.) *Phraseology in Foreign Language Learning and Teaching*, pp. 247–252. John Benjamins, Amsterdam & Philadelphia (2008).
10. Granger, S., Paquot, M.: Disentangling the phraseological web. In: Granger, S., Meunier, F. (eds.) *Phraseology: An Interdisciplinary Perspective*, pp. 27–50. John Benjamins, Amsterdam (2008).
11. Morley, J.: About Academic Phrasebank, <http://www.phrasebank.manchester.ac.uk/about-academic-phrasebank/>, último acceso 3/05/2019 (2019).
12. O’Keeffe, A., McCarthy, M., Carter, R.: *From Corpus to Classroom: language use and language teaching*. Cambridge University Press, Cambridge (2007).
13. Pawley, A., Syder, F. H.: Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: Richards, J. C., Schmidt, R. W. (eds.) *Language and Communication*, p. 191–226. Longman, Nueva York (1983).
14. Römer, U.: The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics* 7, 141–163 (2009).
15. Römer, U.: Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3(1), 95–119 (2010).
16. Ros-El Hosni, L., Swerlowa, O., Klötzer, S., Jentges, S., Sokolowski, K., et al.: *Aussichten A2.1*. Klett, Stuttgart (2010).

17. Sander, I., Braun, B., Doubek, M., Frater-Vogel, A., Fügert, N., Köhl-Kuhn, R., Trebesius-Bensch, U., Vitale, R., Behnes, S., Marquardt-Langermann, M.: DaF kompakt A2. Deutsch als Fremdsprache für Erwachsene. Klett, Stuttgart (2011).
18. Sinclair, J.: Corpus, Concordance and Collocation. Oxford University Press, Oxford (1991).
19. Schümann, A., Schurig, C., Schaefer, B., van der Werff, F. Menschen im Beruf Tourismus A2. Hueber, München (2015).
20. Vandeweerd, N., Keijzer, M. J'ai l'impression que: Lexical Bundles in the Dialogues of Beginner French Textbooks. *Canadian Journal of Applied Linguistics* 21(2), 80–101 (2018).
21. Wray, A.: Formulaic Language and the Lexicon. Cambridge University Press, Cambridge (2002).

Desarrollo de la Fraseología Especializada en Brasil

Cleci Regina Bevilacqua¹

¹ Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil
cleci.bevilacqua@ufrgs.br

Abstract: This article presents an overview of research related to Specialized Phraseology conducted in Brazil in recent years. Various denominations and definitions with respect to our object of study, namely Specialized Phraseological Units (SPU), were collected from Brazilian academic publications (dissertations, theses, specialized journals, books, events, etc.) and systematized. As a synthesis, the two identified types of SPU and their denominations were organized into a concept map. Applications (terminographical products, semi-automatic identification) and the thematic areas studied (Medicine, Environmental Law, Economics, etc.) are also presented. The justification for the study is grounded in the need to find and gather information about the research carried out within the scope of Specialized Phraseology, which allowed us to draw an initial overview of the studies in Brazil. As a result, we hope to be able to disseminate Brazilian scientific publications on the subject.

Keywords: Specialized Phraseology, Specialized Phraseological Units, Brazilian Research.

1 Introducción

El presente trabajo presenta un panorama de las investigaciones relativas a la Fraseología Especializada llevadas a cabo en Brasil. Dichas investigaciones empiezan a desarrollarse a finales de los años de 1990, motivadas sobre todo por la introducción de la Terminología, ámbito en el que se suele incluir la Fraseología Especializada. Eso ocurrió principalmente por la consolidación de grupos de investigación y por la inserción de disciplinas de Terminología en cursos de pregrado y posgrado en Traducción.

Considerando estos aspectos, a partir de la compilación y revisión de la producción científica brasileira de los últimos 15 años, se pudo identificar la existencia de diferentes perspectivas (Terminología, Traducción, Lingüística de Corpus y Procesamiento del Lenguaje Natural) en los estudios relativos al tema y a su objeto de estudio – las Unidades Fraseológicas Especializadas (UFEs) –, que generan distintas denominaciones y definiciones para dicho fenómeno. Se busca identificar esas denominaciones y definiciones asociándolas a sus autores y a los temas y áreas tratados.

Inicialmente, se hace referencia a las fuentes de recogida de las informaciones. Enseguida, se sistematizan las distintas denominaciones y definiciones relativas a las UFEs, buscando mostrar los diferentes tipos y sus principales características. Como síntesis de esta etapa, se presenta un mapa conceptual con la síntesis de los resultados

obtenidos. También se muestra la aplicación de los estudios realizados en el área y a los temas estudiados.

Algunas de las justificaciones que sostienen el trabajo son: el interés por ese tema para la traducción y producción de textos especializados, así como la extracción de UFEs a partir de corpus especializados y su aplicación en la enseñanza de la traducción y de lenguas para propósitos específicos; la ausencia de un relevamiento sobre las investigaciones realizadas en Brasil. Se espera poder contribuir a la construcción de dicho panorama, aunque de forma inicial, y divulgar la producción científica brasileña relativa a la Fraseología Especializada.

2 Las fuentes para la recogida de los datos

Para la recogida de los datos, se hizo un relevamiento general en el Portal de Periódicos de Capes, en repositorios digitales de universidades y en *Google*, utilizándose palabras clave como *colocação(ões) especializada(s)*, *combinatórias léxicas especializadas*, *expressões multipalavras*, *expressões recorrentes*, *fraseologia especializada*, *fraseologismos*, *unidades fraseológicas especializadas*. También se buscaron informaciones en páginas web de proyectos y de grupos de investigación, asociaciones, eventos del área y de áreas afines, libros y revistas nacionales relacionados al estudio del léxico. Consideramos, además, nuestro conocimiento del área y el contacto con los investigadores brasileños.

Algunas de las fuentes consultadas fueron: Universidade de Brasília (UnB), Universidade Estadual Paulista Júlio Mesquita, campus de São José do Rio Preto (UNESP), Universidade Estadual do Ceará (UECE), Universidade Federal do Mato Grosso do Sul (UFMS), Universidade Federal do Rio Grande do Sul (UFRGS), Universidade de São Paulo (USP); proyecto COMET (USP)¹, Projeto Terminológico Cone Sul (TERMISUL, UFRGS)², TexQuim/TexTECC(UFRGS)³, Grupo de Trabalho em Lexicologia, Lexicografia e Terminologia da Associação Nacional de Pesquisa em Letras e Linguística (GTLex-ANPOLL)⁴, Rede Ibero-americana de Terminologia (RITerm) e Rede Panlatina de Terminologia (REALITER)⁵; Congresso Internacional de Fraseologia e Paremiologia, Encontro Nacional de Tradutores, Encontro de Linguística de *Corpus* (ELC); libros (*Avanços na Linguística de Corpus no Brasil*; TAGNIN; VALE, 2008; *Tendências atuais na pesquisa descritiva e aplicada em fraseologia e paremiologia*; ALVAREZ, 2012); publicaciones periódicas (*Alfa*⁶, *Debate Terminológico*⁷, *Domínios da Linguagem*⁸, *TradTerm*⁹).

¹ <http://comet.fflch.usp.br/>

² <http://www.ufrgs.br/termisul/>

³ <http://www.ufrgs.br/textecc/textquim/>; <http://www.ufrgs.br/textecc/>

⁴ <http://www.letras.ufmg.br/gtlex/>

⁵ <https://sites.google.com/view/riterm/riterm>; <http://www.realiter.net/>

⁶ <http://seer.fclar.unesp.br/alfa>

⁷ <http://seer.ufrgs.br/riterm>

⁸ <http://www.seer.ufu.br/index.php/dominiosdelinguagem>

⁹ <http://www.revistas.usp.br/tradterm>

3 Las UFEs: denominaciones y definiciones

El análisis de los datos permitió identificar distintas denominaciones y definiciones para las UFEs, que se agruparon en dos grandes grupos descritos a continuación.

1) UFEs caracterizadas por estructuras semejantes a las colocaciones, es decir, están formadas por una base y un colocado (la unidad que acompaña la base) y pueden describirse según determinadas estructuras morfosintácticas. Esta perspectiva se basa sobre todo en la propuesta de colocación de Hausmann (1979,1990) para las colocaciones de la lengua general y de L’Homme (1998, 2000) y de L’Homme y Bertrand (2000) para los lenguajes de especialidad. Algunos ejemplos son: *adquirir ações, ações subscritas* (ORENHA-OITTANO, 2009), *absorver calor, absorção de calor, calor absorvido* (BEVILACQUA, 2004). La revisión de los autores mostró que hay varias denominaciones utilizadas para este tipo de UFE: *colocação especializada, UFEs, UFEs eventivas, Combinatórias Léxicas Especializadas (CLEs) y expressões recorrentes*.

Algunos de los investigadores que siguen esta perspectiva son: Bertonha e Zavaglia (2015), Bevilacqua (1999, 2004); Bevilacqua et al (2013), Castanho (2011), Cruz (2011), Esperandio (2015), Ferreira (2015), Leinritz (2013), Orenha-Ottaiano (2009), Pacheco (2015), Santiago (2013), Waquil (2013) y Zilio (2009).

2) UFEs caracterizadas como fórmulas, que pueden llegar a frases autónomas o a párrafos completos. Esta perspectiva se fundamenta en las propuestas de autores como Roberts (1994-1995), Parc (1993) y Gouadec (1994). Algunos ejemplos pueden ser: *O não cumprimento de [x] sujeita [y] a [z]*, donde [x] puede ser *lei, decreto, parágrafo*; [y] a los *infrator(es)* y [z] a las *punições (medida, cautelar, advertência, multa, penalidades)*. Las denominaciones encontradas fueron: *colocações especializadas estendidas, combinatórias léxicas especializadas jurídicas, matrizes fraseológicas sem pivô terminológico*. Algunos de los autores que siguen esta perspectiva son: Bevilacqua (1996), Tagnin (2012), Orenha-Ottaiano (2009).

La síntesis de los tipos y denominaciones de UFEs se encuentra en el mapa conceptual presentado en la Figura 1.

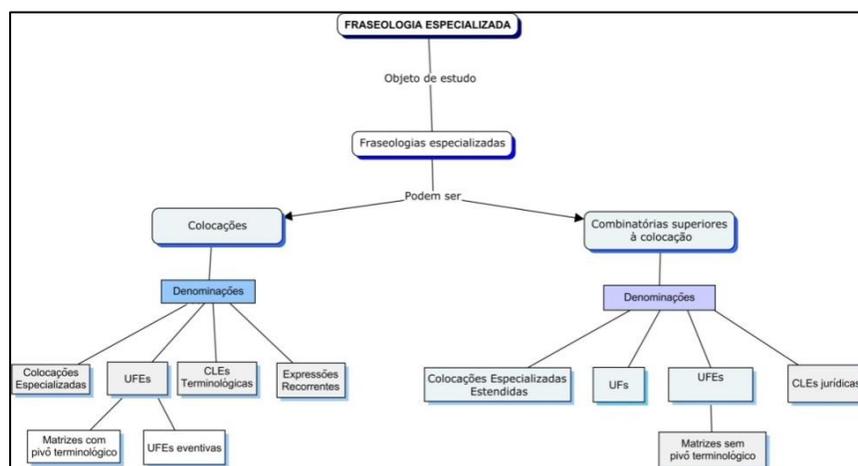


Fig. 1 – Mapa conceptual: tipos y denominaciones de UFEs.

4 Las aplicaciones, temas y áreas investigados

Respecto de las aplicaciones se identificaron dos tendencias: una dirigida a la aplicación y otra a la descripción. Para la primera las temáticas evidencian la interfaz entre: a) la Fraseología Especializada y la Lingüística de Corpus, que trata temas como la constitución de corpus y las herramientas para la extracción e identificación de las UFEs y la búsqueda de equivalentes de las UFEs en lenguas extranjeras; b) la Fraseología Especializada y la Traducción, en la que se destacan temas como la relevancia de las UFEs en el proceso de traducción, las propuestas de diseño de recursos terminológicos para traductores (glosarios, bases de datos) y la identificación de equivalentes en las lenguas de trabajo del traductor.

En lo que se refiere a la descripción, los temas tratados son: la aplicación de modelos lingüísticos para la descripción de las UFEs, aspectos diacrónicos; descripción sintáctico-semántica; equivalencia del portugués para distintas lenguas extranjeras; metáforas, patrones léxico-gramaticales.

Las áreas temáticas tratadas en los trabajos analizados se refieren a: Cardiología, contratos y reglamentos sociales, Culinaria, Derecho Ambiental, Derecho Comercial internacional, Economía, Educación, Educación a Distancia, exportación, fútbol, Medicina, Pediatría, publicidad, Traducción, Traducción Pública, Turismo, etc.

Los datos permitieron destacar que hay una interrelación entre la Fraseología Especializada con los Estudios de Traducción y la Lingüística de Corpus, estableciendo un interdisciplinariedad entre dichas áreas. Asimismo, se puede afirmar que la Fraseología Especializada también es transdisciplinar puesto que se aplica a varios ámbitos del saber – Derecho, Medicina, Economía, etc.

5 Conclusiones

A partir de las informaciones recogidas y de su sistematización se buscó presentar un panorama inicial de la Fraseología Especializada en Brasil. El análisis de los datos reveló que hay una diversidad denominativa para referirse a las UFEs y que, aunque sus propiedades sean comunes y sus conceptos sean muy cercanos, se identificaron dos grandes categorías establecidas principalmente por sus estructuras: las colocaciones especializadas y las fórmulas. Esas categorías están conformes a las propuestas de autores reconocidos en el área, tanto los que se dedican a la fraseología de la lengua general como a la fraseología especializada. Para cada categoría se indicaron sus denominaciones y los autores brasileños. Además, se identificaron las aplicaciones, temáticas y áreas sobre las que se estudian las UFEs, mostrando que la Fraseología Especializada es a la vez interdisciplinar y transdisciplinar.

Con el conjunto de datos relevados, se puede decir que hay una producción académica significativa sobre la Fraseología Especializada en Brasil y la búsqueda por nuevos datos puede revelar que esa producción quizás sea aún mayor. Sin embargo, parece ser que la sistematización de los datos encontrados permitió construir una visión general inicial de los estudios del área en Brasil y, a la vez, posibilitó organizar la terminología utilizada para referirse a las UFEs.

Referências

1. Alvarez, M. L. O. (ed.): *Tendências atuais na pesquisa descritiva e aplicada em fraseologia e paremiologia*, vol. I e II. Pontes, São Paulo (2012).
2. Bevilacqua, C. R.: *A fraseologia jurídico-ambiental*. Tesis. (Maestría en Estudios del Lenguaje). Programa de Posgrado en Letras, Instituto de Letras, UFRGS, Porto Alegre (1996).
3. Bevilacqua, C. R.: Unidades fraseológicas especializadas: estado de la cuestión y perspectivas. Tesina. (Doctorado en Lingüística Aplicada – Léxico). Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona (1999).
4. Bevilacqua, C. R.: *Unidades Fraseológicas Especializadas Eventivas: descripción y reglas de formación en el ámbito de la energía solar*. Tesis. (Doctorado en Lingüística Aplicada – Léxico). Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona (2004).
5. Bevilacqua, C. R. et al: Combinatórias Léxicas Especializadas da Linguagem Legislativa: uma abordagem orientada pelo *corpus*. En: Murakawa, C., Nadin, O. L. (eds.) *Terminologia: uma ciência interdisciplinar*, pp.227-243. Cultura Acadêmica, São Paulo (2013).
6. Bertonha, F. H. C., Zavaglia, C.: Dicionário bilíngue de fraseologismos jurídicos: uma proposta. En: Zavaglia, C., Stupiello, E. (eds.) *Tendências Contemporâneas dos Estudos da Tradução*, vol. 2, pp. 36-64. Ed. UNESP, São José do Rio Preto (2015).
7. Castanho, R. M. C.: *Proposta para a Elaboração de um Glossário de Colocações na Área Médica - Subárea Hipertensão Arterial*. Tesis (Maestría en Estudios Lingüísticos y Literarios en Inglés). USP, São Paulo (2003).
8. Cruz, L. P. R.: *Estudo da tradução de colocações especializadas na área de exportação e agronegócios baseado em um corpus paralelo*. Tesis (Maestría en Estudios Lingüísticos). UNESP, São José do Rio Preto (2011).
9. Esperandio, I. B.: Colocações na legendagem de seriados: Um estudo exploratório. En: Zavaglia, C. Stupiello, E. (eds.) *Tendências Contemporâneas dos Estudos da Tradução*, vol. 2, pp. 139-164. UNESP, São José do Rio Preto (2015).
10. Gouadec, D.: Nature et traitement des entités phraséologiques. *Terminologie et phraséologie*. En: *Acteurs et aménageurs: Actes de la deuxième Université d'Automne en Terminologie*, pp. 167-193. La Maison du Dictionnaire, Paris (1994).
11. Ferreira, V. R.: *Glossário terminológico básico da teoria da tradução: uma experiência com o e-terms*. Tesis (Mestría en Estudios del Lenguaje). Universidade Federal do Mato Grosso do Sul, Campo Grande (2015).
12. Hausmann, F. J.: Un dictionnaire de collocations est-il possible? *Travaux de linguistique et de littérature*, 17 (1), 187-195 (1979).
13. Hausmann, F. J.: Le dictionnaire de collocations. En: Hausmann, F. J. et al. (eds.) *An International encyclopedia of lexicography*, vol. 1, p. 1010-1019. Walter de Gruyter, Berlin- New York (1990).
14. Leipnitz, L.: Fraseologias terminológicas no ensino da tradução. En: Tagnin, S. E. O.; Bevilacqua, C. R. (eds.) *Terminologia*, pp.113-127. HUB, São Paulo (2013).

15. L'Homme, M-C.: Caractérisation des combinaisons lexicales spécialisées par rapport aux collocations de langue générale. En: Fontenelle, T. et al (eds.) *Euralex '98 Proceedings*, vol. II, pp. 513-522. Université de Liège, Liège (1998).
16. L'Homme, M-C.: Understanding Specialized Lexical Combinations. *Terminology* 6(1), 89-110 (2000).
17. L'Homme, M-C., Bertrand, C.: Specialized Lexical Combinations: should they be described as collocations or in terms of selectional restrictions? Página web, <http://olst.ling.umontreal.ca/pdf/berlhom.pdf> , último acceso 2015/10/15.
18. Orenha-Ottaiano, A.: *Unidades fraseológicas especializadas: colocações e colocações estendidas em contratos sociais e estatutos sociais traduzidos no modo juramentado e não-juramentado*. Tesis. (Doctorado en Estudios Lingüísticos). UNESP, São José do Rio Preto (2009).
19. Pacheco, S. A.: *Configurações sintático-semânticas das unidades fraseológicas especializadas: o caso do léxico do exército brasileiro*. Tesis. (Doctorado en Estudios de Lengua). Programa de Posgrado en Letras, Instituto de Letras, UFRGS, Porto Alegre (2015).
20. Parc, F.: Traitement de la phraséologie terminologique tirée des textes législatifs et réglementaires suisses. *Terminologies Nouvelles*, 10, 115-119 (1993).
21. Roberts, R.: Identifying the phraseology of languages for special purposes (LSPs). *Alfa: Actes de langue française et de linguistique*, vol. 7/8, pp. 61-69. Universitat Dalhousiana, Halifax, (1994-1995).
22. Santiago, M.S.: *Unidades fraseológicas especializadas em tutoriais de ambientes virtuais de aprendizagem: proposta de um sistema classificatório com base na valência verbal*. Tesis (Doctorado en Estudios del Lenguaje). Programa de Posgrado en Letras, Instituto de Letras, UFRGS, Porto Alegre (2013).
23. Tagnin, S. E. O.: Fraseologia especializada para tradutores: glossários direcionados pelo corpus. En: Alvarez, M. L. O. (ed.) *Tendências atuais na pesquisa descritiva e aplicada em fraseologia e paremiologia*, vol. 1, pp. 333-344. Campinas, Pontes (2012).
24. Tagnin, S. E. O., Vale, O. A. (eds.): *Avanços da Linguística de Corpus no Brasil*. Humanitas, São Paulo (2008).
25. Waquil, M. L.: *Tradução de textos especializados: unidades fraseológicas especializadas e técnicas tradutórias*. Tesis (Maestría en Estudios del Lenguaje). Programa de Posgrado en Letras, Instituto de Letras, UFRGS, Porto Alegre (2013).
26. Zilio, L.: *Colocações especializadas e 'Komposita': um estudo contrastivo alemão-português na área de cardiologia*. Tesis. (Maestría en Estudios del Lenguaje). Programa de Posgrado en Letras, Instituto de Letras, UFRGS, Porto Alegre (2009).

Orthography in Practice: Corpus-based Verification of Writing Ktetics in MWUs in Croatian

Goranka Blagus Bartolec^[0000-0002-3577-7026] and Ivana Matas Ivanković^[0000-0002-9796-8346]

Institute of Croatian Language and Linguistics, Republike Austrije 16, 10000 Zagreb, Croatia
gblagus, imatas@ihjj.hr

Abstract. Writing upper and lower letter at the beginning of the word is one of the key orthographic problem in Croatian. The two main goals of this research are: 1 to determine whether there are frequent mistakes in writing upper and lower letter in ktetics (adjectives derived from geographic names) in corpus texts, 2 to use corpus tools (regular expressions) for orthographic analysis. According to [1: 112] upper and lower letter errors are defined as spelling errors and can be counted in one of the most common types of orthographic errors in Croatian. The emphasis will be on ktetics as components of various types of MWUs in which the orthographic rules prescribe a lower initial letter. Based on the corpus hrWaC 2.2 (using the Sketch Engine platform), the degree of deviation in the initial letter writing will be determined with regard to the meaning of the individual MWUs. The analysis includes four MWUs that denote different contents, in which the first component is ktetic; *zagrebačka katedrala* ‘lit. Zagrebian cathedral, Eng. Zagreb cathedral’ (sacral object), *osječko sveučilište* ‘lit. Osijekian University, Eng. University of Osijek’ (official institution), *sibirski haski* ‘Siberian Husky’ (dog breed), and *bečki odrezak* ‘Wiener schnitzel’ (types of steak). Given the results obtained, the conclusion will cover both: the types of corpus sources within which orthographic errors are most commonly as well as the possible causes of such errors. The research was made for the project *Rječnik velikoga i maloga početnog slova* ‘Dictionary of upper and lower initial letter’ which is being developed at the Institute of Croatian Language and Linguistics in Zagreb.

Keywords: Croatian, hrWaC, Lower Letter, Orthography, Upper Letter

1 Orthographic Rules and Errors in Corpus Texts

The rules of writing upper initial letter in Croatian, as well as in most other Slavic and European languages, primarily relate to two problems: writing upper initial letter at the beginning of the sentence and writing proper names - personal names and geographic names. These rules have been adopted by the speakers of the Croatian language already at the beginning of the schooling and are mainly used in written practice in the proper manner. In addition to these rules, Croatian orthography also includes other different rules that prescribe the writing upper or lower initial letter in

single words or MWUs depending on their sematic content - upper initial letter is used to describe the names of individual living beings, objects, historical events or geographic objects and areas, and words or MWUs with descriptive meaning in general use or denoting terms that belong to professional or scientific taxonomy are written in a lower initial letter. A large group of such MWUs are those in which the first component is ktetic - an adjective derived from a geographic place/object name (e.g., *zagrebački* 'Zagrebian, related to Zagreb', *pariški* 'Parisian, related to Paris', *švicarski* 'Swiss, related to Switzerland'). As the orthographic rules stipulates, ktetics at the beginning of multiword geographic names should be written in upper initial letter (*Zagrebačka gora* 'Zagrebian mountain' (a mountain north of Zagreb), *Pariška zaval* 'Parisian Basin', *Švicarska Konfederacija* 'Swiss Confederation'). At the beginning of multiword terms and in a general use, ktetics should be written in lower initial letter: *zagrebačka slavistička škola* 'lit. Zagrebian philological school, Eng. Zagreb philological school', *pariški odrezak* 'Steak Parisian', *švicarski franak* 'Swiss franc'). In written practice, however, these rules are often ignored. Wrong written records are common in writing of MWUs for which the orthographic rules prescribe a lower initial letter, but, besides the regular record, there are frequent records with a upper initial letter.

The corpus as a computer collection of written texts of different styles provides a good insight into the state of practice by showing how much the written record follows and how far it varies from the orthographic rules. According to classification made by [1: 112], upper and lower letter errors are type of spelling errors. In addition to spelling errors, punctuation, lexico-sematic errors, stylistic errors, typographical errors are also part of classification of errors mentioned in [1: 112]. All those types of errors are categorized as the errors made by humans and, as part of corpus texts, they can be detected in corpus search. Here, then, we start from the fact that in the case of orthographic errors made by a human, the corpus is not the source of these errors, but is only the platform on which those errors can be collected. In this context, the four limitations described by [2: 22–23] can be taken into account and which users should be aware of if they are using corpus results: "1 A corpus will not give information about whether something is possible or not, only whether it is frequent or not. (...) 2 A corpus can show nothing more than its own contents. (...) all attempts to draw generalizations from a corpus are in fact extrapolations. (...) 3 A corpus can offer evidence but cannot give information. (...) The corpus simply offers the researcher plenty of examples; only intuition can interpret them. 4 Perhaps most seriously a corpus presents language out of its context." Assertions described under 1 and 2 are considered as basic in this research.

2 Frequency Analysis: Ktetics in Corpus Texts

For the analysis, four examples of MWUs with ktetic as the first component were selected: *zagrebačka katedrala* 'lit. Zagrebian cathedral, Eng. Zagreb cathedral' (sa-

cral object), *osječko sveučilište* ‘lit. Osijekian University, Eng. Osijek¹ University’ (official institution), *sibirski haski* ‘Siberian husky’ (dog breed), and *bečki odrezak* ‘Wiener schnitzel’ (type of steak). This research was made for the project *Rječnik velikoga i maloga početnog slova* ‘*Dictionary of upper and lower initial letter*’ which is being developed at the Institute of Croatian Language and Linguistics in Zagreb. Selected examples are prototypical in four thematic contexts of written practice where similar MWUs are frequent: administration, news, culinary (recipes and food), and pets. An objective evaluation of the use of orthographic rules in written practice would be obtained by analyzing a wider range of queries than these four. However, the primary intention here is to show how much corpus tools contribute to a specific search that includes the distinction of upper and lower initial letters. Generally, the use of corpus helps the work at the *Dictionary* for determining the degree of deviation from the rules in the writing of particular content that require upper or lower initial letter. According to given data, it is possible to explain the writing of such content in the dictionary more systematically and include as many such examples in the list of entries.

In Croatian, the way of writing of these multiword units can be subsumed under one of the four specific orthographic rules for writing initial lowercase depending on their meaning: 1 rule for writing sacral objects, 2 rule for writing colloquial names of official institutions, 3 rule for writing zoological or botanical species, and 4 rule for writing kinds of food. For this research, we used the *Croatian web corpus* – hrWaC 2.2² (on Sketch Engine platform) which, as interpreted in [3: 2], has the features of the general and reference corpus and represents language or variety as a whole (vs. specialized corpora). hrWaC was selected for this research as the biggest available corpus of Croatian [4: 30] containing more than 2 billion words.³ This corpus contains different text genres (newspaper texts, administrative and legislative texts, blogs, forums) that are indicators of different styles and levels of use of the Croatian standard language. Because of a wider insight into the research problem, the whole corpus was searched, so it is not used the option Text types functionality. Using the hrWaC, only multiword units in non-initial position in a sentence were selected in order to avoid examples with upper initial letter at the beginning of the sentence. Such a search has included two basic regular expressions [5], depending on whether we searched for attestations with upper or lower initial letter (Table 1). Regular expressions have varied with respect to what we wanted to get for a particular multiword unit.

¹ The city in eastern Croatia.

² https://old.sketchengine.co.uk/corpus/first_form?corpname=preloaded/hrwac22_rft1; last accessed 2019/04/19.

³ There are two other corpora for the Croatian language: *Hrvatski nacionalni korpus / Croatian National Corpus* (http://filip.ffzg.hr/cgi-bin/run.cgi/first_form) as a general corpus, and *Hrvatska jezična riznica / Croatian Language Repository* (<http://riznica.ihjj.hr/>) which is limited to newspaper and literary texts. These two corpora are also available on the Sketch Engine platform.

Table 1. Regular expressions for MWUs *ktetics* + *noun* with lower or upper letter form

Form of initial letter	Regular expression
Upper initial letter	[word!="\."][lemma="ktetic" & word="[A-Z]*"][lemma="noun" & word="[a-z]*"]
Lower initial letter	[word!="\."][lemma="ktetic" & word="[a-z]*"][lemma="noun" & word="[a-z]*"]

2.1 MWU *zagrebačka katedrala*

The orthographic rule [6: 38] stipulates that types of sacral objects (church, cathedral, mosque, synagogue, temple) should be written in a lower initial letter. According to this rule, it is correct to write *zagrebačka katedrala*. Attestations for forms *Zagrebačka katedrala* i *zagrebačka katedrala* were selected in hrWaC using regular expressions:

```
[word!="\."][lemma="zagrebački" & word="[A-Z].*"][lemma="katedrala" & word="[a-z].*"]
```

```
[word!="\."][lemma="zagrebački" & word="[a-z].*"][lemma="katedrala" & word="[a-z].*"]
```

Frequency results are presented in Table 2.

Table 2. Results for MWU *zagrebačka katedrala* in hrWaC

Form of writing	Frequency	%
<i>Zagrebačka katedrala</i>	441	20
<i>zagrebačka katedrala</i>	1791	80

2.2 MWU *osječko sveučilište*

Descriptive colloquial forms of public institutions which are not in official use according to orthographic rule should be written in lower initial letter [6: 30]. Since the official name of the university in Osijek is *Sveučilište Josipa Jurja Strossmayera u Osijeku* ‘Josip Juraj Strossmayer University of Osijek’, MWU *osječko sveučilište* is colloquial and unofficial form and should be written in a lower initial letter. Attestations for forms *Osječko sveučilište* i *osječko sveučilište* were selected in hrWaC using regular expressions:

```
[word!="\."][lemma="osječki" & word="[A-Z].*"][lemma="sveučilište" & word="[a-z].*"]
```

[word!="\."][lemma="osječki" & word="[a-z].*"][lemma="sveučilište" & word="[a-z].*"].

Frequency results are presented in Table 3.

Table 3. Results for MWU *osječko sveučilište* in hrWaC

Form of writing	Frequency	%
<i>Osječko sveučilište</i>	71	36
<i>osječko sveučilište</i>	127	64

2.3 MWU *sibirski haski* and *bečki odrezak*

The orthographic rule [6: 36] stipulates that zoological species and kinds of food should be written in a lower initial letter. According to this rule, proper written records are *sibirski haski* (for dog breed) and *bečki odrezak* (type of steak). Attestations for forms *sibirski haski*, *Sibirski haski*, *bečki odrezak*, and *Bečki odrezak* were selected in hrWaC using regular expressions:

[word!="\."][lemma="sibirski" & word="[A-Z].*"][lemma="haski" & word="[a-z].*"]

[word!="\."][lemma="sibirski" & word="[a-z].*"][lemma="haski" & word="[a-z].*"]

[word!="\."][lemma="bečki" & word="[A-Z].*"][lemma="odrezak" & word="[a-z].*"]

[word!="\."][lemma="bečki" & word="[a-z].*"][lemma="odrezak" & word="[a-z].*"]

Frequency results are presented in Table 3.

Table 3. Results for MWU *sibirski haski* and *bečki odrezak* in hrWaC

Form of writing	Frequency	%
<i>Sibirski haski</i>	18	27
<i>sibirski haski</i>	48	73
<i>Bečki odrezak</i>	12	10
<i>bečki odrezak</i>	111	90

3 Ktetics in Corpus Texts: Examples of Good or Bad Orthographic Practice?

Based on results obtained for all four MWUs it is evident that written attestations which confirm orthographic rules prevail in the corpus texts. Statistically, the correct writing of ktetics at the beginning of MWUs prevails in the range of 64% (*osječko sveučilište*) up to 90% (*bečki odrezak*). These frequency results confirm the relatively high level of correct writing of MWUs with the ktetics as first component. It is therefore possible to conclude the following:

1 corpus text writers know the orthographic rules that prescribe the writing of a upper or lower initial letter in the ktetics, but also other rules that include the contents of the MWUs included in this research, whether or not they contain ktetics. Speakers of the Croatian language in written practice often have doubts regarding the writing of sacral objects, especially buildings or places for different religious ceremonies, so it was expected that corpus texts would overwrite erroneous attestations, that is, with upper initial letter. The searching, however, confirmed that have attestations with the correct written records have higher frequency.⁴

2 It was also expected that the erroneous written records will prevail in writing colloquial forms of official institutions, because, in written practice, such forms are often taken as official although they are not.⁵ Here, we can distinguish two types of corpus texts as the key reason why correct orthographic records are prevalent in the obtained results - newspaper texts (Croatian daily newspapers) and official sites of various public institutions and religious communities. In general, in such texts both the higher level of spelling knowledge and the higher level of language culture is represented because it takes into account that the texts are aligned with the Croatian standard language. As the second reason, it is possible to point out the fact that interest in orthography has increased in Croatia in recent years because, apart from the printed orthographic manuals, there are numerous public available free internet pages with orthographic and language advice topics (e.g. <http://pravopis.hr/>, <http://bolje.hr/>, <http://jezicni-savjetnik.hr/>, <http://hjp.znanje.hr/>). In this way, one can quickly and easily reach the appropriate information on how to properly write.

3 Although corpus research has established that attestations of good orthographic practice are statistically predominant, corpus texts also contain written records that do not comply with spelling rules. Such written records are, as the study has shown, in a statistically lower ratio, but are not negligible - in the MWUs involved in this research, they range from 10% to 36%. Here are some possible explanations as to why, apart from the correct written records, there are also records that deviate from the orthographic rules: 1 when writing some contents, e.g. religious buildings, it is assumed that, if something is written with a upper initial letter, has a greater importance

⁴ Lower initial letter also prevails for *varaždinska katedrala* (488 results) 'Varažđian cathedral, Eng. Varažđin cathedral' in relation to *Varaždinska katedrala* (93 results).

⁵ Lower initial letter also prevails for *zadarsko sveučilište* (530 results) 'lit. Zadarian university, Eng. University of Zadar' and *riječko sveučilište* (191 results) 'lit. Rijekian university, Eng. University of Rijeka', over upper initial letter in *Zadarsko sveučilište* (443 results) and *Riječko sveučilište* (145). The results were reversed for the *zagrebačko sveučilište* (727 results) 'lit. Zagrebian university, Eng. University of Zagreb' i *splitsko sveučilište* (176 results) 'lit. Splitian university, Eng. University of Split', in relation to *Zagrebačko sveučilište* (1124 results) and *Splitsko sveučilište* (556 results).

in the individual minds of speakers, but actually is a deviation from orthographic rule, 2 when writing some contents, such as religious buildings or colloquial and unofficial names of official institutions, the speakers do not know or have not checked the official names of the institutions and write such names in the upper initial letter instead of lowercase letter, and 3 as a possible reason, the influence of orthographic rules from other languages (most commonly English and German) can be mentioned in writing some single words and MWUs. For this reason, some of the MWUs involved in this research (*bečki odrezak*, *sibirski haski*) sometimes are written in upper initial letter, which is contrary to the orthographic rules in Croatian for writing zoological species and types of dishes. For a stronger argumentation of this assertion, the parallel English and Croatian, and German and Croatian Corpora should certainly be included in the research, which was not carried out in our work.

4 Finally, writing upper and lower initial letter is an important orthographic problem in Croatian. Only four orthographically interesting MWUs were included as prototypes in this research. For more insight into the orthographic practice of the Croatian language based on the corpus, it is necessary to include much more MWUs. The corpus as a source of large collection of texts gives a wide insight into how orthographic rules for writing upper and lower initial letter are implemented in practice, i.e. in written use. Good corpus tools, such as regular expressions for upper and lower letter, facilitate the availability of the requested results, suggesting that the corpus-based approach becomes an integral part in orthographic researches.

References

1. Jakubiček, M., Bušta, J., Hlaváčková, D., Pala, K.: Classification of Errors in Text. In: RASLAN 2009: Recent Advances in Slavonic Natural Language Processing, pp 109–119. Masaryk University, Brno (2009).
2. Hunston, S.: Corpora in applied linguistics. Cambridge University Press, Cambridge (2002).
3. Nesselhauf, N.: Corpus Linguistics: A Practical Introduction, <http://www.as.uni-heidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf>, last accessed 2019/04.
4. Ljubešić, N., Klubička, F.: {bs,hr,sr}WaC –Web corpora of Bosnian, Croatian and Serbian. In: Proceedings of the 9th Web as Corpus Workshop (WaC-9), pp 29–35. Association for Computational Linguistics, Gothenburg (2014).
5. Regular expressions, <https://www.sketchengine.eu/user-guide/user-manual/concordance-introduction/regular-expressions/>, last accessed 2019/04.
6. Jozić, Ž.: Hrvatski pravopis. Institut za hrvatski jezik i jezikoslovlje, Zagreb (2013).

A Didactic Sequence for Phrasemes in L2 French

Maria Francesca Bonadonna^[1] and Silvia Domenica Zollo^[2]

¹ University of Verona, Lungadige Porta Vittoria, 41, 37129 Verona, Italy,
mariafrancesca.bonadonna@univr.it

² University of Verona, Lungadige Porta Vittoria, 41, 37129 Verona, Italy,
silvia.zollo@univr.it

Abstract. This paper deals with vocabulary teaching in French as a foreign language by focusing on the case of phraseological units. More precisely, it aims to design a didactic sequence for phrasemes within the theoretical framework of Explanatory Combinatorial Lexicology. Our study is based on a corpus-based experience of phraseology teaching conducted in a class of Italian university students learning French as a foreign language. Results indicated that, at the end of the course, students were able to describe the morphosyntactic structure and the use of phraseology in general or specialized language, but that they frequently made mistakes on these elements, in particular in writing, and that there was some confusion over the distinction between different types of phrasemes. Therefore, we propose a didactic sequence that focuses on a sample of phraseological units connected with discursive functions in argumentative texts and that includes the use of online resources in order to give a metalinguistic commentary. Our purpose is to draw a distinction between idioms and collocations, according to a progressive acquisition of metalexical competencies.

Keywords: Language Teaching; Teaching experiences; FLE; LEC; Phraseology.

1 Introduction

Lexical competence cannot be underestimated in the acquisition or learning of any foreign language. The lexicon is key to the development of language competence and can be built by working on multiple linguistic facets of words, whether phonetic, morphological, syntactic or semantic (Charmeux, 2010; Mel'čuk, Polguère, 2007; Picoche, 1993; Scott, Nagy, 2009). Some studies have focused on the more specific issues of teaching collocations (Lewis, 2000; Binon, Verlinde, 2004; Frassi, 2018; Tutin, Grossmann, 2002) and of phraseological units (Cavalla, Labre, 2009) by showing that their learning is necessarily the basis of any complete lexical training and that a void among L2 learners could constitute a real obstacle to understanding and producing a language. Promoting the learning of phraseology is therefore imperative.

In this paper we deal with the teaching of phrasemes in French as a foreign language. Our study is based on the results of a corpus-based experience on the teaching of phraseological units to a group of Italian-speaking students (level B1 of the CEFR)

studying a French Lexicology course at the second university cycle in Italy. Students had to learn their structure and usage in order to obtain the course certificate of achievement, the main assessment of which involved the writing of an argumentative essay with a metalinguistic dimension. Our experience shows that: 1. phraseological units are quite complex and varied lexical structures that still seem to be poorly mastered by Italian-speaking students. In classroom practices, Italian-speaking students frequently made mistakes on these elements, in particular in writing; 2. there seems to be some confusion over the distinction between different types of phrasemes.

After a brief theoretical-methodological overview of the didactics of the lexicon and phraseology within Explanatory Combinatorial Lexicology, we present the hypotheses and objectives that lead us to the realization of a didactic sequence (Cavanagh, Blain 2010; Cavanagh 2010) based on a systematic didactics of phraseology, on the use of ICTs¹ and on online corpora. We seek to link work on the lexicon and work on the argumentative text, by moving away from a “utilitarian” conception of phrasemes.

2 Theoretical Framework

The design of the didactic sequence is based on the principles of Explanatory Combinatorial Lexicology (ECL), which is the lexical module of the Meaning-Text Theory (MTT) (Mel’čuk, Polguère, 2007). According to this theory, lexicon is the central component of linguistic description and, as a result, of language teaching. A series of theoretical and methodological principles have been formulated around the structured teaching of lexicon and some tools have been developed for a practical use, both for first language and second language learning. The teaching of lexicon implies the development not only of lexical competence — the knowledge of lexical units — but also of metalinguistic and metalexical knowledge. This knowledge includes a set of notions about the structure itself of the lexicon, such as *meaning*, *polysemy*, *semantic compositionality*, etc. (Tremblay, 2009). We want to focus on the notion of phrasemes and, more precisely, on the distinction between full phrasemes and semi-phrasemes, which are rigorously defined in MTT. On the one hand, a full phraseme or idiom is a multilexemic lexical unit with a non-compositional meaning and a fixed syntactic structure: for example, *en souffrance* and *levée de boucliers* in French (Mel’čuk, Polguère, 2007, 27). On the other hand, a semi-phraseme or collocation is a combination of more lexical units, whether a lexeme or a full phraseme: the base, selected freely to express a particular meaning, and a collocate, selected as a function of the base. For example, the base *argument* can combine with the collocates *massue*, *de poids* in collocations like *argument massue*, *argument de poids* (Mel’čuk, Polguère, 2007, 20). While full phrasemes are recorded as separate lexical entries in ECL dictionaries, collocations do not receive a lexical entry.

Moreover, the learning of metalexical notions should follow a progression from simple to more complex notions (Mel’čuk, Polguère, 2007). For instance, Polguère

¹ Information and Communications Technology. Computer tools such as ICTs have been encouraged for several reasons, the first being related to the corpora used, which are digitized and therefore more easily accessible by computer. A second, more didactic reason is linked to the development of autonomy and learning, which is the main focus here.

and Tremblay have designed a five-step progression for future primary-school teachers (2014, 1184). After pointing out that language teaching practices and lexicographic tools make a poor use of collocations, Frassi has presented a progressive sequence for the teaching of collocation, and of its different types, based on its semantic and syntactic features (Frassi, 2018). In order to apply its theoretical principles, ECL has also created some practical tools that can be used for the teaching and learning of L1 and L2 French, such as the *Lexique actif du français* and the *DicoPop*.² Both theoretical reflections and lexicographic tools by ECL aim at text synthesis: this means that the viewpoint of Explanatory Combinatorial Lexicology on language and on language teaching is that of production or encoding.

3 Development of the Didactic Sequence

3.1 Hypotheses and Objectives

Our working hypothesis is as follows: starting from phraseological units, we can work on written production in a different way. In other words, students can be led to shift away from their usual practices (writing a list of phraseological units and “passively” memorizing them), which result in a very seldom use of phrasemes that are appropriate to the requested textual genre. Our aim is to invite the students to recognize and reuse phrasemes in their own productions. Another objective is to train on computer media for the use of linguistic phenomena which can contribute to improving the phraseological level of Italian-speaking students. We hence have two types of linguistic and didactic objectives: to develop the ability to locate phrasemes taken from a corpus of language science articles using the *ScienQuest* database³ and to describe them using metalexical competencies; to encourage the systematic implementation and recognition of phrasemes and their reuse, with a view to active memorization.

Our sequence is based on the distinction between phrasemes and semi-phrasemes. Although the borderline between these two categories is not always easy to draw, the learners will focus on their distinctive features, in order to understand them in a systematic way. It is not excluded that in future they will deal with the wide spectrum of phraseology and consider these categories in a more complex perspective, according to a progressive learning of metalexical notions.

3.2 Specific Needs of Learners and Methodological Approach

Obviously, the realization of such an activity requires the prior analysis of the specific needs of learners. At present, we have 54 argumentative essays produced by Italian-speaking students. The examples collected in these productions suggest that students are partially familiar with some phrasemes (most likely because they have read or

² *DiCoPop*, <http://olst.ling.umontreal.ca/dicopop> (Accessed in January, 2019).

³ The use of corpora, both for linguistic description and for the implementation of teaching activities, leads to considering these activities from the perspective of placing the lexical units to be taught in a permanent context. Thus, it is important for us to develop the use of corpora in French as a foreign language in order to multiply the examples of use of the language units taught.

heard them somewhere), but they need help to fix them. Here are some examples: *À différence du texte premier...*, *Dans ce texte, il y a la présence d'un phénomène de...*, *pour ce qui regarde la féminisation, nous voyons que...* In order to be effective, the learning of phrasemes must necessarily involve a reasoned and structured teaching of lexicon, according to syntactic, semantic and pragmatic criteria that contribute to the structural coherence of the argumentative essay. To do this, we adopt both an onomasiological and semasiological approach allowing the students to retain the meaning and form of the phraseme and its discursive function. On the didactic level, this approach will also be accompanied by a corpus extraction of the phrasemes.

3.3 The Didactic Sequence

The didactic path follows a standard progression where three main complementary steps are articulated and completed: the stage of exploration of phrasemes from a semantic point of view; the deepening of linguistic values and observation of the use of phrasemes in discourse through the use of corpora and the development of metalexical knowledge; text synthesis and reuse of the phrasemes discussed in an argumentative essay. These three steps are articulated in different workshops, which are as follows:

Step 1 – Exploration of phrasemes.

1) General comprehension of phrasemes – Number of workshops: 1; Time: 70 minutes. The workshop focuses on the recognition and overall understanding of a number of phrasemes provided by the teacher, using general questions to guide the learners towards their meaning. As the learners' linguistic level allows, we ask them from the very first exercises to arrange the phrasemes by their discursive function, which helps them to grasp their meaning through context. Here are some examples:

Table 1. Classification of discursive functions.

Discursive functions	Example
Exprimer son point de vue	<i>jouer en faveur de, avoir un impact moindre, etc.</i>
Nommer, définir, discuter, commenter une définition d'un terme ou d'un concept	<i>cet article affirme que, l'auteur doit envisager, etc.</i>
Exprimer une temporalité	<i>toucher à sa fin, au cours des dernières années, etc.</i>
Organiser son discours	<i>nous analysons ici, d'autre part, dans le cas de, comme l'illustre, etc.</i>

Step 2 – Metalexical knowledge.

2) Meaning in corpus – Number of workshops: 1; Time: 70 minutes. In a second step, we plan to develop tracking activities through the decomposition of the meaning not of the units of the phraseme, but of the referent to which the entire phraseme refers. Once the students have understood how these lexical structures work, we verify their ability to extract new phrasemes in corpus. On the home page of the *ScienQuest* corpus “EEIDA français”,⁴ students choose the discipline, the modality and the text portions that interest them. The learners thus begin their exploration and analysis on the syntactic, lexico-semantic and discursive levels. Here are some examples:

Table 2. Linguistic analysis of phrasemes proposed by the students.

Introduire une transition	Phraseme	Construction	Discours
	<i>Nous examinons ici</i>	Introducteur d'énoncé	Technique
	<i>D'autre part</i>	Adj+N	Multiregistre
	<i>Nous intéresser à</i>	Pron+V+prép.	Technique
	<i>Dernier point mais pas le moindre</i>	Constr. Adjectivale	Multiregistre
	<i>On peut affirmer que</i>	Constr. Verbale	Universitaire

3) Definition of idiom and collocation in ECL – Number of workshops: 2; Time: 140 minutes (70 minutes for each workshop). After dealing with phrasemes as a whole, this activity aims at distinguishing idioms from collocations according to the rigorous definition offered within the ECL framework. As students who possess some metalexical competencies are presupposed to understand these two notions properly, teachers should verify that students already possess these competencies – as it would be the case for our public of students of Lexicology⁵ – or should introduce them to these skills in a preliminary step. For instance, students should possess the notions of lexical unit, meaning, base, collocate, etc.⁶, in order to understand the notion of collocation.

During the second workshop, students are asked to identify the free lexical combinations, the idioms and the collocations within each discursive function of step 1. For example, as for the discursive function *Exprimer son point de vue*, students must distinguish between the idiom *jouer en faveur de* and the collocations *résultat prometteur*, *constitue une avancée*, *moindre impact*. At the end of this activity, they

⁴ *Corpus Études interdisciplinaires et interlinguistiques du discours académique (EIIIDA français)* <https://corpora.aiakide.net/scientext19/> (Accessed in 13 january 2019).

⁵ Also for students of Lexicology, it could be necessary to revise some lexicological notions or to present some new ones.

⁶ "Si l'on veut faire l'étude de la notion de collocation, d'ailleurs au carrefour de la sémantique et de la syntaxe, on doit donc d'abord maîtriser les notions qui la définissent et la caractérisent : *unité lexicale, syntagme, locuteur, sens, combinatoire restreinte, base [d'une collocation] et collocatif*" (Polguère, Tremblay, 2014, 1184).

will prepare a list of idioms (for example *toucher à sa fin*) and of collocations (*conduire une analyse, travaux antérieurs, tendance actuelle*, etc.).

4) Analysis of idioms and collocations – Number of workshops: 1; Time: 70 minutes. Students are then asked to conduct a syntactic and semantic analysis of a sample of idioms and collocations to justify their choice. As for idioms, they are asked to test the non-compositional meaning and the fixed syntactic structure of the selected items. They are supposed to identify the base and the collocate with respect to collocations. Here are some examples: *résultat prometteur* > *résultat* = base and *prometteur* = collocate; *tendance actuelle* > *tendance* = base and *actuelle* = collocate; *avancer l'hypothèse* > *hypothèse* = base and *avancer* = collocate.

5) Use of lexicography – Number of workshops: 1; Time: 70 minutes. Following this analysis, another workshop is developed around the lexicographic treatment of phrasemes in online dictionaries and resources. Students are asked to observe if the two phenomena are treated as a whole, under the category of phraseology, or whether they are clearly separated. Moreover, they have to verify if idioms constitute separate entries and whether collocations are mentioned under the heading of the base or the collocate. Students also consult two tools, the *Lexique actif du français* (Mel'čuk, Polguère, 2007) and the *DicoPop*, developed in the theoretical framework of ECL; the supervision of teachers is suggested if students are not familiar with ECL.

Step 3 – Text synthesis and reuse of the phrasemes

6) Production of an argumentative text - Number of workshops: 1; Time: 70 minutes. The sequence ends with an activity of assistance in writing an argumentative text using the *Antidote* tool, a grammatical correction and writing assistance software. This software offers a series of questions that accompany the learners during the writing phase of a text, forcing them to write sentences from which a final text will be generated. The principle is based on interactivity between the learner and the software by associating the users' answers to the software's questions with the statements proposed by the tool. More specifically, students are asked to use a selection of idioms and collocations into the argumentative text.

7) Assessment - Number of workshops: 2; Time: 80 minutes (60 minutes for the written production + 20 minutes for the oral examination). After the completion of the didactic sequence, a final assessment will evaluate how much progress students have achieved in relation to the phraseological units. It involves both the evaluation of the written production and an oral examination: as for the final production, the re-use of phrasemes in an argumentative text will be considered. During the oral examination, a particular attention will be paid to the metalexical knowledge acquired by students in relation to the phraseological units discussed during the Lexicology course.

4 Conclusion

The didactic sequence we suggest is based on an experience of phraseology teaching that highlighted the need both of a progressive learning and of a sharp distinction between different types of phrasemes. Nonetheless, it will be fundamental to test it in a class in order to understand if further adjustments are required. Students should develop a more in-depth knowledge of different kinds of phraseological units after acquiring a certain consciousness about phraseological phenomena as a whole. So, the development of another didactic sequence aiming to explore different kinds of idioms and different kinds of collocations (about collocations, see Frassi, 2018) would be appropriate.

The didactic concept behind this sequence is to develop activities that combine several linguistic phenomena; it is not a question of setting up activities that are detached or focused on a specific point (a particular collocation), or the use of specific verbs, but rather several of these phenomena, combining syntax, lexicon and semantics, without forgetting the predominant role of corpora, which allow the context to be set and thus the identification of both the structures and pragmatic functions of each element of the discourse. These few reflections on phrasemes have allowed us to understand that their presence and that of a wide range of linguistic and textual means such as a progressive learning of metalexical notions, corpora and text writing software can contribute to the construction of semantic coherence in the writing of argumentative texts.

References

1. Binon, J., Verlinde, S.: Les collocations : clef de voûte de l'enseignement et de l'apprentissage du vocabulaire d'une langue étrangère. *Romanesque*, 29 (2), 16–24 (2004).
2. Cavalla C., Labre V.: L'enseignement en FLE de la phraséologie du lexique des affects. In A. Tutin & I. Novakova (eds.), *Le lexique des émotions et sa combinatoire lexicale et syntaxique*, Grenoble: Ellug, 297–316 (2009).
3. Cavanagh M.: *Stratégies pour écrire un texte explicatif*, Montréal: Chenelière Education (2010).
4. Cavanagh M., Blain S.: Élaborer une séquence didactique à l'écrit : selon quels principes théoriques ? *Enjeux*, 77, 83–100 (2010).
5. Charmeux, E.: *Enseigner le vocabulaire autrement*. Lyon: Chronique sociale (2014).
6. Frassi, P.: L'enseignement/apprentissage de la collocation entre contraintes grammaticales et contenu sémantique. *Études de linguistique appliquée*, 189, 63–84 (2018).
7. Frassi, P., Tremblay, O.: Il Réseau Lexical du Français: una banca dati per l'apprendimento del lessico francese. *Linguaggio e apprendimento linguistico. Metodi e strumenti tecnologici*. Milano: Studi AItLA 4, 155–172 (2016).
8. Grossmann, F.: Didactique du lexique : état des lieux et nouvelles orientations. *Pratiques*, 149–150, 163–183 (2011).
9. Lewis M. (eds.) *Teaching Collocation: Further Developments in the Lexical Approach*. Hove: Language Teaching Publications (2000).
10. Mel'čuk, I., Polguère, A.: *Le Lexique Actif du Français*. Louvain-la-Neuve: Duculot (2007).
11. Picoche, J.: *Didactique du vocabulaire français*. Paris: Nathan (1993).

12. Polguère, A.: *Lexicologie et sémantique lexicale. Notions fondamentales*. Montréal: Les Presses de l'Université de Montréal (2016).
13. Scott, J., Nagy, W.: Developing word consciousness. *Essential readings on vocabulary instruction*. Newark, DE: International Reading Association, 106–117 (2009).
14. Tremblay, O., Polguère, A.: Une ontologie linguistique au service de la didactique du lexique. *SHS Web of Conferences 8. 4e Congrès Mondial de Linguistique Française*. Berlin: EDP Sciences, 1173-1188, (2014), Homepage, <10.1051/shsconf/20140801383>. <hal-01026114v2>, last accessed 2019/1/20.
15. Tutin, A., Grossmann, F.: Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif. *Revue française de linguistique appliquée*, vol. VII(1), 7–25 (2002).

Procesos de Reconocimiento e Interpretación de Unidades Fraseológicas Metafóricas y Factores Influyentes

Silvia Cataldo

Universidad de Alicante, San Vicente del Raspeig 03690, España
silvia.cataldo90@gmail.com

Resumen. El objetivo de este trabajo es delinear un cuadro global de las modalidades de recepción de unidades fraseológicas metafóricas (*idioms*), presentando los factores que mayormente influyen en su detección e interpretación y en el esfuerzo cognitivo necesario para llevar a cabo estos procesos. Después de definir la metáfora como un mecanismo propio del pensamiento de conceptualizar la realidad de forma convencional o innovadora, que se manifiesta en la lengua a través de formulaciones libres o lexicalizadas, se introducen los *idioms*, unidades fraseológicas con valor figurado por expresar asociaciones entre dominios distintos. Los factores que intervienen en su reconocimiento y comprensión y en el esfuerzo cognitivo que estos suponen, son heterogéneos y guardan relación con una dimensión gramatical, una semántico-conceptual, la frecuencia de uso y el grado de familiaridad, el acceso directo al sentido figurado o la necesidad de analizar literalmente la expresión. Estos últimos dos en particular, influyen en los tiempos de comprensión y, a su vez, están relacionados con los aspectos mencionados anteriormente, además de con la ambigüedad de la expresión (si acepta más interpretaciones), la predictibilidad de las palabras que la componen y el contexto en el que ella aparece. Todos los factores señalados, como resulta claro, no son autónomos, sino que se entremezclan, creando una condición de dependencia recíproca.

Palabras clave: Idioms, Fraseología, Cognición.

1 La Metáfora en la Mente

A lo largo del siglo XX, poco a poco, se deja de concebir la metáfora como un fenómeno lingüístico con función decorativa, privilegiando su carácter cognitivo: a partir de las observaciones de Richards (1936), Black (1962), Lakoff y Johnson (1980), se propone una visión de la metáfora como mecanismo de transferencia de características entre dominios distintos, que puede ser convencional por un lado, si responde a formas de comprender una determinada realidad comunes en varias culturas, en una comunidad dada, en grupos reducidos dentro de ella o de una situación enunciativa en concreto, o incluso utilizadas normalmente por un solo individuo, (*cf.* Kövecses 2005; 2010a; 2010b; 2010c), e innovadora por otro lado.

En el marco de este trabajo, se hará referencia a conceptualizaciones convencionales, sin considerar su grado de difusión intra- o intercultural.

2 La Metáfora en la Lengua: los *Idioms*

Las metáforas en la mente comentadas en el párrafo anterior se pueden manifestar en la lengua como formulaciones libres en caso de conceptualizaciones puntuales y como libres o lexicalizadas en correspondencia de conceptualizaciones convencionales. Estas últimas, a su vez, pueden constar de palabras sueltas, tal y como se observa en las polisemias, o de unidades fraseológicas, las cuales pueden incluso sufrir manipulaciones creativas internas, si se interviene en sus componentes, o externas, por ejemplo, si se insertan en un contexto inusual (*cf.* Corpas 1996: 235-250; *cf.* Langlotz, 2006: 179-182). El presente estudio se concentra en las unidades fraseológicas con valor metafórico (denominadas también *idioms*) que no han sufrido modificaciones.

3 Modalidades de Reconocimiento e Interpretación de *Idioms*

En esta tercera sección se presentan algunos de los factores que influyen en el reconocimiento y en la interpretación de los *idioms* y en el esfuerzo cognitivo que dichos procesos mentales comportan. En primer lugar, se toma en consideración la relación cronológica, no siempre lineal y previsible, entre reconocimiento de la condición lexicalizada de la expresión y de su metafóricidad, e interpretación; en los subpárrafos sucesivos, en cambio, se tratan los aspectos gramaticales, semántico-conceptuales y relativos a la frecuencia de uso relacionados con la detectabilidad y con la correcta comprensión de los *idioms*; por último, se consideran los principales factores involucrados en el esfuerzo cognitivo requerido al oyente o al lector para el procesamiento de unidades fraseológicas metafóricas.

3.1 Relación Cronológica entre Reconocimiento e Interpretación

En lo que atañe a la relación cronológica entre el reconocimiento de una unidad fraseológica y su interpretación, Timofeeva (2012: 409-410) parafrasea, interpreta y completa la visión de Corpas (2003): la autora supone que cuando se conoce el significado de una unidad fraseológica figurada hay simultaneidad, mientras que con expresiones homófonas, que se pueden comprender tanto literal como metafóricamente, la interpretación puede preceder y guiar hacia la identificación; además, admite la posibilidad de que se detecte el fraseologismo a pesar de no conocer su significado y de que la identificación pueda incluso no tener lugar. Aunque Timofeeva delinea su reflexión en una óptica traductológica, su discurso podría adaptarse a cualquier lector, ya que el primer paso de una traducción es la lectura y la comprensión del texto. Por lo visto, resulta clara la incapacidad de establecer *a priori* si la detección se manifiesta y cuándo, tratándose de un proceso

vinculado a factores subjetivos, relativos a la competencia fraseológica y a la sensibilidad lingüística del individuo, contextuales, gramaticales y semántico-conceptuales. Debido a esta imprevisibilidad, en el marco del presente trabajo, solo en ciertos casos se podrá distinguir entre reconocimiento e interpretación, mientras que en la mayoría de las circunstancias tales mecanismos se considerarán en bloque.

3.2 Factores Gramaticales Involucrados en el Reconocimiento de los *Idioms*

Con respecto a la dimensión formal, Goatly (1997: 78-88) considera lo importante que puede ser, en la identificación de una metáfora, la categoría gramatical del dominio utilizado para comprender una determinada realidad: si se trata de un nombre, la posibilidad de que se detecte la metáfora es alta, pero va reduciéndose gradualmente si la transferencia conceptual se activa mediante un verbo, un adjetivo, un adverbio o una preposición. Sería a través de preposiciones que se manifiesta la mayoría de las metáforas en las lenguas, debido a su capacidad de adaptarse a usos temporales y abstractos (Steen, 2011: 52). Aunque la aportación de Goatly parece más adecuada en caso de polisemias, no se excluye que pueda aplicarse también a los *idioms*, por lo menos cuando su significado es parcialmente composicional y su metaforicidad reside solo en una palabra: según su teoría, en *ser el segundo plato de alguien*, por ejemplo, por concentrarse la metáfora en un sintagma nominal a través del cual se alude al hecho de ser una persona la segunda opción para otra, la metaforicidad resulta quizás más fuerte que en las expresiones *caer enfermo*, utilizada para referirse al enfermar alguien, y *no estar muy católico*, con la que se alude a un mal estado de salud, donde el valor figurado del verbo *caer* y del adjetivo *católico*, aunque evidente, podría percibirse con menor vigor. Sin embargo, es importante precisar que la percepción de la metaforicidad en los casos mencionados está vinculada también a otros aspectos que se comentarán más adelante. En lo que se refiere a las demás categorías gramaticales, no es fácil encontrar unidades fraseológicas cuya metaforicidad derive de un adverbio, a no ser que se haga referencia a algunas colocaciones, o de una preposición. Considérese la colocación *profundamente dormido*, donde el carácter figurado del adverbio combinado con *dormir* para describir la intensidad del sueño, presenta un grado de lexicalización muy alto, que en ocasiones no permite detectar conscientemente la metáfora; de forma en parte análoga, la expresión *en un santiamén*, del todo figurada por indicar que algo se lleva a cabo o se realiza en muy poco tiempo, presenta, en su interior, un uso figurado de *en* en el que reside la conceptualización extremadamente común del tiempo en términos de espacio.

En cuanto a la dimensión formal, además, Baker (1992: 65) sugiere que posibles estructuras agramaticales dentro de un *idiom* pueden señalar su condición de unidad fraseológica.

3.3 Factores Semántico-conceptuales Involucrados en el Reconocimiento y en la Interpretación de los *Idioms*

De entre los factores semántico-conceptuales relevantes en la identificación de la condición figurada y en la interpretación de *idioms*, se pueden mencionar: a) el posible valor de falsedad de un enunciado, considerando que para Grice (1975) la metáfora, independientemente de su condición lexicalizada, representa una violación de la máxima de calidad, por querer el hablante decir algo diferente a lo que su enunciado expresa literalmente, contando con la colaboración del interlocutor en darle a sus palabras una interpretación coherente; b) el grado de metaforicidad de la expresión figurada, que a su vez es alto en presencia de asociaciones de dominios semánticamente distantes por pertenecer a campos semánticos diferentes, sobre todo si son uno concreto y uno abstracto, y de una clara contradictoriedad de la imagen (Goatly 1997), que se supone que podría dar lugar a un fuerte conflicto conceptual. Aunque idealmente el conflicto conceptual no debería manifestarse en los *idioms*, que por definición son estables en la lengua y responden a conceptualizaciones ya disponibles, su posible baja frecuencia de uso, o la familiaridad limitada por parte del oyente o del lector, podrían hacer que una expresión lexicalizada se perciba como contradictoria. Hay que tener en cuenta, a este propósito, que es difícil que un individuo conozca todas las unidades fraseológicas de su lengua, lo que puede conllevar que las identifique y las interprete correctamente gracias a la transparencia de la conceptualización, al contexto o a sus propias capacidades de crear relaciones conceptuales, que las identifique por motivos formales y no las sepa interpretar, que las interprete correctamente sin enterarse de su condición lexicalizada por considerarlas innovadoras en la mente y libres en la lengua.

El desconocimiento de un *idiom*, de su condición lexicalizada o de su significado por parte de un individuo, se podría también deber a la presencia de variantes diatópicas o diastráticas: es posible que en una lengua existan unidades fraseológicas metafóricas limitadas territorialmente o propias de ciertos grupos sociales, desconocidas en otras zonas o por personas no pertenecientes a esos grupos. La comprensión en tales circunstancias se parece de alguna forma a la que puede caracterizar la interpretación en la traducción, cuando las transferencias conceptuales disponibles en la lengua fuente y en la lengua meta son diferentes y el traductor se enfrenta con un *idiom* que no existe tal cual en la otra lengua y que no forma parte de su bagaje de conocimientos lingüísticos. Mandelblit (1995), al respecto, observa tiempos de reacción (y supuestamente esfuerzos cognitivos) menores en unidades fraseológicas que reflejan conceptualizaciones compartidas entre las lenguas involucradas en el proceso traductor, sobre todo si son análogas también lexicalmente, frente a las que reflejan conceptualizaciones distintas. Además de diferencias entre dos universos lingüístico-conceptuales, como pueden ser los representados no solo por dos lenguas, sino también por dos variantes de un mismo idioma, Bazzanella (1999: 155-156) sugiere que la falta de coincidencia de las experiencias de autor y lector hace que sus conceptualizaciones prototípicas de un determinado dominio sean distintas, originando dificultades en la interpretación o malentendidos. Esto, sin duda válido para las metáforas originales, puede aplicarse de alguna forma a las unidades

fraseológicas figuradas: es posible, por ejemplo, que los hablantes de una misma lengua, en el uso y en la recepción de un *idiom* versátil, por ser capaz de adaptarse a más contextos, seleccionen preferentemente significados diferentes de entre los que se le suelen atribuir. Considérese la expresión española *torcer la cabeza*, que puede referirse tanto al enfermar como al morir: a no ser que el contexto desambigüe de manera inequívoca el sentido de la unidad fraseológica metonímica, por conceptualizar condiciones físicas mediante un posible gesto que las caracteriza habitualmente, dos individuos podrían propender por dos interpretaciones distintas dependiendo de si asocian con más frecuencia el movimiento de la cabeza a la enfermedad o a la muerte. De forma parecida, la expresión italiana *fare la cicala* (hacer la cigarra), podría aludir a la tendencia, por un lado, a gastar dinero sin pensar en el futuro, haciendo referencia a la fábula de la cigarra y de la hormiga, y por otro, a hablar mucho y de futilidades: también en este caso, sin el soporte del contexto, dos hablantes podrían no entenderse si uno acostumbra a conceptualizar la cigarra como un desperdiciador de riquezas y el otro como un gran hablador.

3.4 Frecuencia de Uso y Grado de Familiaridad

Algunos de los aspectos tratados en el párrafo anterior, además de adscribirse a los factores semántico-conceptuales mencionados, guardan cierta relación también con la frecuencia de uso de los *idioms* y con cuánto el receptor está familiarizado con ellos.

En primer lugar, es importante considerar que las unidades fraseológicas, a pesar de su fuerte estabilidad en la lengua, no presentan todas el mismo grado de difusión y algunas son utilizadas más a menudo que otras: claramente, las probabilidades de reconocimiento y adecuada interpretación serán superiores en las más comunes. Por otra parte, sin embargo, su alta frecuencia de uso podría obstaculizar la identificación de su carácter metafórico y llevar al oyente o al lector a acceder a su significado figurado sin percibir las transferencias conceptuales subyacentes. Se supone que existe un vínculo estricto entre frecuencia de uso del *idiom* y cuánto el individuo está familiarizado con él, aunque es cierto que el grado de familiaridad depende también de factores personales, ya que un hablante podría estar acostumbrado a utilizar preferentemente una unidad fraseológica determinada en general poco difusa y, por consiguiente, tener más posibilidades de reconocerla y comprenderla (o incluso de comprenderla sin percibir su metafóricidad) con respecto a otras personas.

3.5 Esfuerzo Cognitivo

En esta sección se trata el proceso de reconocimiento e interpretación de unidades fraseológicas metafóricas desde la perspectiva del esfuerzo cognitivo que este requiere al receptor. Varios estudiosos citados por Sjørup (2013), concentrándose en la comprensión de textos escritos, consideran que el esfuerzo cognitivo es directamente proporcional a la duración de la fijación ocular en la unidad textual, y por esto en los últimos años se han incrementado las investigaciones de *eye-tracking*, basadas en el seguimiento de los movimientos oculares del traductor en la pantalla. Aunque muchas de ellas se dedican a la observación de los aspectos cognitivos

relacionados con la labor traductora, algunos de sus resultados se pueden aplicar al común lector.

Carpenter, Miyake y Just (1994: 1083), por ejemplo, observan que en correspondencia de expresiones ambiguas, la capacidad de la memoria de trabajo interviene en los tiempos necesarios para su procesamiento, ya que cuanto más reducida es, más se tarda en manejar las múltiples interpretaciones: esta es la situación que puede presentarse en caso de *idioms* que, en el contexto en el que aparecen, pueden adquirir claves de lectura distintas, literales y metafóricas o sólo metafóricas. Resumiendo los resultados de diferentes estudios y refiriéndose a la comprensión de textos en general, Sjørup (2013: 86-92) menciona varios elementos que se ha demostrado que influyen en el movimiento de los ojos, algunos de los cuales podrían considerarse válidos para los *idioms*: (i) ambigüedades léxicas o sintácticas, (ii) la ya mencionada escasa frecuencia de uso o familiaridad por parte del lector, en ambos casos con una duración mayor de las fijaciones oculares o un número más alto de ellas, (iii) predictibilidad de una palabra en un contexto determinado, que si es alta supone un esfuerzo cognitivo menor, con una reducción de los tiempos de fijación ocular, y que claramente dependerá también de cuánto el lector está familiarizado con la expresión.

El esfuerzo cognitivo, además, está vinculado al tipo de interpretación que se activa al empezar el procesamiento del *idiom*, que puede ser literal o metafórica. Según los partidarios de la *standard pragmatic view*, el receptor reconoce la metafóricidad de una expresión después de enterarse de la incoherencia de su significado literal, mientras que los que sostienen la *direct access view*, consideran que es posible acceder directamente al sentido figurado de ciertas expresiones. La primera parece respaldar la idea de Grice (1975) de que para interpretar correctamente una expresión metafórica, el oyente o el lector necesita identificarla como falsa, lo cual supondría tiempos de comprensión más largos; sin embargo, quienes sostienen la validez de la *direct access view*, afirman que en ocasiones se comprende una expresión en su sentido figurado incluso cuando esta podría aceptar una interpretación literal. Según Gibbs (1994: 425, 427), aunque pueda surgir del análisis semántico de los componentes, la interpretación metafórica podría no tener en cuenta el sentido literal de cada palabra: en un experimento, efectivamente, observa que con expresiones que admiten interpretaciones tanto idiomáticas como literales, se suele seleccionar el sentido convencional figurado y pasar a una comprensión literal solo si este no se considera plausible (Gibbs 1980: 155), como si algunos *idioms* estuvieran almacenados en la memoria como unidades léxicas. Sin embargo, el límite entre literalidad y metafóricidad en la interpretación de unidades fraseológicas figuradas parece matizado, ya que Cacciari y Tabossi (1988) y Cacciari (2001: 313-314) suponen la posibilidad de un análisis literal de los *idioms* hasta un punto de reconocimiento a partir del cual son procesados como bloques únicos.

Se ha observado también que la dimensión contextual juega un papel esencial tanto en la selección de una interpretación literal o metafórica, como en el acceso directo a esta en presencia de metáforas ambiguas y, por consiguiente, en el esfuerzo cognitivo por parte del lector: al respecto, Inhoff, Lima y Carroll (1984) han comprobado que un contexto amplio previo a la expresión metafórica permite comprenderla

automáticamente en su sentido figurado, sin requerir tiempos superiores a los empleados en el procesamiento de un uso literal, y que estos se reducen si el contexto que precede la metáfora es metafórico y no literal. Aceptando todas las situaciones descritas por lo que atañe a la relación entre interpretaciones literales y metafóricas, se puede concluir que los tiempos de comprensión aumentan en caso de que se acceda al análisis semántico de los componentes para interpretar metafóricamente la unidad fraseológica, lo cual se traduce en un esfuerzo cognitivo mayor.

4 Conclusiones

Las reflexiones presentadas en este estudio pueden resumirse en los siguientes puntos: 1) la metáfora es un mecanismo de transferencias conceptuales de características entre dos dominios y 2) puede ser convencional o creativa en el pensamiento y realizarse de manera libre o lexicalizada en la lengua; 3) un tipo de metáforas lexicalizadas son los *idioms*, unidades fraseológicas de carácter figurado; 4) al procesar un *idiom*, el receptor podría detectarlo e interpretarlo simultáneamente, detectarlo antes de interpretarlo, interpretarlo antes de detectarlo, o interpretarlo sin reconocerlo como unidad fraseológica. 5) De entre los factores que intervienen en el procesamiento de los *idioms*, los que guardan relación con la detección de su condición metafórica y de su estado lexicalizado son respectivamente 5.1) la categoría gramatical del dominio utilizado para comprender una determinada realidad y 5.2) la composición interna en caso de estructuras gramaticalmente incorrectas, mientras que los que influyen en el reconocimiento de su condición figurada y en la interpretación son 5.3) aspectos semántico-cognitivos, relativos al posible valor de falsedad de una expresión y al grado de metaforicidad, que depende de la distancia semántica entre los dominios asociados y de la contradictoriedad de la imagen, y que está vinculado también al nivel de familiaridad del hablante con la expresión y a si para él esta refleja una conceptualización usual, sobre todo en caso de variantes diatópicas o diastráticas, 5.4) la frecuencia de uso y el grado de familiaridad del *idiom*, que si son altos podrían llevar a una correcta interpretación sin que se perciba la metaforicidad. 6) Estos aspectos intervienen en el esfuerzo cognitivo necesario para el procesamiento de una unidad fraseológica metafórica, cuya unidad de medida suele ser la duración del tiempo de fijación ocular durante la lectura, el cual depende también 6.1) de la ambigüedad de la expresión (si se puede interpretar tanto literal como metafóricamente), 6.2) de la predictibilidad de las palabras que componen el *idiom* y 6.3) de si su procesamiento empieza con una interpretación literal para luego pasar a una lectura metafórica o directamente con una comprensión figurada. En esta última cuestión juega un papel fundamental el contexto, además de factores personales como la competencia fraseológica del hablante (importante también para el reconocimiento de la condición lexicalizada de la unidad textual), su nivel de familiaridad con la expresión y la frecuencia de uso del *idiom*.

De la breve lista presentada resulta evidente que los factores involucrados en la detección y en la interpretación de unidades fraseológicas metafóricas son

heterogéneos y, al mismo tiempo, están estrictamente conectados y entremezclados entre ellos, lo cual hace que sea muy difícil prever la manera en la que el procesamiento de los *idioms* se realiza.

Referencias

1. Baker, M.: In Other Words: A Coursebook on Translation. Routledge, New York (1992).
2. Bazzanella, C.: La metáfora tra mente e discorso: alcuni cenni. *Lingua e stile*, 34(2), 150-158 (1999).
3. Black, M.: Models and metaphors: Studies in language and philosophy. Cornell University Press, Ithaca (1962).
4. Cacciari, C., P. Tabossi: The Comprehension of Idioms. *Journal of Memory and Language* 2, 668-683 (1988).
5. Cacciari, C.: *Psicologia del linguaggio*. Il Mulino, Bologna (2001).
6. Camp, E: Metaphor in the Mind: The Cognition of Metaphor. *Philosophy Compass* 1(2), 154-170 (2006)
7. Carpenter, P. A., A. Miyake, M. A. Just: Working memory constraints in comprehension: Evidence from individual differences, aphasia, and aging. In: Gernsbacher, M. A. (ed.) *Handbook of Psycholinguistics*, pp. 1075-1122. Academic Press, San Diego (1994).
8. Corpas Pastor, G.: *Manual de fraseología española*. Gredos, Madrid (1996).
9. Corpas Pastor, G: Diez años de investigación en fraseología: análisis sintáctico-semánticos, contrastivos y traductológicos. Vervuert, Madrid (2003).
10. Gibbs, R.: The process of understanding literary metaphor. *Journal of Literary Semantics* 19, 65-94 (1990).
11. Gibbs, R.: Figurative Thought and figurative language. In: M. A. Gernsbacher (ed.) *Handbook of Psycholinguistics*, pp. 411-446. Academic Press, San Diego (1994).
12. Gibbs, R.: A new look at literal meaning in understanding what is said and implicated. *Journal of Pragmatics* 34, 457-486 (2002).
13. Gibbs, R.: Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition* 8, 149-156 (1980).
14. Goatly, A.: *The language of metaphors*. Routledge, London (1997).
15. Grice, H. P.: Logic and Conversation. In: Cole, P., J. L. Morgan (eds.) *Syntax and Semantics 3: Speech Acts*, pp. 41-58. Academic Press, New York: (1975).
16. Inhoff, A. W., S. D. Lima, P. J. Carroll: Contextual effects on metaphor comprehension in reading. *Memory and Cognition* 7(6), 558-567 (1984).
17. Kövecses, Z.: *Metaphor in culture: universality and variation*. Cambridge University Press, Cambridge (2005).
18. Kövecses, Z.: A new look at metaphorical creativity in cognitive linguistics. *Cognitive Linguistics* 21(4), 663-697 (2010a).
19. Kövecses, Z.: Metaphor and Culture. *Acta Universitatis Sapientiae, Philologica* 2(2), 197-220 (2010b).
20. Kövecses, Z.: Metaphor, Creativity, and Discourse. *DELTA* 26, 719-738 (2010c).
21. Lakoff, G., M. Johnson: *Metaphors we live by*. The University of Chicago Press, Chicago/London (1980).
22. Langlotz, A.: *Idiomatic Creativity. A cognitive-linguistic model of idiom-representation and idiom-variation in English*. John Benjamins Publishing Company, Amsterdam & Philadelphia (2006).

23. Mandelblit, N.: The cognitive view of metaphor and its implications for translation theory. In: Thelen, M., B. Lewandowska-Tomaszczyk (eds.) *TRANSLATION AND MEANING*, Part 3, pp. 483-495. Hogeschool Maastricht, School of Translation and Interpreting, Maastricht (1995).
24. Richards, I. A.: *The Philosophy of Rhetoric*. Oxford University Press, New York/London (1936).
25. Sjørup, A. C.: Cognitive effort in metaphor translation. An eye-tracking and key-logging study. Copenhagen Business School, Copenhagen (2013).
26. Steen, G.: The Contemporary Theory of Metaphor: Now New and Improved! Review of *Cognitive Linguistics* 9(1), 26-64 (2011).
27. Timofeeva, L.: Sobre la traducción fraseológica. *ELUA* 26, 405-432 (2012).

Processing European Portuguese Verbal Idioms: From the Lexicon-Grammar to a Rule-based Parser*

Ana Galvão^{1,3}[0000-0002-0045-012X], Jorge Baptista^{2,3}[0000-0003-4603-4364], and Nuno Mamede^{1,3}[0000-0001-6033-158X]

¹ Instituto Superior Técnico, Universidade de Lisboa
Av. Rovisco Pais 1, P-1049-001 Lisboa, Portugal,
a.s.galvao@tecnico.ulisboa.pt

² Universidade do Algarve - FCHS, Campus de Gambelas, P-2005-139 Faro, Portugal,
jbaptis@ualg.pt

³ L2F - Spoken Language Laboratory, R. Alves Redol 8, P-1000-029 Lisboa, Portugal,
Nuno.Mamede@tecnico.ulisboa.pt

Abstract. Processing verbal idioms is a challenging task for Natural Language Processing systems because they are syntactically analysable strings, with a well-formed structure, identical to that of distributionally free sentences, but whose meaning is for the most part non-compositional. This paper presents recent advances in processing European Portuguese verbal idioms. From a lexicon-grammar matrix, containing +2,500 verbal idioms and +100 (structural, distributional and transformational) properties, parsing rules are automatically generated, within the framework of a rule-based incremental parser. They are then integrated in STRING, a fully-fledged natural language processing system for Portuguese. The system now identifies not only the idioms' base forms, but also the sentences resulting from some productive and very general transformations (passive, pronominalisation), admitted by some of these idioms. Other improvements include: a newly developed lexicon-grammar *validator*, a new *generation module* for transformations' examples, and a new, more granular, *evaluation* module. An *intrinsic* evaluation achieves an overall recall of 92.5%.

Keywords: Verbal idioms · Frozen sentences · European Portuguese · Lexicon-Grammar · Natural Language Processing.

1 Introduction

Verbal idioms (e.g. *não mexer um palha* lit.: 'do not move a straw', 'be idle or indifferent') are a type of *frozen sentences* where the verb and at least one of its arguments are frozen together. By 'frozen' we mean that strong combinatorial constraints can be observed between the verb and at least one of its arguments. The syntactic properties and the overall meaning of the idiom cannot be derived from properties and the individual meaning of its component elements, when they are used independently. Therefore,

* Research for this paper was partially supported by national funds through Fundação para a Ciência e a Tecnologia (ref. UID/CEC/50021/2019).

this information must be encoded in the lexicon, to be precise, in the *lexicon-grammar* of the language [8,9]. In this theoretical and methodological framework, the meaning unit is not the word but the *elementary sentence*, in this case, the verbal idiom, along with its relevant syntactic-semantic (i.e. distributional, structural and transformational) properties.

Verbal idioms can be construed as a special type of *multiword expressions* [5] and constitute a large set of the lexicon-grammar of many languages [10,11], though their frequency in texts is often very low. Processing verbal idioms is a challenging task for Natural Language Processing (NLP) systems [18] because they are syntactically analysable strings, with a well-formed structure, identical to that of distributionally free sentences, but whose meaning is for the most part non-compositional. Processing multiword expressions, including verbal idioms, is essential to represent the meaning of a text in an adequate way. The low frequency of many verbal idioms in corpora makes spotting them a difficult task [13], and much prior has been dedicated to identifying them in texts [14,15,16]. However, the focus of this paper will not be on *identification* (in a lexicographic perspective), but rather on the *processing* of an *already built* computational lexicon (a lexicon-grammar) of verbal idioms [2,3], particularly of transformationally-derived, equivalent sentence-forms (for lack of space, this lexicon-grammar will not be presented here). In fact, little relevance has been given to *transformations* of verbal idioms, that is, sentences that are derived from their base form by general formal changes such as *passive* or *pronominalisation* of free arguments.

This is the major contribution of this paper. It presents the improvements introduced in processing European Portuguese verbal idioms, within the development of a Portuguese NLP system, STRING [12]⁴, allowing it now to identify the most common transformations of verbal idioms. These improvements include: (i) a newly developed *validator* of the lexicon-grammar matrix, to help linguists represent in a consistent and systematic way the verbal idioms in this computational lexicon; (ii) a new *example generation* module, that produces natural language examples for the transformations allowed by each verbal idiom; (iii) a new *rule generation* module, automatically producing the parsing rules to identify both base sentences and their transformations in texts; and (iv) a new, more granular and expedite, *evaluation* module, for an intrinsic assessment of the system's performance.

2 Validator

Since the encoding of linguistic information is a manual procedure, which is a very time-consuming and error-prone task, a *Validator* was built, written in Perl, to check the formal consistency of the matrix. The validator takes as input the CSV-converted lexicon-grammar matrix and performs the following checks, outputting the corresponding error messages: (i) *cell content validation*: checks if the content of the cell in a given column is consistent with the predefined values for that column; (ii) *class consistency cross-validation*: depending on the class of the idiom, the number of relevant columns can vary; the system checks the consistency of the properties with the overall class definition; (iii) *related properties cross-validation*: consistency among related properties,

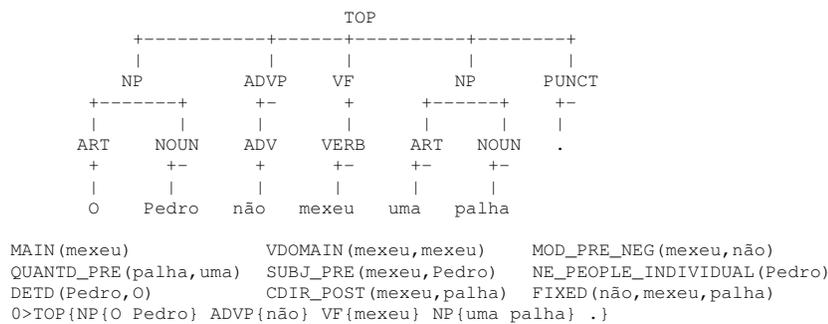
⁴ <https://string.l2f.inesc-id.pt/>

represented in different columns is cross-checked. Several dozens of these rules were manually crafted. Based on the error messages outputted by the validator, it is possible to detect most input errors, which are then manually corrected.

3 Rules Generation Process

In this Section, a brief, overall view of the STRING system [12] is presented first, in order to better frame the task of syntactic analysis (parsing) and the identification of verbal idioms. The STRING system performs all basic operations in a pipeline of modules, from tokenization, text segmentation (sentence splitting) and part-of-speech (PoS) tagging [7,20], and rule-based and statistical PoS disambiguation [6,7]. The syntactic analysis is carried out by *XIP*, the Xerox Incremental Parser [1], using a rule-based grammar specifically built to process Portuguese. It is in this module that verbal idioms are processed. The parser processes texts sentence by sentence. Firstly, it groups words into elementary syntactic phrases or *chunks*, such as noun phrases (NP); this *shallow parsing* operation is called *chunking*. Then, the system extracts syntactic dependencies between the chunks' heads, e.g. the SUBJ (subject); a CDIR (direct complement) dependency between a NP's head and a verb; and an "umbrella" dependency MOD (modifier) linking the nominal head of a PP to a verb (or an adjective to a noun, or an adverb to verb).

This is the output of a sentence with the verbal idiom *não mexer um palha* 'be idle or indifferent'. Each word is associated to a part-of-speech node and then grouped into chunks: NP (noun phrases, twice), ADVP (adverbial phrase) and VF (finite verb phrase). Below the chunk tree, the list of extracted dependencies is presented.



Since the syntactic structure of verbal idioms is by and large identical to that of ordinary, distributionally free verbs; and since even very general transformations, such as the passive or the pronominalization, can be allowed in some cases; we adopted the strategy, as presented in [3,4,17], to first allow the parser to perform a general-purpose analysis of the sentences, as shown above; and, then, use the result of this parse to capture the verbal idiom lexical-syntactic pattern. This is represented by another dependency, FIXED. The (fully expanded) rule for this idiom is the following:

```

if ( VDOMAIN( #?, #2[lemma:mexer] ) & MOD[neg,pre] (#2, #3) & CDIR[post] (#2, #4[surface:palha] )
    & QUANTD (#4, ?[surface:uma] )
    FIXED (#3, #2, #4)

```

The rule first matches a main verb (VDOMAIN) *mexer* ‘move’ with a negation (*neg*) modifier and a direct complement (CDIR) *palha* ‘straw’ and then produces the FIXED dependency. A rule is automatically generated for each verbal idiom in the matrix based solely in the information encoded therein. To do so, each relevant column value contributes with a condition to the rule (relevant columns depend on the verbal idiom class); and each main constituent’s head is associated to a variable; conditions are concatenated by operator ‘&’.

4 Processing Transformations

This is one of the important developments introduced in this paper. In case any transformations apply, the `if()` structure is branched in alternative sets of conditions (noted ‘||’). For example, for the sentence: *O jornalista bombardeou a atriz com perguntas inconvenientes* ‘The reporter bombarded the actress with impertinent questions’, the reduction of the free direct complement to an accusative pronoun (PRON_A): *O jornalista bombardeou-a com perguntas inconvenientes* ‘The reporter bombarded her with impertinent questions’, is expressed by the rule:

```
if ( VDOMAIN(#?, #2[lemma:bombardear]) &
    ( CDIR[post](#2, #3[UMB-Human]) || CLITIC(#2, ?[acc]) ) &
    MOD[post](#2, #4[surface:perguntas]) & PREPD(#4, ?[surface:com]) )
    FIXED(#2, #4)
```

More precisely, the CDIR can alternate (‘||’) with a CLITIC dependency between the verb and a personal pronoun in the accusative case [*acc*] (irrespective of the pronoun being before or after the verb). A similar procedure was used for the dative (PRON_D) and the reflex (PRON_R) pronominalisations. In the case of the passive transformation, *A atriz foi bombardeada com perguntas inconvenientes pelo jornalista* ‘The actress was bombarded with inconvenient questions by the reporter’, the corresponding new rule is:

```
if ( VDOMAIN(#?, #2[pass-ser, lemma:bombardear]) & SUBJ(#2, ?[UMB-Human]) &
    MOD[post](#2, #3[surface:perguntas]) & PREPD(#3, ?[surface:com]) )
    FIXED(#2, #3)
```

The VDOMAIN condition relies on the previous identification of the passive construction pattern with auxiliary verb *ser* ‘be’. A conversion table is used to associate the passive sentence constituents to the correspondent variables in the new rule.

A *configuration file* makes it possible to determine which restrictions are to be applied to the rule generation process (unless otherwise determined, fully expanded rules are produced, as shown above). The controllable restrictions apply to determiners, prepositions and both left and right modifiers of the frozen head noun; and to the distributional constraints to any of the free constituents, both the subject and/or the complements.

5 Generation of Transformation-derived Examples

The example generation process starts by verifying for each idiom whether its description in the lexicon-grammar matrix allows for any transformation and if so, the system

reads the content of each constituent. A conversion table is then used to associate the idiom's constituents to the transformed sentence constituents, which will correspond to a given set of variables in the (new) rule. For example, the direct complement (variable #3) of a verb (#2) becomes the subject (#1) of a passive sentence (see the passive rule above). About 1,170 transformationally-derived sentences were generated for a set of 7 transformations (4 types of pronominalization, the dative restructuring and 2 types of passive constructions), and they were manually revised by a linguist. This was done in several iterations, until satisfactory results were achieved.

6 Evaluation

A newly-built *Evaluation module* was used for an *intrinsic* evaluation of the system's performance. The evaluation uses the lexicon-grammar manually produced examples and the automatically generated, transformation-derived examples, along with the expected output for each sentence, i.e. the `FIXED` dependency with all its arguments. These examples constitute our evaluation corpus. It must be stressed that the examples were produced not taking into account any of the processing steps prior to parsing so that many types of errors may accumulate along this processing pipeline.

The new evaluation module improved the granularity of the system's evaluation concerning verbal idioms, as it considers now 3 criteria for the successful extraction of the idioms' dependency: (i) the extraction of the `FIXED` dependency; (ii) the production of the correct number of arguments for the dependency; (iii) the correct identification of the lexical elements for each argument of the dependency. In case no `FIXED` dependency is extracted, the system returns 'FAILED'. Notice that the previous evaluation module [3] only returned whether the `FIXED` dependency was extracted or not. Besides, the new module processes all sentences in a single batch, so it takes much less time (6 min.) to obtain the results, contributing to a more efficient development of the lexicon-grammar.

Table 1 shows the results for the manually produced sentences per verbal idioms' class. Overall, recall varies from 86.8% for the more relaxed criterion of just capturing the `FIXED` dependency; to 85.3, when the number or the dependency's arguments (`NB-ARG`) is considered; down to 78.2% for a complete match of the dependency's arguments (`ARG`). It is noteworthy to mention that, though the criteria are progressively more strict, the 8.6% drop in the system's performance is relatively small. Next, Table 2 shows the results for the automatically generated, transformation-derived examples, per transformation.

Most errors were found to be due to previous steps of the processing pipeline. For example, the incorrect disambiguation of *a*, either as the definite article 'the' (fem. sg.) or as the preposition *a* 'to'; or the incorrect attribution of lemma to ambiguous verb forms (*foi: ser/ir* 'be/go'). Compound (or multiword) lexical units also hinder the process, as the system gives precedence to them (e.g. *por conta de* 'because' vs. *A Ana vive por conta de_o Pedro* 'Ana depends on Pedro for a living'). Few errors were found due to parsing. For example, the chunking rules failed to produce a `PP` for the sequence *entre nós* 'among us' in the idiom *O Pedro já não está entre nós*, lit. 'Pedro is no longer among us', 'Pedro has died'. Many of these errors have been corrected since, either by

Table 1. Results for verbal idioms’ identification: manually produced sentences.

Class	Total	#FIXED	%	#NB-ARG	%	#ARG	%
CADV	16	7	43.8	7	43.8	5	31.3
C0	21	15	71.4	15	71.4	12	57.1
C1	503	484	96.2	481	95.6	448	89.1
CAN	182	156	85.7	156	85.7	153	84.1
CDN	46	37	80.4	37	80.4	36	78.3
C1P2	291	274	94.2	266	91.4	228	78.4
C1PN	259	224	86.5	216	83.4	206	79.5
CNP2	176	152	86.4	151	85.8	149	84.7
CP1	718	635	88.4	628	87.5	558	77.7
CPN	106	74	69.8	71	67.0	63	59.4
CPP	195	130	66.7	126	64.6	115	59.0
CPPN	36	28	77.8	27	75.0	26	72.2
CV	12	6	50.0	4	33.3	4	33.3
TOTAL	2,561	2,222	86.8	2,185	85.3	2,003	78.2

Table 2. Results for verbal idioms’ identification: Automatically generated, transformation-derived sentences.

Transformation	Total	#FIXED	%	#NB-ARG	%	#ARG	%
PronA	187	170	90.9	169	90.4	165	88.2
PronD	178	131	73.6	130	73.0	129	72.5
PronPos	324	268	82.7	266	82.1	265	81.8
Rdat	192	107	55.7	106	55.2	106	55.2
PassSer	185	142	76.8	141	76.2	139	75.1
PassEstar	83	70	84.3	69	83.1	68	81.9
Total	1,170	909	77.7	902	77.1	884	75.6

improving the system’s general parsing rules, e.g. the PP *entre nós* ‘among us’ chunking rule; or by manually producing alternative rules to the rule generation module in order to take into account multiword lexical units, e.g. *viver por conta de* ‘depend on’ (this often entailed changing the verbal idiom’s class); or simply by using a non-ambiguous verb form in the examples. Naturally, after this process, the overall results, after a 2nd run evaluation, improved significantly, as shown in Table 3 (the system current status).

7 Conclusion and Future Work

This paper expanded previous work on the automatic processing of European Portuguese verbal idioms. The lexicon-grammar of verbal idioms, a matrix with the linguistic description +2,500 frozen sentences, represented by +100 distributional, structural and transformational properties, is automatically converted into the corresponding parsing rules, so that the system may identify these idiomatic expressions in texts.

Table 3. Results for the 2nd run evaluation.

Sentences	Count	FIXED	%	NB-ARG	%	ARG	%
base	2,542	2,429	0.956	2,400	0.944	2,337	0.919
transformed	1,157	1,088	0.940	1,083	0.936	1,083	0.936
Total	3,699	3,517	0.951	3,483	0.942	3,420	0.925

Each idiom is provided with a manually produced example, to illustrate that construction’s base form. The paper introduced several new developments, foremost producing rules that capture the expressions derived from the base forms of those idioms by application of very general transformations (pronominalization, dative restructuring and passive constructions). The system also generates, for these transformationally derived sentences, natural language examples (1,157), which can then be used to test the system’s performance. An *intrinsic* evaluation was carried out, and has shown very positive results: for the strictest criterion, an overall recall of 78.2% and 75.6% for the manually produced and for the automatically generated sentences, respectively. Most errors are due to the previous modules of the processing pipeline, particularly the PoS tagger and the statistical and rule-based disambiguator. Manual correction of these errors made it possible to achieve, in the strictest criterion, a recall of 91.9% and 93.6% for each type of examples, and an overall performance of 92.5%. In the near future, we intend to integrate the corresponding lexicon-grammar of Brazilian Portuguese [19] and perform an *extrinsic* evaluation using (or adapting) the Portuguese corpus developed in-house [3] and that built for the PARSEME project ⁵ [15,16].

References

1. Ait-Mokhtar, S., Chanod, J., Roux, C.: Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering* **8**(2/3), 121–144 (2002)
2. Baptista, J., Correia, A., Fernandes, G.: Frozen Sentences of Portuguese: Formal Descriptions for NLP. In: *Workshop on Multiword Expressions: Integrating Processing* (in EACL 2004). pp. 72–79 (2004)
3. Baptista, J., Fernandes, G., Talhadas, R., Dias, F., Mamede, N.: Implementing European Portuguese Verbal Idioms in a Natural Language Processing System. In: *Corpas Pastor, G. (ed.) Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives* (Proceedings of EUROPHRAS 2015). pp. 102–115 (2016)
4. Baptista, J., Mamede, N., Markov, I.: Integrating verbal idioms into an NLP system. In: *Computational Processing of the Portuguese Language (PROPOR 2014)*. LNAI/LNCS, vol. 8775, pp. 251–256. Springer (2014)
5. Constant, M., Eryigit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Multiword expression processing: A survey. *Computational Linguistics* **43**(4), 837–892 (2017)
6. Diniz, C.: *RuDriCo2 - Um Conversor Baseado em Regras de Transformação Declarativas*. Master’s thesis, Universidade Técnica Lisboa - IST (2010)

⁵ <https://typo.uni-konstanz.de/parseme>

7. Diniz, C., Mamede, N., Pereira, J.D.: RuDriCo2 - a faster disambiguator and segmentation modifier. In: Simpósio de Informática - INForum. pp. 573–584. Universidade do Minho, Portugal (2010)
8. Gross, M.: Une classification des phrases «figées» du français. *Revue Québécoise de Linguistique* **11-2**, 151–185 (1982)
9. Gross, M.: Lexicon-grammar. In: Brown, K., Miller, J. (eds.) *Concise Encyclopedia of Syntactic Theories*, pp. 244–259. Pergamon, Cambridge (1996)
10. Lamiroy, B.: Le lexique-grammaire: essai de synthèse. *Travaux de Linguistique* **37**, 7–23 (1998)
11. Lamiroy, B. (ed.): *Les expressions verbales figées de la francophonie: Belgique, France, Québec et Suisse*. OPHRYS, Paris (2010)
12. Mamede, N., Baptista, J., Diniz, C., Cabarrão, V.: STRING - A Hybrid Statistical and Rule-based Natural Language Processing Chain for Portuguese. In: Abad, A. (ed.) *International Conference on Computational Processing of Portuguese (PROPOR 2012) - Demo Session (2012)*, <http://www.inesc-id.pt/ficheiros/publicacoes/8578.pdf>
13. Manning, Chris; Schütze, H.: *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1st edn. (May 1999)
14. Pecina, P.: Lexical association measures and collocation extraction. *Language Resources and Evaluation* **44**, 137–158 (2010)
15. Ramisch, C., et al.: Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. pp. 222–240. ACL (2018), <https://www.aclweb.org/anthology/W18-4925>
16. Ramisch, C., Ramisch, R., Zilio, L., Villavicencio, A., Cordeiro, S.: A Corpus Study of Verbal Multiword Expressions in Brazilian Portuguese. In: *Computational Processing of the Portuguese Language (PROPOR 2018)*. LNAI/LNCS, vol. 11122, pp. 24–34 (2018)
17. Rassi, A., Santos-Turati, C., Baptista, J., Mamede, N., Vale, O.: The fuzzy boundaries of operator verb and support verb constructions with *dar* “give” and *ter* “have” in Brazilian Portuguese. In: *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP2014 in COLING 2014)*. pp. 92–101. ACL (2014)
18. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: *Proceedings of Computational Linguistics and Intelligent Text Processing*. LNAI/LNCS, vol. 2276, pp. 1–15. 3rd International Conference CILing-2002, Springer, Berlin (2002)
19. Vale, O.A.: *Expressões Cristalizadas do Português do Brasil: uma proposta de tipologia*. Tese de Doutorado, Universidade Estadual Paulista, Araraquara (2001)
20. Vicente, A.: *LexMan: um Segmentador e Analisador Morfológico com Transdutores*. Master’s thesis, Instituto Superior Técnico, Universidade de Lisboa, L²F/INESC-ID, Lisboa, Portugal (2013)

Constructive Linguistics for Computational Phraseology: the Esperanto Case

Federico Gobbo^{1,2}[0000-0003-1748-4921]

¹ University of Amsterdam, the Netherlands

² University of Turin, Italy

F.Gobbo@uva.nl

<http://uva.nl/profile/f.gobbo/>

Abstract. This paper presents the application of the constructive adpositional grammars (CxAdGrams) to phraseological units, through the special case study of Esperanto. Constructive linguistics is an approach to human language analysis that considers constructions, themselves being paradigms of language-in-use, as the first units. Unlike other constructional approaches, constructive linguists apply formalisms in understanding linguistic phenomena. The adpositional paradigm is the most developed formalism in constructive linguistics, which is understandable by humans and machine-readable at the same time. The term ‘constructive’ should also be understood in formal terms, as the adpositional paradigm is based on constructive mathematics, and in particular on topos-theory. From a theoretical perspective, CxAdGrams describe human languages in terms of constructions, described adpositional trees (in short, adtrees). This paper aims to explain why such an interpretation of constructions in terms of adtrees can be useful for a deeper understanding of phraseology. Esperanto is the case study chosen so to give an empirical base to CxAdGrams. In particular, we illustrate the problematisation of its phraseology as well as the advantages of Esperanto in setting up workable prototypes in a short time.

Keywords: Constructive Linguistics · Computational Phraseology · Adpositional Grammars · Adpositional Argumentation · Esperanto.

1 Introduction

Understanding how language is structured is one of the most fascinating and challenging endeavour that human beings have ever done. Many approaches are possible, and many approaches were proposed throughout the flow of human history. Because of the computational turn, in the 21st century our conceptualisation of language is changed and is still changing, and, for this reason, linguistics should propose robust theories that treat language in terms of information, to be understood by humans and read by machines at the same time.

Initially proposed in [14], constructive linguistics is a relatively new approach to human language that follows such informational tenet. It is important to note that, in this perspective, the word ‘constructive’ has both a mathematical and

a linguistic specific meaning at the same time. In short, on the one hand, constructive mathematics is a way to develop mathematics that strictly preserves the information content of any statement [2]. On the other hand, cognitive sciences show that humans are able to communicate as they can read intentions, i.e., infer what the listener is expecting from the speaker, and find patterns, i.e., they can categorise sensibilia mapping them into the mind; they learn intention-reading and pattern-finding by imitation of other humans [21]. In fact, humans use language to gain somebody's attention or to share their mental state. In order to do so, a human language can be described in terms of a map of social conventions of a specific speech community. This individual and collective process of categorisation leads to the emergence of linguistic *constructions*, which are patterns of form-meaning correspondence based on language usage. For this reason, human languages can be described as collections of constructions. We consider constructions as the hypernym of phraseological units and other linguistic phenomena. In general, (oral) discourses and (written) texts are split into units – such as sentences and phrases – which are instantiations of linguistic constructions; phraseological units are a specific type of such units. For the purposes of this paper, we will delve into phraseological units only.

Phraseological units are at the crossroad of grammaticalisation and lexicalisation, which are two complementary processes that can be found in any living human language [3]. While grammaticalisation is a syntactotelic process, lexicalisation is a synthetic process. In other words, grammaticalisation goes from the lexicon to the syntax, affecting lexical items both in their phonological material and in their meaning (which tends to be lost). Conversely, the process of lexicalisation makes constructions lose their flexibility and compositionality, and eventually, they acquire idiosyncratic content. The most extreme result of lexicalisation is the formation of idiomatic expressions, which are not analysable anymore, but should be taken as fixed. Therefore, under the perspective of constructive linguistics, phraseological units are in the middle of the continuum of constructions, where at one extreme we find idioms while at the other one we find word-playing, portmanteaus, dynamic metaphors, and, in general, creative language usage.

Let us show an example of a phraseological unit found in the middle of the continuum of constructions. The following quotation is from the political pamphlet *Gli Stati Uniti d'Europa* [United States of Europe] [16], written by one of the founding fathers of the European Union, the Italian antifascist intellectual Ernesto Rossi (author's English translation immediately below):

Clemenceau diceva che la guerra è una cosa troppo seria per essere lasciata ai generali. Noi dobbiamo dire che la pace è una cosa troppo seria per essere lasciata ai diplomatici. [16, p. 96]

[Clemenceau used to say that war is too serious a matter to be left to the generals. We should say that peace is too a serious matter to be left to the diplomats.]

The quotation shows the same construction – in this particular case, a phrase schema – in two different instantiations. The fixed part is *... say that ... is too serious matter to be left to ...* while the analysable parts for the respective sentences are the triples {Clemencau, war, generals} and {we, peace, diplomats}. In the next section, we illustrate how such a phraseological unit can be expressed in terms of adpositional trees.

2 Adpositional trees for phraseological units

Unlike purely constructionist approaches to language such as Radical Construction Grammar [4], constructive linguists do not avoid formalisms, instead they embrace them in constructive mathematical terms. The most developed paradigm in constructive linguistics is based on the concept of *adposition*. In this context, the term has to be intended in two different ways at the same time.

The first way to intend adpositions is linguistic. However, it should be underlined that, here, an adposition is not only a mere hypernym of prepositions, postpositions, and the like, but also and mainly a generalisation of functional words that connect lexemes and other semantically loaded material. The second way to intend adpositions is mathematical. Adpositions represent purely structural information, and they are placed as hooks under the upmost root that sustain the trees that represent constructions.

So far, the adpositional paradigm has been applied in various branches of human languages, generating constructive adpositional grammars (CxAdGrams), which are abstract and general and language-dependent at the same time. In the field of morphology and syntax, CxAdGrams were applied to purely constructional analysis, adapting the Tesnerian notion of valency, actant and grammar character to the key notion of adposition.³ In the field of semantics and pragmatics, CxAdGrams were applied to discourse analysis of therapeutic conversations, through the representation of Searle’s speech act theory of social world construction [17] in terms of pragmatic adtrees [14]. Up to the author’s knowledge, there is still no application of CxAdGrams in the field of phraseology.

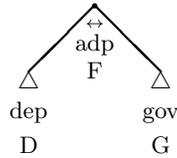


Fig. 1. The generic abstract adpositional tree in its standard form

³ Strictly speaking, the adpositional paradigm does not generate dependency grammars, although there is a relation of ancestry between Tesnière’s Structural Syntax and CxAdGrams [15].

Figure 1 shows the minimal, abstract adtree, in its standard form. Adpositions (adp) represent the relation between two linguistic elements. Linguistic relations are asymmetrical, and they are understood in terms of dependency (dep, conventionally on the left) on a governor (gov, on the right). Each element is tagged in terms of grammar characters: D and G respectively for dependants and governors, while F stands for ‘final’, as it is the result of their structural relation. Adtrees are recursive; the triangles \triangle on the leaves are a convenient way to represent subtrees without indulging in details. Finally, adpositions convey information prominence between dependants and governors – in the form of the arrow \leftrightarrow on top of the hook.⁴ It is worth noting, that every adtree can be flattened, for instance for the purpose of coding, through trivial finite-state automata that do the linearisation. Figure 2 (immediately below) shows the generic abstract adtree in its linearised form.

$$\text{adp}_F^{\leftrightarrow}((\text{dep})_D, (\text{gov})_G)$$

Fig. 2. The generic abstract adpositional tree in its linearised form

Let us see the phrase schema of Ernesto Rossi’s example, previously stated, in terms of adtrees. For sparing space, Figure 3 represents only the instantiation by the triple {Clemencau, war, generals}.⁵ Epsilons (ϵ) represent syntactic relations, i.e., where no morpheme is found. The right arrow \rightarrow above *Clemencau* indicates that the information prominence is above the dependent instead of the governor in that particular subtree. The usefulness of triangles which hide non-relevant information for the analyst – but always retrievable, thanks to the constructive mathematical foundation – is immediate to the reader. For instance, in Figure 3 the morphological information of the word *generals* as well as the linguistic details of the verbal forms *is too serious matter to be left to* and *used to say* are of no interest as the purpose of this adtree is to put in evidence the phrase schema underlying this phraseological unit, i.e., in linguistic terms, what is grammaticalized, and hence fixed – the skeleton of the phrase schema – and what is conveying the lexical information, that is the triple {Clemencau, war, generals}.

The abstract grammar characters {D, G, F} shown in Figures 1-2 are instantiated as verbants, nominals, adjuncts, circumstantials, respectively {I, O, A,

⁴ For more details on the constructive mathematical aspects of adpositional grammar, readers are invited to check Appendix B of the book presenting the mathematical foundation of CxAdGrams in terms of Grothendiek’s toposes [14].

⁵ It is worth noting that punctuation is included in CxAdGrams, being themselves adpositions between sentences. In such a way, potentially, a whole large text – like Dante’s *Divina Commedia* – can be represented as a single, enormous adtree.

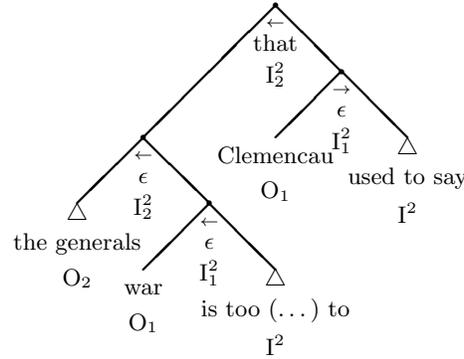


Fig. 3. The standard address of the example *Clemenceau...*

E}.⁶ In principle, any consistent part-of-speech tagging convention can be used with addresses; it suffices to put the tags on the bottom of the leaves (such as O_1 under *Clemenceau* and *war* in Figure 3) and of the hook (such as I_2^2 under *that*).⁷

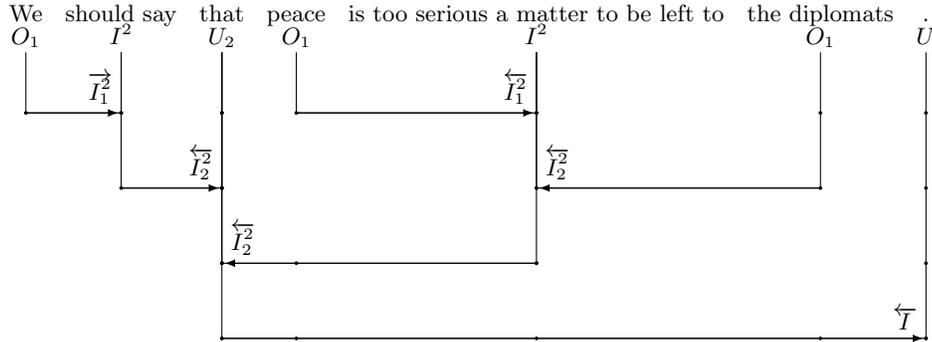


Fig. 4. The path-like address of the example *We should say...*

We present here a third way to represent addresses, which we propose here to call “path-like” addresses, in the absence of a better naming. This representation has the important advantage of respecting the linguistic word order, and therefore it can be useful for educational purposes, as shown in the Linguistic Atelier in Montessori primary schools in Milan, Italy [13]. Structurally speaking, this representation preserves the information, and thus it respects the fundamen-

⁶ Such labelling of grammar characters is borrowed from the original Tesnière’s Structural Syntax [19]. Unfortunately, the letters, which Tesnière took from Esperanto, were not kept in the English translation [20].

⁷ Readers interested in delving into this particular convention can refer to [15, 14].

tal tenet of constructive linguistics, i.e., the foundation on constructive mathematics. In the example below, the fifth grammar character is needed, in order to indicate underspecified or unifying elements (hence, the vowel U), typically grammaticalised morphs or punctuation elements. Figure 4 shows the phrase schema presented before instantiated with the triple {we, peace, diplomats}. Of course, path-like adtrees do not show syntactic relations in terms of epsilons (ϵ) under the hook because they are driven by concrete linguistic material, i.e., by morphs.

3 Phraseology and Esperanto

Esperanto is the most interesting product of Interlinguistics, the branch of linguistics dealing with planned languages, i.e., languages that are written in their fundamental structural traits before even to be spoken [10]. Unlike all other planned languages proposed in the last two centuries, Esperanto succeeded in forming a stable community of language users, with a relevant critical mass, and so it shows emerging sociolinguistic traits that are a challenge for theoretical linguistics.⁸ According to the corpus-based grammar of Esperanto by Gledhill [7], its high morphological regularity, especially in derivation, permits to drastically reduce the learning efforts, both for humans and for machines, even if this does not imply that in absolute terms Esperanto is simpler than other human languages [11]. While presenting its phraseology, Gledhill [7] notes that “many Esperantists are uncomfortable with the idea of variation and near-synonymy in the vocabulary of the language, but as Janton (1994) has pointed out[,] multiple vocabularies are an integral part of Esperanto’s system of register and style.” This may be a reason why in Esperanto studies phraseology is relatively an understudied aspect.

In order to reinforce its language project, Ludwik Lejzer Zamenhof in 1910 published *Proverbaro Esperanta*, a collection of Esperanto proverbs extracted from a comparative analysis of four major European languages (French, German, Polish, Russian) made by his father Mordechai Mark. That book can be considered the base of Esperanto phraseology. Because of the language ideology of ethnic neutrality surrounding Esperanto – which is rather complex [12] – some translations were not straightforward. Let us show one tricky example. Entry number 7 in the *Proverbaro* corresponds to the English phraseological unit ‘it’s Greek to me’, which is construed around the idea of ‘language of Otherness’. In particular, it contains four different proposals for expressing such phraseological unit in Esperanto.

- 7a (539 too [sic]). Ĝi estas por mi ĥina scienco.
- 7b. Ĝi estas por mi volapukaĵo.
- 7c. Nun finiĝas mia klereco.
- 7d. Venis fino al mia latino.

⁸ For a discussion on the possible definition of such peculiar community of language practice, see at least [18].

If we take ethnic neutrality as the standpoint, the problem becomes obvious: you cannot blame Greeks for their “strange” language (following English), and analogously you don’t blame Chinese (following the Spanish *me suena a chino*) or Arabic (following the Italian *per me è arabo*), because Esperanto speakers can be English, Greeks, Chinese, Arabs, Spanish and Italians alike. For this reason, proposal 7a, which refers to Chinese (i.e., *ĥina*) was discarded in practice. Proposal 7c literally means “now it arrived to the end of my knowledge”, lowering too much the pragmatic force found in the phrase schema, because it does not involve a language of Otherness, and therefore it did not work either. Proposal 7d mentions Latin, but Latin cannot always play the role of Otherness: for example, the Dutch expression *Ik ben aan het eind van mijn latijn*, which is very similar to proposal 7c, both meaning literally “there is an end to my Latin”, more or less, in Dutch is used to convey the information ‘I have no energy anymore’, which is completely different from a pragmatic point of view. For this reason, it survives only in the most prestigious register of Esperanto intellectuals. Proposal 7b actually won, as the blamed language of Otherness is Volapük, a language project planned before Esperanto which gained some success in the early days of Esperanto, but it did not work so well. Eventually Volapük entered the Esperanto culture as the language of Otherness – on Volapük, see at least [6].

In Esperanto, phraseological units are the result of the negotiation of meaning between speakers immersed most of their lives in other language environments (there are no Esperanto monolinguals). Zamenhof’s proposal 7c shows that endogenous phraseological solutions are possible. In other terms, there are phraseological units referring specifically to the history, habits, ways of life of Esperanto speakers – as shown by Fiedler in her fundamental work [5]. On the other hand, many of the phraseological expressions found in colloquial Esperanto language use come from europeanisms, i.e., units that are commonly represented in most European languages. In his study on metaphors in Esperanto, Astori [1] shows the proposal by Hungarian Esperanto speakers proposed to introduce *dormi kiel lakto*, literally ‘to sleep like milk’ for the europeanism ‘to sleep like a baby’ (i.e., profoundly), did not work, being too specifically linked to the Hungarian *Weltanschauung*. Conversely, the europeanism, *dormi kiel ŝtono*, literally, ‘like a stone’ is of common use as an alternative expression in the colloquial Esperanto register.

4 A final remark

This position paper shows that CxAdGrams are apt to represent phraseological units, and that a first testbed for a consistent linguistic analysis could be done through Esperanto. The recent proposal of Adpositional Argumentation could analyse Ernesto Rossi’s example as an argument from comparison framed into the Period Table of Arguments, with a considered added-value to the annotated corpus to be done [8, 9, 22].

References

1. Astori, D.: Metafore nell'esperanto. In: Astori, D. (Ed.) *La metafora e la sua traduzione*, pp. 133–148. Bottega del libro, Parma (2016)
2. Bridges, D., Richman, F.: *Varieties of Constructive Mathematics*. Cambridge University Press, Cambridge (1987)
3. Cabrera Moreno, J.C.: On the Relationships Between Grammaticalization and Lexicalization. In: Giacalone Ramat, A. and Hopper, P.J. (Eds.), *The limits of grammaticalization*. pp. 211–229. John Benjamins, Amsterdam (1998)
4. Croft, W.: *Radical Construction Grammar. Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford (2001)
5. Fiedler, S.: *Plansprache und Phraseologie Empirische: Untersuchungen zu reproduziertem Sprachmaterial im Esperanto*. Peter Lang, Bern (1999)
6. Garvía, R.: *Esperanto and its rivals: the struggle for an international language*. Penn Press, Chicago (2015)
7. Gledhill, C.: *The Grammar of Esperanto. A Corpus-based description*. Lincom Europa, München, 2 edn. (2000)
8. Gobbo, F., Wagemans, J.H.M.: Building argumentative adpositional trees: Towards a high precision method for reconstructing arguments in natural language. In: *Proceedings of the Ninth Conference of the International Society for the Study of Argumentation*. pp. 408–420 (2019)
9. Gobbo, F., Wagemans, J.H.M.: A method for reconstructing first-order arguments in natural language. In: *Proceedings of the 2nd Workshop on Advances in Argumentation in Artificial Intelligence (AI³ 2018)*. pp. 27–23 (2019)
10. Gobbo, F.: *Interlinguistics, a discipline for multilingualism*. Amsterdam University Press, Amsterdam (2015)
11. Gobbo, F.: Are planned languages less complex than natural languages? *Language Sciences* **60**, 36–52 (2017)
12. Gobbo, F.: Beyond the nation-state? the ideology of the esperanto movement between neutralism and multilingualism. *Social inclusion* **5**(4), 38–47 (2017)
13. Gobbo, F.: *Language Games Children Play: Language Invention in a Montessori Primary School*, pp. 1–14. Springer International Publishing, Cham (2019)
14. Gobbo, F., Benini, M.: *Constructive Adpositional Grammars. Foundations of Constructive Linguistics*. Cambridge Scholars Publishing, Newcastle upon Tyne (2011)
15. Gobbo, F., Benini, M.: Dependency and valency. from structural syntax to constructive adpositional grammars. In: In K. Gerdes, E. Hajiov and L. Wanner (Eds.), *Computational Dependency Theory*. pp. 113–135. IOS Press, Amsterdam (2013)
16. Rossi, E.: *L'Europa di domani: Un progetto per gli Stati Uniti d'Europa*. A cura di Mauro Rubino. Stilo editrice, Bari (2014),
17. Searle, J.R.: *Making the Social World: The Structure of Human Civilization*. Oxford University Press, Oxford (2010)
18. Stria, I.: *Esperanto Speakers - an Unclassifiable Community?* Wydawnictwo KUL (2015), instytut Pedagogiki na Katolickim Uniwersytecie Lubelskim Jana Pawła II w Lublinie. Księga Jubileuszowa
19. Tesnière, L.: *Éléments de syntaxe structurale*. Klincksieck, Paris (1959)
20. Tesnière, L.: *Elements of Structural Syntax*. John Benjamins, Amsterdam (2015)
21. Tomasello, M.: *Constructing a language. A Usage-Based Theory of Language Acquisition*. Harvard University Press, Harvard (2003)
22. Wagemans, J.H.M.: Constructing a Periodic Table of Arguments. In: *Argumentation, Objectivity, and Bias: Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation (OSSA)* (2016)

Corpus-Based Empirical Research on Collocation beyond Existing Grammatical Rules: *Make Angry/Mad* as an Example

Ai Inoue¹[0000-0002-4577-753X]

¹ National Defense Academy, 1-10-20 Yokosuka, Kanagawa, Pref. 2398686, JAPAN
aiinoue@nda.ac.jp

Abstract. This corpus-based phraseological research focuses on the transition from the traditional, transitive usage of the existing collocation, *make somebody angry/mad*, to the intransitive utilisation of a collocation, *make angry/mad*. The present study elucidates three principal points. First, *make angry/mad* is established by the analogy of the semantically similar collocation *get angry/mad*. In terms of the syntactic factor, *make* takes the pattern ‘make + an adjective’ [SVC], and *get angry/mad* utilises the same syntactic construction. Second, the linguistic phenomenon is the subject of this study because it is more important to convey the intended meaning without causing a misunderstanding than it is to adhere to existing rules. In other words, the notion of the linguistic economy works in the case of *make mad/angry*. Finally, the research outcome reveals that existing linguistic theories and rules account for only a small part of English usage, most of which evidences numerous uses beyond normative theories and rules.

Keywords: Corpus-based empirical study, Resurgent collocation, Analogy

1 Introduction

This corpus-based empirical research focuses on the collocation *make angry/mad*, which is a seemingly minor error but is an independently used phraseological unit (hence PU) ¹.

It has been widely acknowledged that *make angry/mad* has been regarded as a mistake although it appears as is shown in (1) (italicised by the author as in the following).

- a. Mr-MAHDI: ..., so I did not want them to *make angry*, or I just did not want to make any risk about myself, so I had to postpone the project. (COCA, 2005, SP)

¹ The study uses phraseological units (PUs), a comprehensive term that includes idioms, collocations, phrasal verbs, formulae, proverbs, discourse particles and fixed expressions.

- b. The two biggest bolts (from the 1860 Democrats and the 1912 Republicans) both cost the majority party the presidency. Perhaps it is true that whom the gods would destroy, they first *make mad*. (COCA, 2010, MAG) (1)

(1) creates a hypothesis that *make angry/mad* functioning intransitively is thought to be established by the analogy of *get angry/mad*. The study proves the hypothesis from quantitative and qualitative viewpoints using corpora.

2 Literature Review

This section offers an explanation of *make somebody (sb for short) angry/mad* in previous research.

Make angry/mad has not been fully discussed in previous research, but Swan (2016) [1] mentions that *make* working as an intransitive verb is used in the pattern ‘make C’ (C is an adjective), but that the pattern is old-fashioned and obsolete.

3 Corpora Used in the Study

This study uses the data from the Corpus of Contemporary American English (COCA), British National Corpus (BNC), *WordBanksOnline* (WB) from a synchronic perspective, and from the Corpus of Historical American English (COHA) from a diachronic perspective. I accessed COCA on February 25th, 26th, 27th and 28th and March 1st, 2nd, 5th, 6th and 7th in 2019. In Sections 5, data obtained from COCA show the register, where each example is used. The acronyms MAG for magazine, SP for spoken, and WR for written. I accessed COHA on March 6th in 2019.

4 Research Methods the Study Adopts

This descriptive research places more emphasis on linguistic realities rather than on existing linguistic rules or theories. It adopts the research methodology of attempting to examine actual linguistic phenomena without depending on any major linguistic theory. This investigation is grounded in the theory of semantic syntax which is a continuation of the tradition revised and developed under the influence of various linguistic theories developed in the United States and elsewhere. Its thesis is that the meaning of a word or a phrase is closely related to the syntactic feature of the word or the phrase.

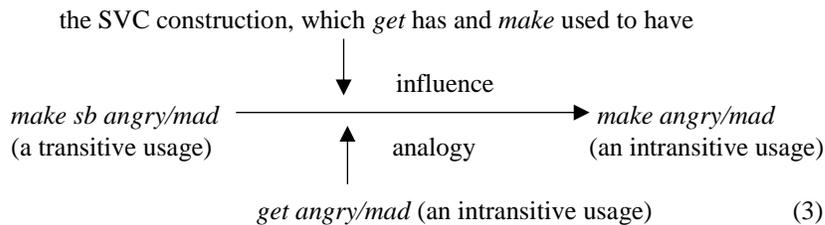
5 Quantitative and Qualitative Results of *Make Angry/Mad*

I investigate which tense and aspect *make angry/mad* is used based on the data obtained from the corpora. The results show that *make angry/mad* appears only in present tense and past tense although it is not necessarily used frequently compared to the frequency of *get angry/mad*. The examples below are quoted from the corpora.

- a. Mr-MAHDI: ..., so I did not want them to *make angry*, or I just did not want to make any risk about myself, so I had to postpone the project. (COCA, 2005, SP)
- b. The two biggest bolts (from the 1860 Democrats and the 1912 Republicans) both cost the majority party the presidency. Perhaps it is true that whom the gods would destroy, they first *make mad*. (COCA, 2010, MAG)
- c. It is said that those whom the gods wish to destroy, they first *make mad*, and it is clear that Tory Ministers are mad. (BNC, 1992, WR) (2)

(2) suggests the following points: (i) *make angry/mad* is used intransitively to express anger the same as *get angry/mad*; (ii) a careful examination of (2) shows that *make mad* in (2b,c) is described in a proverb dictionary, so *make mad* originates from an old proverb and remains fossilised; (iii) there is little semantic difference between *make angry* and *make mad*; (iv) *make angry/mad* is perhaps not a minor error, due to the fact that it is observed in written registers of the corpora and (v) *make angry/mad* appears both in American English and British English.

I maintain that *make angry/mad* comes to intransitive function as follows: *make angry/mad* is established by the analogy of *get angry/mad*; in other words, the original construction [make sb angry/mad] changes into *make angry/mad*, which is influenced by *get angry/mad* without causing a semantic change. In other words, *make angry/mad* is formed due to a syntactic-based contributing factor. Also, *make* has the old-fashioned and obsolete construction [make C]. The construction has a syntactic influence on forming *make angry/mad*. In other words, it is safe to assert that *make angry/mad* is established by these two contributing factors, i.e. semantic and syntactic contributing factors. The diagram of *make angry/mad* is illustrated in (3).



Consequently, (3) proposes that the intransitive meaning of ‘anger’ is reflected in the syntactic pattern *make angry/mad*, and that least effort of linguistic economy underlies to form *make angry/mad* because the original syntactic pattern *make sb angry/mad* is simplified into *make angry/mad*.

6 Informants’ Elicitation

I asked native speakers of English (a Canadian, an Australian, two English people, two Americans) to cooperate in the following two investigations to examine whether the results shown in the previous section are supported: ① to investigate whether the

sentences where *get(got) angry/mad* in (2) changes into *make(made) angry/mad* are acceptable; and ② to investigate whether *make angry/mad* in (2) is acceptable.

As for ①, all informants replied that the sentences in (2), where *get(got) angry/mad* changes to *make(made) angry/mad*, are acceptable. This means that *make angry/mad* is regarded same as *get angry/mad*. As for ②, all informants said that *make angry/mad* in (2) is acceptable and is used to mean ‘anger’.

Hence, the informants’ elicitation shows that the results provided in the study (i.e. *make angry/mad* is used intransitively and is same as *get angry/mad*) are supported.

7 Use of *Make Angry/Mad* from a Diachronic Standpoint

This section examines the diachronic use of *make angry/mad* based on the data obtained from COHA. *Make angry/mad* is observed only in present tense and past tense, as the above example shows. Thus, I retrieved *make(made) angry/mad* used in the same tenses in COHA.

Historically, *make(made) angry/mad* began being observed in the 1810s, although it did not appear frequently.

8 Conclusion

The corpus-based collocational study suggests that *make angry/mad* derived from *make sb angry/mad* is established as a resurgent collocation (i.e. a part of PUs) due to the influence of *get angry/mad*. The phenomenon may have arisen from the concept that it prioritises to express meanings over to adhere to grammatical rules. In other words, linguistic economy works and supports our smooth communication as a background. The study reminds us that only part of our language can be accounted for by rules or theories, and in fact most parts of our languages are diversely used beyond the rules or theories. This especially holds true for PUs, and it is not surprising to find PUs that are not explained by the rules or theories.

Acknowledgement

This research was made possible by the Grant-in-Aid for Young Scientists (B) (Grant number 17K13480). I would like to thank the Japan Society for the Promotion of Science.

Reference

1. Swan, M.: Practical English usage. 4th edn. Oxford University Press, Oxford (2016).

Vivid Phrasal Idioms and the Green New Deal

Teaching Idioms to EAP Students Via Authentic Contexts

Melissa Larsen-Walker

¹ University of South Florida, Tampa FL 33615, USA
mlarsenw@mail.usf.edu

Abstract. Vivid phrasal Idioms (VP Idioms), meaning non-compositional, figurative phrases such as “spill the beans,” occur frequently in English, particularly in conversation [23]. They occur also in the presentational mode, as exemplified by newscasts and political speeches. Yet research suggests that course books for ELLs inadequately address idioms [16, 19, 20]. The current study investigates the effects of exposing learners to VP Idioms in an authentic audiovisual context as recommended by previous research [9, 19, 20, 22]. The authentic context used in this study features U.S. Representative Ocasio-Cortez narrating a story that explains the *Green New Deal*, a plan which mandates an aggressive course of action to combat Global Climate Change. The instructional intervention and data collection took place within one in-tact classroom, wherein the participants passed through both the experimental (authentic context) and control conditions. Participants accessed an instructional website for images and definitions of the six key VP Idioms, three each for the control and experimental phases. Then they viewed “A Message from the Future with Alexandria Ocasio-Cortez” [17] and read the associated script. Student-written dialogues, each of which contain one of the key VP Idioms, were the main data source. Results and pedagogical implications will be discussed.

Keywords: English for Academic Purposes, Idioms, Teaching and Assessment

1 Introduction

Figurative expressions, which are idiomatic in that their meaning is not transparent, are ubiquitous throughout the English language, in both oral and written form. For English Language Learners (ELLs) these idioms both open a door to fluent conversation and an opportunity miscommunication. For example, if a contradictory circumstance occurs, a speaker might respond, “every cloud has a silver lining” or say that the circumstance, “has a silver lining.” The listener cannot decode the meaning from the constituent words since the institutionalized meaning of the phrase does not equate with its literal meaning. Nevertheless, research has found that idiomatic expressions such as “take the bull by the horns,” etc. occur frequently in spoken English [19][20][21][22]. While such expressions occur more frequently in speech than in writing [3][23], there is evidence to suggest that some of these figurative phrases,

such as “cross the Rubicon,” are appropriate to more formal registers of usage, including oral presentations and even academic writing [32].

Regardless of the level of formality, a mastery of idioms requires an awareness of their cultural context. The non-native speaker (NNS) of English who lacks knowledge of these idioms is likely to be baffled by them when s/he meets these phrases in a conversation, oral presentation, or text. This can occur even among learners who have attained a high level of English proficiency. Idioms, many of which have word origins rooted in historical fact, deeply reflect culture [16] [19] [20] [27]. Learners who acquire awareness of the cultural context of a word or phrase are likely to avoid such confusion and acquire what Lontos [19] [21] calls Idiomatic Competence, the ability to use an idiom with relatively little effort and in an appropriate way. Sinclair also comments upon this ability to participate fluently in L2 conversation being contingent upon cultural awareness, including the following: “a subliminal mastery of phraseology, the ability to make linguistic and textual inferences, and a knowledge of aspects of culture which are not signaled anywhere in the text, but which are nonetheless known” [30]. Having to rely solely on L1 knowledge of idioms and (often futile) attempts to decompose an English idiom leads to misinterpretation unless this idiom is nearly identical to one found in the learner’s L1 [10] [20] [21] [22]. As instructors who endeavor to elicit the type of fluency that Sinclair describes, we ought to teach idioms in such a way that learners grasp the full connotation, which includes observing how native speakers use them.

1.1 Definition of Terms

The current study explores Vivid Phrasal Idioms, which should be defined. Moon (1997) states that numerous terms for idioms and other phrases exist, yet linguists have not agreed upon any standardized definitions of terms or categories. To address this ambiguity and as a focus of his research, Dr. John Lontos coined the name, Vivid Phrasal Idiom (VP Idiom) which includes many phrases such as “jump the gun,” “bite the bullet,” etc. The characteristics of VP Idioms are: (1) They are non-compositional, as explained above; these phrases cannot be decomposed so as to figure out their meaning based on the constituent words [11] [13] [28]. (2) VP Idioms are not limited to any particular verb tense. For instance, one can “spill the beans,” or say that “He has spilled the beans,” etc. (3) However, most of these phrases are what Fernando and Flavell (1981) call “transformationally deficient;” this means that they are usually inflexible insofar as substituting lexis. For example, the idiom “play second fiddle” cannot be transformed to “play second violin” without losing its figurative meaning. Hence, VP Idioms are conventionalized. These figurative expressions must be phrasal or sentential, so single-word figurative expressions are not included in this category. (4) The learner can easily visualize the idiom given that it has a concrete and literal counterpart, which led Lontos to call them “vivid.” (5) As mentioned, these idioms have dual meaning, in that a literal meaning exists. Nevertheless, analysis of the structure, syntax, semantics of the literal expression is insufficient to decipher the idiom’s figurative meaning [21].

2 Motive for the Study

Despite the importance of VP Idioms, evidence suggests that that teaching materials designed for L2 English learners do not adequately address these phrases [16] [19] [20]. Still, survey data reveals that the learners acknowledge the benefits of acquiring fluency in the use of idioms [16] [19] [20]. This gap in contemporary materials development and practice has motivated the researcher to investigate approaches for teaching idioms.

Various registers of usage employ VP Idioms, and while they occur most frequently in informal, spoken contexts [23], they also occur in other registers through which L2 learners will need to communicate. Myers (2006), analyzed communicative events (written or spoken) into three modes: conversational, presentational, and academic. Understanding these modes is integral to the rationale for teaching idioms. At the beginning of life, one hears and begins to speak L1, in the mode that will remain permanently (in most cases) the register through which one communicates with family and close friends. The interlocutor in a conversation aims primarily to immerse himself/herself into the community and maintain relationships. This mode includes non-specific language and vernacular forms. Conversely, academic communication (in speech or writing) has less emphasis on establishing and maintaining relationships. Rather, it requires precision, correctness, and distancing of the author's self and his/her opinion. A central goal of academic communicative events is to convey scientific objectivity, while elaborating upon the research that has been conducted by others.

Presentational speech events fall into a category neither fully academic nor informal in that the purpose and conventions of speech in this mode differs from either informal speech or academic writing. Examples of communication in the presentational modes include political speeches and news reports. This mode does not permit non-standard usage like conversation yet aims for a deeper connection with the audience than academic discourse. Applying Myers [25] research to learning English idioms, the immigrant or international student may be hampered not only in attempts at conversation but also in his/her understanding of political speeches and news reports if s/he fails to grasp an idiom's meaning.

Various approaches to idioms instruction have been taken, including the use of images associated with the literal meaning of each target idiom, etymologies, conceptual metaphors, etc. The plethora of English idioms makes it necessary for instructors or materials designers to thoughtfully select which idioms to teach due to time constraints. Teaching the most commonly occurring idioms has been recommended by some researchers and pedagogues [19]. Attention is more likely to be engaged if the selection of idioms is at least partly based on student interests [19]. Recent research recommends providing students with examples of idioms as they occur in authentic contexts, such as YouTube videos, etc. [16] [19] [20] [21]. This exploratory study investigates how the use of authentic contexts affects English for Academic Purposes (EAP) students' ability to produce them.

- (1) How does the use of authentic video clips (TV excerpts), printed definitions, and dialogues impact the learner's production of Vivid Phrasal Idioms? The learner's production has been measured by the authenticity and appropriateness of each idiom within the context of the participant's written dialogue.
- (2) What is the impact (if any) of using the authentic contexts on learner's receptive knowledge of idioms as compared to the effect of teaching them without it? The difference in effectiveness would be measured by comparing the posttest scores of the idioms taught through the control condition with the scores of the idioms taught through the experimental condition.

All instructional materials, links and supplementary information were provided for students via a researcher-designed website for learning idioms. <https://idioms.sitey.me/> This site was available to participants throughout the control and experimental phases of the instructional intervention.

3 Literature Review

This concise review of literature has two foci. First, it summarizes results of studies wherein researchers investigated methods for teaching idioms and relevant theory. Next, it explains the Green New Deal because this is the topic of the authentic context presented to the participants. Beginning with instructional interventions, numerous studies have explored the approach of providing students with an etymological elaboration of the idiom, which may come before presenting them with its meaning [1] [5] [6]. With rare exceptions, this approach has been found to be effective whether the study uses the Boers CALL tool, as in the studies above, or is implemented within a classroom without it [2] [27]. A more common approach is to present students with a picture of the literal meaning of the idiom [12] [15] [34]. Those textbooks that address idioms usually contain such associated images. While some researchers have found this to be effective, others have not. For example, results of a study by Boers and colleagues [8] suggest that providing students with associated pictures helps them to remember the meaning of the idiom but not its precise form. As noted above, idioms lack flexibility, so fluency in idiom usage, i.e. Idiomatic Competence, is contingent upon mastery of form as well as meaning. The accessibility of video via internet and YouTube has made it practical to use authentic contexts, in print as well as video, as a material to teach students about how native speakers and highly proficient NNS use idioms. Finally, there have been studies that investigate a multi-modal approach, which combines having students generate their own images with providing authentic contexts [13] [36]. This literature review touches upon the most common and replicated types of instructional interventions. As such, it is not exhaustive. The topic of the audio-visual context given to the participants is described next.

Possibly the most pressing issue worldwide for the 21st century is Global Climate Change. While the politics of the current U.S. administration denies its existence and has endeavored to strengthen the business of coal extraction and exploration for petro-

leum, many citizens in the U.S. grasp the reality and severity of this problem and support measures to fight the worst effects Global Climate Change, which has been caused by the burning of fossil fuels. The controversial Green New Deal exists as a plan rather than a bill ready to be discussed and voted upon by the U.S. Congress. The plan includes a provision mandating the implementation of a fossil-fuel free power in the U.S. ten years after becoming a law [26] [32], which is ambitious considering the extent to which the world's largest polluter consumes fossil fuels for home and industrial uses, not to mention transportation. The four major goals are: (1) It would mandate the transformation of the United States' energy sources to renewables, such as solar and wind. This plan includes creating a smart-grid, retrofitting power plants, and updating transportation, both mass transit and the options for individual use. "The draft proposal also mentions upgrading every residential and industrial building across the country for state-of-the-art energy efficiency and decarbonizing manufacturing, transportation, and agriculture" [32]. (2) It aims to restructure the economy so that even service sector workers can earn a living wage. (3) This plan would refute the argument of those who say that transforming America's energy would mean unemployment for numerous coal miners and petroleum workers. Rather the Green New Deal will provide re-training for those who currently work in toxic, pollution creating jobs (such as oil, gas, coal industries) and will pay them to be part of the restructuring process. (4) Finally, it will provide funding for communities affected by these changes. The video, "A Message from the Future," [17] provided audio-visual support for the EAP students in learning about the Green New Deal and a few idioms that occur within it.

4 Method

"A Message from the Future with AOC" was presented to students because of the importance of the topic, the prevalence of idioms, and the fact that the students could hear the script being read aloud accompanied by images illustrating the content of the narrative. This follows Paivio's [29] Dual Coding Theory, which posits that three channels of cognitive processing exist. Furthermore, having images associated with verbal input enables the learner to process the content through two channels, enhancing comprehension. The researcher collected posttest data in addition to collecting and analyzing the students use of idioms in students' dialogues.

This study builds upon the exemplary research by Freyn and Gross [13], yet with a smaller and less linguistically homogeneous group of students. Their study of Ecuadorian ELLs included four classrooms, two of which both observed idioms spoken in authentic contexts and created their own images. Given the smaller sample, the current study uses one group of learners who have passed sequentially through the control and the authentic context conditions. Three of the six target idioms have been taught through each of the conditions listed above. The target idioms include: "put the cart before the horse," "cross the Rubicon," "bite the bullet," "go to bat for someone," "the lion's share," and "kick off." The participants, who were in an advanced Level 5 writing course, are between the ages of 19 and 45. The native languages of partici-

pants are Spanish (12), Vietnamese (2), and Russian (1). The idioms were selected on the basis of likelihood to enhance TOEFL performance, such as “bite the bullet” and “cross the Rubicon,” frequency, such as “put the cart before the horse” and “bite the bullet,” and the appearance of the idioms in the audio-visual authentic context.

The instructional intervention proceeded as follows: On the first day, students became acquainted with the instructional website, including the purpose for learning idioms and the meaning of Idiomatic Competence. They then learned the first three idioms, based upon the definition, explicitly provided, along with examples. Finally, they wrote a dialogue based upon one of these idioms, “cross the Rubicon,” “bite the bullet,” and “put the cart before the horse.” On the next day of instruction, the students watched, “A Message from the Future.” Then they reviewed the script and watched it again after which the instructor/researcher clarified the definitions for any non-idiomatic words that they did not understand. Next, students accessed the page on the instructional website that contains the target idioms, wherein they could read the definition, followed by exemplification and class discussion. Finally, the students wrote their own dialogues using one of three idioms presented under the experimental condition, “the lion’s share,” “go to bat for _____,” and “kick-off.” By the end of data collection, each student had written two dialogues, one for each set of three idioms. The researcher evaluated each written dialogue based upon a rubric that measured writing skills in addition to the appropriateness of idiom use. Posttest data was collected one week later; there was no pretest.

5 Results and Discussion

Despite the small sample size, the researcher analyzed both the learner-generated dialogues and the posttest scores, endeavoring to determine the effectiveness of the intervention. Posttest results for idioms taught under the experimental (authentic context) condition were $N = 14$, $M = 93$ $SD = 1.20$; likewise, for the control idioms. Regarding research question #2, based upon posttest scores, there appears to be no difference in learning gains between the control and the experimental condition, i.e. authentic contexts. However, factors exist that may have compromised the validity of the assessment. Students were tested soon after instruction, there were only a total of six items, and all of these items were multiple choice matching of the definition with the idiom. The researcher refrains from drawing any conclusion *based on test scores* due to the small number of participants (14) and brevity of the test. The results for students’ dialogues show a different pattern. Despite some grammatical errors, the student writers were able to employ the VP Idioms appropriately within the dialogues, particularly those containing idioms that were taught under the experimental condition.

Analysis reveals that in some cases, students were able to use the idioms appropriately. Of the three control idioms, only dialogues containing “cross the Rubicon” were rated as having appropriate and naturalistic use of the idiom, whereas examples of appropriate usage appeared in all dialogues for the experimental items. Excerpts from two of these written dialogues are shown below. Given the limitations of space,

other dialogues do not appear. The current study served as an exploration and pilot study. The dissertation research, which includes four of the idioms from the pilot test, plus five more and additional web pages, has both a pretest and posttest, which should reveal changes in students' receptive knowledge of the key VP Idioms. Future research with a larger sample may reveal more about the effectiveness of using authentic contexts in teaching VP Idioms.

“The lion’s share” – Short summary: The characters are an office manager and her supervisor for the Chevrolet Company. The regional sales departments were competing to see which salesman could sell the most cars.

Boss: Hey Hellen, could you send me the report of all those sales?

Hellen: Of course, sir, here it is!

Boss: Perfect, let me see the papers. Hmm, ok Carl has sold 28 cars. Brenda has 25 cars sold [sic]. Steve has 31 cars sold [sic]. And wow! Chris has 40 cars sold [sic]. Incredible! He got **the lion’s share** of the business! He deserves the prize. Please Hellen, call all the groups to say who the winner is.

Hellen: Yes, sir, our dealership is excited. Could you tell me what the prize is?

Boss: Of course, Hellen, the prize is a new car for the winner.

“Cross the Rubicon” - Short summary: The main characters in the dialogue are two young adult immigrants to the U.S. One of them tells her friend about her desire to move out of the home of her father and American step-mother.

Beatriz: Since I finished the high school, I am the only one that “have time to clean the house” [sic] so I am the only one that does it. Also, they try to compare me with my siblings, but they have the advantage of language because they were born in the US.

Marla: Oh, I didn’t know that it was like that. . .

Beatriz: I try my best, but it is never enough. I work, I go to the school, I almost do everything in the house . . . I just want to leave my house. I don’t know where I’m going, but I don’t care!!

Marla: Bea, you must calm down and think carefully about your situation. You are only 20 and your work is not stable enough to be independent.

Beatriz: I know that Marla, but it’s hard.

Marla: If you leave your house you are going to **Cross the Rubicon**, and you will be living by yourself. That is difficult if you are not prepared. I think it is better if you talk to your family.

References

1. Bagheri, M.S. & Fazel, I. Effects of etymological elaboration on the EFL learners' comprehension and retention of idioms. *Journal of Pan-Pacific Association of Applied Linguistics*, 14(1), 45-55 (2010).
2. Baleghizadeh, S. & Bagheri, M.M. The effect of etymology elaboration on EFL learners' comprehension and retention of idioms. *The Southeast Asian Journal of English Language Studies*, 18(1), 23 – 32 (2012).

3. Biber, D., Johansson, G., Leech, Conrad, S. & Finegan, E. *Longman Grammar of Spoken and Written English*. Longman, Harlow, England (1999).
4. Boers, F. Applied linguistics perspectives on cross-cultural variation in conceptual metaphor. *Metaphor and Symbol*, 18(4), 231-238 (2003).
5. Boers, F., Demecheleer, M., & Eyckmans, J. Cross-cultural variation as a variable in comprehending and remembering figurative idioms. *European Journal of English Studies*, 8(3), 375-388 (2004).
6. Boers, F., Eyckmans, J. & Stengers, F. Presenting figurative idioms with a touch of etymology: more than mere mnemonics. *Language Teaching Research* (11)1, 43–62 (2007), doi: 10.1177/1362168806072460.
7. Boers, F., Lindstromberg, S., Littlemore, J., Stengers, H. and Eyckmans, J. Variables in the mnemonic effectiveness of pictorial elucidation. In Boers, F. and Lindstromberg, S. (Eds.), *Cognitive linguistic approaches to teaching vocabulary and phraseology* (pp. 189–216). Berlin/New York: Mouton de Gruyter (2008).
8. Boers, F., Piriz, A.M.P., Stengers, H., Eyckmans, J. Does pictorial elucidation foster recollection of idioms? *Language Teaching Research*, 13(4), 237-382 (2009).
9. Cakir, I. How do learners perceive idioms in EFL classes? *Ekev Academic Review*, 15(47), 371-381 (2011).
10. Carrol, G., Conklin, K. & Gyllstad, H. Found in translation: The influence of the L1 on the reading of idioms in a L2. *Studies in Second Language Acquisition*, 38, 403–443 (2016), doi: 10.1017/S0272263115000492.
11. Fernando, C., & Flavell, R. *On Idiom: Critical views and perspectives*. Exeter linguistics studies (vol. 5). Exeter: University of Exeter (1981).
12. Fotovatnia, Z. & Khaki, G. The effect of three techniques for teaching English idioms to Iranian TEFL undergraduates *Theory and Practice in Language Studies*, 2(2), 272-281 (2012), doi:10.4304/tpls.2.2.272-28.
13. Freyn, A. & Gross, S. An empirical study of Ecuadorian EFL learners' comprehension of English idioms using a multi-modal approach. *Theory and Practice in Language Studies*, 7(11) (2017).
14. Galal, M.M. "My patience is exhausted" and "Nafida Sabrii": A conceptual metaphor account of "patience" idioms in English and Arabic. *International Journal of English Linguistics*, 4(4) (2014), doi: 10.5539/ijel.v4n4p22.
15. Hagshenas, M.S.M. & Hashemian, M. A comparative study of the effectiveness of two strategies of etymological elaboration and pictorial elucidation on idiom learning: A case of young EFL Iranian learners. *English Language Teaching*, 9(8), 140-151 (2016), doi: 10.5539/elt.v9n8p140.
16. Khan, O., & Daskin, N. C. 'You reap what you sow' Idioms in materials designed by EFL teacher-trainees. *Novitas ROYAL (Research on Youth and Language)*, 8(2), 97-118 (2014).
17. Klein, N., director, Lewis, A. & Ocasio-Cortez, A., authors, Crabapple, A., illustrator. *A Message from the Future with Alexandria Ocasio-Cortez*. YouTube, The Intercept, 17 Apr. 2019, www.youtube.com/watch?reload=9&v=d9uTH0iprVQ&feature=youtu.be.
18. Kong, Y. The study of English and Chinese numerical idioms and their translation. *Journal of Language Teaching and Research*, 5(2), 446-451 (2014).

19. Liantas, J.I. Developing idiomatic competence in the ESOL classroom: A pragmatic Account. *TESOL Journal* 6.4, 621-658 (2015).
20. Liantas, J.I. Toward a critical pedagogy or idiomaticity. *Indian Journal of Applied Linguistics*, 34(1-2), 11-30 (2008).
21. Liantas, J.I. Vivid phrasal idioms and the lexical-image continuum. *Issues in Applied Linguistics*, 13(1), 71-109 (2002).
22. Liantas, J. Why teach idioms? A challenge to the profession. *Iranian Journal of Language Teaching Research* 5(3), 5-25 (2017).
23. Liu, D. Idioms: Description, comprehension, acquisition and pedagogy. In Eli Hinkel (ed.) *ESL & Applied Linguistics Professional Series*. London: Routledge Taylor and Francis Group (2008).
24. Moon, R. *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Clarendon Press. (1997).
25. Myers, M. *Changing our minds: Negotiating English and literacy*. Urbana, IL: National Council for the Teachers of English (2006).
26. Nelson, C. "What's the 'Green New Deal' and why do environmentalists want it?" *MPR News*. (19 Nov. 2018). <https://www.mprnews.org/story/2018/11/19/green-new-deal-omar-ocasio-cortez>
27. Noroozi, I. & Salehi, H. The effect of the etymological elaboration and rote memorization on learning idioms by Iranian EFL learners. *Journal of Language Teaching and Research*, 4(4), 845-851 (2013).
28. Nunberg, G., Sag, I.A. & Wasow, T. Idioms. *Language*, 70(3), 491-538 (1994).
29. Paivio, A. *Mental Representations: A Dual-Coding Approach*, Oxford University Press, New York (1990).
30. Prodromou, L. Idiomaticity and the non-native speaker. *English Today* 74, 19(2), 42-48 (2003), doi: 10.1017/S0266078403002086.
31. Samani, E.R. & Hashemi, M. The effect of conceptual metaphors on learning idioms by L2 learners. *International Journal of English Linguistics*, 2(1) (2012).
32. Smith, H. What is this *Green New Deal* anyway? Alexandria Ocasio-Cortez has a plan to tackle climate change (28 Nov. 2018). <https://www.sierraclub.org/sierra/what-green-new-deal-anyway-alexandria-ocasio-cortez>
33. Stirling, B. *500 words, phrases, and idioms for the TOEFL iBT plus typing strategies*. Los Angeles: Nova Press (2015).
34. Summary of the Green New Deal. Green Party of the United States (2019). <https://gp.us.org/organizing-tools/the-green-new-deal/>
35. Tabatabaei, O. & Mirzaei, M. Comprehension and idiom learning of Iranian EFL learners. *Journal of Educational and Social Research*, 4(1), 46-56 (2016), doi:10.5901/jesr.2014.v4n1p45.
36. Vasiljevic, Z. Imagery and idiom teaching (effects of learner-generated illustrations and etymology). *International Journal of Arts & Sciences*, 8(01), 25-42 (2015).
37. Vasiljevic, Z. Teaching idioms through pictorial elucidation. *The Journal of Asia TEFL*, 9(3), 75-105 (2012).

Extracción Terminológica Basada en Corpus Para la Traducción de Fichas Técnicas de Impresoras 3D

Ángela Luque Giráldez and Míriam Seghiri¹

¹ University of Málaga, 29010 Málaga, Spain

Angelaluquegiralddez@gmail.com
Seghiri@uma.es

Abstract. En el presente trabajo presentaremos una metodología para la creación de un glosario bilingüe y bidireccional (inglés-español/español-inglés) que será de utilidad para la traducción fichas técnicas de impresoras 3D. La extracción de los términos que integrarán el mencionado glosario se realizará a partir de un corpus bilingüe creado a tal efecto al que denominaremos 3DCOR. De este modo, aunamos el formato preferido por los traductores (durante la fase documental), como es el glosario, pero cuya implementación se basará en el recurso ideal para los investigadores, como es el corpus. En cuanto al campo elegido, como hemos apuntado, será técnico, y más concretamente aquel de las impresoras 3D por su novedad y auge en el mercado (cfr. TMT Deloitte, 2019), y, en lo que respecta al género, se ha seleccionado la ficha técnica, pues es uno de los que más demanda genera (cfr. Resolución del Consejo Europeo 98/C 411/01).

Keywords: Corpus paralelo, Extracción terminológica, Glosario, Traducción técnica, Ficha técnica, de mpresoras 3D

1 Introducción

La tecnología avanza a gran velocidad hasta llegar al punto de que ya se comercializan impresoras 3D que pueden ser adquiridas por cualquier usuario común. De hecho, según estudios de TMT Deloitte [1], la industria de la impresión 3D está creciendo a un ritmo aproximado del 12,5 % al año y se prevé que tan solo en 2019 las ventas superen los 2 700 millones de dólares, llegando a los 3 000 millones en 2020. Esta situación, no cabe duda, contribuirá al aumento de las traducciones técnicas en este campo y, en concreto, de sus manuales de instrucciones y fichas técnicas, ya que la *Resolución del Consejo de 17 de diciembre de 1998 sobre las instrucciones de uso de los bienes de consumo técnicos (98/C 411/01)*¹ establece en su quinto artículo lo siguiente:

¹ La presente *Resolución del Consejo de 17 de diciembre de 1998 sobre las instrucciones de uso de los bienes de consumo (98/C 411/01)* puede consultarse en: <https://europa.eu/!Hj67Ug>.

Los consumidores deberán poder acceder fácilmente a las instrucciones de uso al menos en su propio idioma oficial de la Comunidad de manera que el usuario pueda leerlas y comprenderlas con facilidad.

Por razones de claridad y facilidad de uso, cada versión lingüística deberá estar separada de las demás.

Las traducciones deberán basarse sólo en el idioma original y tener en cuenta las características culturales distintivas de la zona en la que se usa el idioma correspondiente; esto requiere que las traducciones sean hechas por expertos con la formación adecuada, que utilicen el idioma de los consumidores a los que está destinado el producto, y que, en la medida de lo posible, sean sometidas a una prueba de comprensión de los consumidores.

De esta manera, la necesidad de traducciones de calidad de los manuales y fichas técnicas viene dada, además, por imperativo legal. Sin embargo, en el campo que nos ocupa, el de las impresoras 3D, al tratarse de un producto novedoso, es de prever que el traductor se encontrará con una escasez de recursos para abordar su traducción. Y es así como surge el objetivo principal del este trabajo: la creación de un glosario bilingüe y bidireccional (inglés-español/español-inglés) que será de utilidad para la traducción fichas técnicas de impresoras 3D. La extracción de los términos que integrarán el mencionado glosario se realizará a partir de un corpus bilingüe creado a tal efecto. De este modo, aunamos el formato preferido por los traductores (en particular los noveles) durante la fase documental, como es el glosario, pero cuya implementación se basará en el recurso ideal para los investigadores, como es el corpus (cfr. Seghiri, 2015 [2]).

2 Definición de *Ficha Técnica*

Autores como Gamero (2001) [3] no recogen la denominación de *ficha técnica* dentro de los géneros del campo de la técnica. No obstante, son cada vez más los investigadores, como Byrne (2012) [4], que señalan la ficha técnica como un género independiente con características propias. Así, podemos definir la ficha técnica como un documento que describe las características principales, la composición y las aplicaciones de un producto y que aporta información detallada sobre una serie de aspectos del producto en cuestión. La información suele aparecer presentada en tablas y difícilmente aparecen oraciones completas. Por último, el foco predominante es el foco expositivo, a diferencia de los manuales, donde predomina el foco exhortativo.

3 Creación de un Corpus de Impresoras 3D

Para proceder a la creación del corpus a partir del cual se implementará el glosario es necesario tener claro su diseño para, seguidamente, aplicar un protocolo de compilación y alineación, todo ello siguiendo los postulados expuestos por Seghiri (2006, 2015 y 2017) [5, 2, 6].

3.1 Diseño

Es frecuente encontrar en la red fichas técnicas escritas originariamente en lengua inglesa junto a sus traducciones a diferentes lenguas. Aprovechando este hecho, nos proponemos compilar un corpus *paralelo* de fichas técnicas de impresoras 3D disponibles en Internet; es decir, el corpus estará integrado por bitextos, ya que los documentos que se incluirán serán fichas técnicas originales (escritas originariamente en inglés) y sus traducciones (al español); y, por consiguiente, también será *bilingüe* y *monodireccional*. A su vez, es un corpus *virtual*, pues está integrado por textos descargados exclusivamente de la red, y *textual*, dado que se incorporarán las fichas al completo (y no fragmentos de estas); por último, será *especializado*, y más concretamente *técnico*, en el género de fichas técnicas y la temática de impresoras 3D.

3.2 Protocolo de Compilación y Alineación

Una vez que se ha diseñado el corpus, se procederá en dos fases (cfr. Seghiri, 2006, 2015 y 2017) [5, 2, 6]: la primera fase se dedicará al protocolo de compilación dividida en cinco pasos, a saber, búsqueda, descarga, formato, almacenamiento y determinación de la representatividad cuantitativa; posteriormente, le seguirá una segunda fase consistente en la alineación de los bitextos.

Fase 1: Compilación. Una vez establecido el diseño del corpus, se compilará el corpus, en cinco pasos, con objeto de asegurar la representatividad cualitativa y cuantitativa de la muestra:

Búsqueda: El primer paso consiste en el acceso a la información y la localización de las fichas técnicas que se incluirán en el corpus. Como se trata de un corpus virtual, se descargarán los textos exclusivamente de Internet. Para ello, nos dirigiremos a páginas de empresas de comercialización de impresoras 3D, como HP, Bq, BCN3D o XYZ Printing, por mencionar solo algunas de las más relevantes.

Descarga: El segundo paso consiste en la descarga de las fichas técnicas de las impresoras 3D, que puede llevarse a cabo de forma manual (recurriendo a las teclas Ctrl+G), o bien se puede ir más allá a través del empleo de programas, como BootCat², que permiten la descarga de textos en lotes desde una página web determinada mediante el uso de palabras clave.

Formato: El tercer paso consiste en dar el mismo formato a todos los textos para que el programa de gestión de corpus pueda interrogarlos. En este sentido, todas las fichas técnicas que se han descargado se encuentran en formato HTML (.html) o PDF (.pdf), por lo que es necesario convertirlas a ASCII o texto plano (.txt). Para ello, puede sencillamente copiarse el texto y pegarlo en un archivo .txt. Si la ficha en PDF se encontrara encriptada o bloqueada, este proceso se puede llevar a cabo con la ayuda de un programa de reconocimiento de OCR, como Abby FineReader³.

² El programa BootCat puede descargarse en: <https://bootcat.dipintra.it/?section=download>.

³ El programa Abbyy FineReader puede descargarse en: <https://www.abbyy.com/es-la/bajar/>.

Almacenamiento: El cuarto paso consiste en codificar y archivar todos los textos descargados en carpetas y subcarpetas. Para ello, se creó, en primer lugar, una carpeta llamada «Fichas técnicas» que se divide en dos subcarpetas, una para cada lengua de trabajo: para los textos en español, denominada «ES», y para los textos en inglés, llamada «EN». Dentro de estas carpetas destinadas a cada lengua se han creado paralelamente dos más, una llamada «PDF-HTML», en la que se incluirán los documentos en su formato original, y otra llamada «TXT», en la que se almacenarán los textos ya convertidos a texto plano. Una vez estructuradas las carpetas, se organizaron los textos siguiendo una codificación, que permita su organización y explotación en paralelo (así como futuras ampliaciones del corpus, incluso a otras lenguas). De esta forma, la codificación ideada es la siguiente:

- Número: 01, 02, 03, etc.
- Original o traducción: texto original (TO)/texto meta (TM)
- Lengua: español (ES)/inglés (EN)
- Género: fichas técnicas (FT)

Así, el primer texto descargado en lengua inglesa se denominará 01TOENFT, el segundo 02TOENFT y así sucesivamente, mientras que sus traducciones se codificarán como 01TMESFT, 02TMESFT, etc. respectivamente. La codificación también puede llevarse a cabo de forma automática gracias a programas como Lupas Rename⁴.

Tras la aplicación de los cuatro primeros pasos hemos asegurado la representatividad cualitativa de la muestra compilada y el resultado ha sido la creación de un corpus *paralelo monodireccional* (compuesto por fichas técnicas redactadas originariamente en inglés y sus traducciones al español), *virtual* (integrado exclusivamente por documentos electrónicos), *bilingüe* (inglés-español) y *textual* (recoge fichas completas), que se encuentra integrado por 110 fichas técnicas, de las cuales 55 son en inglés y 55 en español.

El último paso sería, una vez asegurada la calidad, determinar si se ha alcanzado la representatividad desde el punto de vista de la cantidad a través del empleo del programa ReCor⁵. Esta herramienta fue diseñada por Corpas Pastor y Seghiri (2007) [9], por la que recibieron el Premio de Tecnología de la Traducción de España en 2008, y sirve para determinar el tamaño mínimo de un corpus dado; así, en palabras de Corpas Pastor y Seghiri (2007) [9], para calcular el tamaño mínimo del corpus se establece:

[...] el umbral mínimo de representatividad a partir de un algoritmo (N-Cor) de análisis de la densidad léxica en función del aumento incremental del corpus [...]. Se analizan gradualmente todos los archivos que componen el corpus, extrayendo

⁴ El programa Lupas Rename puede descargarse en: <https://es.ccm.net/download/descargar-456-lupas-rename>.

⁵ El programa ReCor se encuentra patentado y su licencia de uso gratuita puede solicitarse a través del siguiente correo electrónico: alinares@uma.es.

información sobre la frecuencia de palabras tipo (*types*) y las ocurrencias o instancias (*tokens*) de cada archivo del corpus.

Tras subir los documentos al programa, este devuelve unas gráficas con las que se ha podido determinar que el subcorpus español es representativo a partir de 47 documentos, con 2 504 *types* y 12 765 *tokens*. De la misma manera, el subcorpus en lengua inglesa es representativo a partir de 48 documentos, 2 387 *types* y 12 056 *tokens*.

El resultado obtenido tras esta primera fase es un corpus representativo tanto desde el punto de vista cuantitativo (gracias a la aplicación de ReCor) como cualitativo (gracias al protocolo de diseño y compilación seguidos), al que denominaremos 3DCOR.

Fase 2: Alineación. Para la extracción terminológica a partir del corpus de bitextos emplearemos, como veremos más adelante, LexTerm⁶. Este programa requiere de una alineación previa de los textos del corpus (cada original con su respectiva traducción), y que supone la segunda fase del proceso. Para alinear los archivos recurriremos a LF Aligner⁷ que presenta una interfaz y un funcionamiento sencillos. Este programa detecta automáticamente los segmentos equivalentes y permite una revisión manual en la que se pueden unir o separar los segmentos a través de los botones «Merge», «Split», «Shift up» y «Shift down».

4 Extracción de Unidades Terminológicas

Alineado el corpus de bitextos 3DCOR, utilizaremos el software Lexterm para llevar a cabo la extracción terminológica.

4.1 Identificación de Candidatos a Término

Una vez alineados los textos del corpus, se puede proceder a la identificación de los candidatos a término con el programa LexTerm. Para ello, subiremos los textos al programa y seleccionamos en la barra de herramientas la opción «n-gramas», con objeto de que se active la búsqueda de términos. El programa creará, así, una lista con todos los candidatos a términos encontrados. En la primera columna aparece un cuadrado para seleccionar (o no) los candidatos a término y exportarlos a una futura lista; en la segunda columna se indica el número de veces que aparece el término en cuestión en el texto; en la tercera columna se recoge el término en cuestión, que puede modificarse de forma manual si se necesita; y, en la cuarta y última columna, se albergan las posibles traducciones localizadas. En este punto cabe indicar que, para que el programa extraiga las traducciones de cada término, se debe hacer clic en el botón «Tond» de la barra de herramientas, donde aparecerá una pequeña pestaña que contiene

⁶ El programa Lexterm puede descargarse en: <http://aulaint.es/software-libre-para-traductores-e-interpretres/herramientas-linguisticas/>.

⁷ El programa LF Aligner puede descargarse en: <https://lf-aligner.soft112.com/>.

esta opción. Si no aparece el equivalente adecuado o se duda sobre su validez, se puede ver el término en su contexto pinchando en «Cerca» y, además, siempre cabe la posibilidad de escribir el equivalente de forma manual.

Cuando se ha finalizado la selección de los candidatos a término, se puede guardar el corpus y también se puede exportar una lista con los términos seleccionados en un archivo TXT. En este caso, se guardarán en un archivo TXT todos los términos pertenecientes a los 55 textos que conforman el corpus con objeto de crear, como veremos a continuación, un glosario.

4.2 Creación de un Glosario Bilingüe y Bidireccional

Una vez analizados y extraídos los listados de términos seleccionados, procederemos a la creación del glosario bilingüe y bidireccional para la traducción de fichas técnicas de impresoras 3D.

En primer lugar, se unirán los 55 archivos para manipularlos de forma más sencilla utilizando la herramienta online Files Merge⁸ gratuita y de fácil uso.

A continuación, se abre el archivo TXT que aparecerá con los términos separados por una tabulación y se copiarán a un archivo Excel, en dos columnas, la primera en inglés y la segunda en español. Dado que se han unido todos los archivos en un único documento (ahora en Excel), los términos aparecerán desordenados, por lo que habrá que ordenarlos. Para ello, seleccionaremos en la barra de herramientas de Excel la pestaña «Datos», que contiene la opción «Ordenar» en la que se seleccionará el criterio «A a Z». Así, se ordenará alfabéticamente la columna de términos en lengua inglesa (junto a sus respectivos equivalentes). Procederemos a eliminar los términos repetidos y, con ello, quedará listo el glosario inglés-español.

Para obtener el glosario español-inglés basta simplemente con cambiar el orden de las columnas y elegir de nuevo la opción «Ordenar de A a Z». Así, quedarán ordenados alfabéticamente los términos en español junto a sus equivalentes en lengua inglesa, conformando el glosario en su versión español-inglés.

Por último, además de en Excel⁹, el glosario bilingüe y bidireccional creado se ha almacenado en formato PDF¹⁰ y en Word¹¹, con objeto de que el traductor pueda utilizar tres formatos de salida, .xls, .doc y .pdf, en función de sus preferencias.

5 Conclusiones

El presente trabajo ha tenido como objetivo principal la creación de un glosario bilingüe y bidireccional (inglés-español/español-inglés) para la traducción de fichas técnicas de impresoras 3D. Este glosario ha sido creado a partir de un corpus diseñado para tal fin, al que hemos denominado 3DCOR. Así, hemos aunado el recurso documental preferido por los traductores, el glosario, con aquel preferido por múltiples

⁸ El programa Files Merge puede utilizarse online en: <https://www.filesmerge.com/sp/>.

⁹ El glosario en Excel puede consultarse en: <https://bit.ly/2lufSPV>.

¹⁰ El glosario en PDF puede consultarse en: <https://bit.ly/2IR6GoH>.

¹¹ El glosario en Word puede consultarse en: <https://bit.ly/2lwYav7>.

investigadores, como es el corpus. El resultado ha sido la creación de un glosario compuesto por 148 términos, y sus respectivos equivalentes, que esperamos que sea de utilidad para realizar traducciones directas o inversas de fichas técnicas de impresoras 3D, así como de otros géneros análogos (como, por ejemplo, los manuales de instrucciones) y temáticas (como aquellas de lápices o escáneres 3D). Asimismo, el corpus 3DCOR podría abrir múltiples líneas de investigación desde el punto de vista monolingüe y monocultural, como desde el punto de vista de la traducción. Por último, la metodología aquí presentada puede ser aplicada para la creación de cualquier otro glosario bilingüe o multilingüe basado en corpus.

6 Agradecimientos

El presente volumen ha sido realizado en el seno de la red temática TRAJUTEC y de la red docente de excelencia TACTRAD (Ref. 719/2018), ambas de la Universidad de Málaga, así como en el marco de los proyectos VIP (Ref. FF12016-75831-P), NOVATIC (Ref. PIE15-145, UMA), INTERPRETA 2.0 (Ref. PIE17-015, UMA), El traductor autónomo: fiscalidad, impuestos y empleabilidad (Ref. 433/2019, UMA) y PROFETA (Ref. PIE19-033, UMA).

Referencias

1. Technology, media & telecommunications Deloitte, <https://www2.deloitte.com/content/dam/Deloitte/ec/Documents/technology-media-telecommunications/Deloitte-ES-TMT-Trends-2019-Folleto.pdf>, last accessed 2019/05/06.
2. Seghiri, M.: Determinación de la representatividad cuantitativa de un corpus ad hoc bilingüe (inglés-español) de manuales de instrucciones generales de lectores electrónicos/Establishing the quantitative representativeness of an E-Reader User's Guide ad hoc corpus (English-Spanish). En: Sánchez Nieto, M. (ed.). *Corpus-based Translation and Interpreting Studies: From description to application*. pp. 125-146. Frank & Timme, Berlín (2015).
3. Gamero Pérez, S.: *La traducción de textos técnicos*. Ariel, Barcelona (2001).
4. Byrne, J.: *Scientific and Technical Translation Explained. A Nuts and Bolts Guide for Beginners*. St. Jerome Publishing, Manchester (2012).
5. Seghiri, M.: *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad*. SPICUM, Málaga (2006).
6. Seghiri, M.: Metodología de elaboración de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus para la traducción de manuales de instrucciones de televisores. En *Babel*, 63 (1), pp. 43-64 (2017).
7. EAGLES: Preliminary Recommendations on Corpus Typology. En *EAGLES* (1996), <http://www.ilc.cnr.it/EAGLES96/corpus/corpus.html>, last accessed 2019/05/10.
8. Seghiri, M.: *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad*. SPICUM, Málaga (2006).
9. Corpas Pastor, G., Seghiri, M.: Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor». *Procesamiento del Lenguaje Natural* 39, 165-172 (2007).

Corpus Analysis of Complex Names with Common Nouns in Croatian

Ivana Matas Ivanković^[0000-0002-9796-8346] and Goranka Blagus Bartolec^[0000-0002-3577-7026]

Institute of Croatian Language and Linguistics, Republike Austrije 16, 10000 Zagreb, Croatia
{imatas, gblagus}@ihjj.hr

Abstract. The goal of this corpus-based research is to see can the complex names with common nouns in their composition be extracted from Croatian hrWaC v2.2 corpus by using regular expressions, i.e. to what extent the capital letter (not the one after the full stop, the exclamation mark or the question mark) can be taken as an indication of a name. Common noun can be used as a regular noun or as a constituent of a complex name, which, on one hand, makes it difficult to tag them automatically, and on the other hand, affects the lexicographic description. With the help of regular expressions, we searched for capitalized common nouns and for sequences in which a capitalized attribute is on the first place and the common noun follows it. After analyzing 1000 examples in each search, we divided results into two groups: names and sequences with an uppercase letter that are not names. Some of the causes of extracting “false” names are technical (e.g. interpunction: separating sentences with paragraph mark (¶), lack of interpunction at the end of sentence; whole parts of text written in upper case...), and some of them lie in the texts crawled for hrWaC, which are not written in accordance with Croatian orthography.

Keywords: Complex Names, Croatian Orthography, Corpus Search.

1 Introduction

Proper names denote a particular person, place, organization, ship, animal, event, or other individual entity. They can be divided into proper nouns (single words like *Europe*, *Mars*...) or complex names (phrases, multiword expressions like *United States of America*, *Faculty of Humanities and Social Sciences*...). Uppercase letter is usually a sign of proper name. In Croatian, uppercase letter stands at the beginning of the sentence and quotation, at the beginning of names (personal names, surnames and nicknames, names of animals, geographical names, names of inhabitants and members of the nation, other names, and possessive adjectives derived from the name) and at the beginning of words of respect and honor. In this work, we have focused on common nouns as a part of complex names. Our goal was to see can the complex names with common nouns in their composition be extracted from Croatian hrWaC

v2.2 corpus¹ by using regular expressions, i.e. to what extent the capital letter (after excluding the capital letter after the full stop, the exclamation mark or the question mark) can be taken as an indication of a name. Common nouns can be used as regular nouns or as constituents of complex names, which, on one hand, makes it difficult to tag them automatically, and on the other hand, if a noun is a part of complex name, “a new word sense must be created in the lexicon” [3: 443]. In lexicography, especially in determining the collocations of a word, it is important to establish the difference between a regular multiword expression and a complex name. Extracting of names from corpus could also be useful help in complementing a *Dictionary of upper and lower initial case in Croatian*, which is being developed in the Institute of Croatian Language and Linguistics.

The rest of the work is structured as follows. In the second chapter the detailed parameters of the corpus search are presented, the results of the search are analyzed in the third chapter, and conclusion is in the fourth chapter.

2 Analysis

Proper nouns are tagged in hrWaC and the search of proper nouns² provided us with results like (first ten sorted by frequency and lemma): *Hrvatska, Zagreb, Ivan, EU, Europa, Zadar, Split, Rijeka, Hrvat, HDZ*. Wider list of 50 examples shows similar results: among them are the names of people (*Ivan, Marija, Ante*), cities (*Zagreb, Zadar, Split*), continents (*Europa* ‘Europe’, *Amerika* ‘America’), states (*Hrvatska* ‘Croatia’, *Srbija* ‘Serbia’), nations (*Hrvat* ‘Croat’, *Srbin* ‘Serb’), football team (*Hajduk*). Two things should be noted: 1) Abbreviations tagged as proper nouns gave us abbreviated names of states and unions, like *EU, BiH* (*Bosna i Hercegovina* ‘Bosnia and Herzegovina’), *SAD* (*Sjedinjene Američke Države* ‘United States of America’), abbreviated names of parties, like *HDZ* (*Hrvatska demokratska zajednica* ‘Croatian Democratic Union’), *SDP* (*Socijaldemokratska partija* ‘Social Democratic Party’), but it also gave us *TV*, which is the abbreviation for the common noun ‘television’. 2) Lemma “Bi”, listed on 35th place by frequency, as a lemma does not exist in Croatian, and in HrWaC it stands for Bog (in examples like: *Taj put započinje krštenjem po kojem možemo **Boga** nazivati Ocem...* ‘This path begins with baptism by which we can call God as Father’) and for *Bin* (*Laden*) (in examples like *U njoj se opisuju aktivnosti Osame **Bin** Ladena...* ‘In it, the activities of Osama Bin Laden are described...’).

Since proper names are mainly correctly tagged and easy to find in HrWaC, we have focused on complex names consisting of a common noun, which form a large number of proper names. According to Croatian orthography [4], uppercase letter comes in one-word names and in multiword names, where the first word is capitalized as well as the word that itself is a name and a possessive adjective derived from the

¹ hrWaC 2.2 is Croatian web corpus by Tomaž Erjavec and Nikola Ljubešić, crawled in 2011 and 2013, cleaned, deduplicated, tagged with the Croatian specification from MULTTEXT-East v5, word sketches created by Nikola Ljubešić. More about HrWaC see in [1] and [2].

² [tag="Np.*"]

name. This is why we searched for capitalized common noun or capitalized word before the common noun. The beginning of the sentence is marked with capital letter, so we have excluded capitalized words after the full stop, the exclamation mark or the question mark. In Search 1 (S1), we looked for capitalized common noun with the help of regular expression `[word!="\.\|\?!\"]``[word="[A-Z].*"&tag="Nc..."]`, and in Search 2 (S2), we looked for the sequence in which a capitalized attribute is on the first place and the common noun follows it (`[word!="\.\|\?!\"]``[word="[A-Z].*"``[tag="Nc..."]`)³. The possibility of one more word between capitalized word (`[word!="\.\|\?!\"]``[word="[A-Z].*"``][tag="Nc..."]`) gave us very general results, so we did not examine sequences with one or more words between the attribute and the noun any further. In each search, we have examined 1000 examples.⁴ To avoid large number of examples from one source, examples were shuffled.

3 Results

Although we searched primarily for complex names that consist of capitalized common noun at the beginning of the name (S1) or of the attribute written in capital letter followed by the common noun (S2), our search (especially S1) gave us also some one-word names (e.g. *Primorje* ‘part of Croatia by the coast). In numbers, we got 539 names and complex names in S1 and 369 complex names in S2. On the other hand, we obtained sequences with the uppercase letter that are not names, 461 of them in S1 and 631 in S2. It seems like a lot, so in 3.2 we presented the reasons for these results.

3.1 Names and complex names

Names and complex names obtained by the search can be divided into two major groups in accordance with rules for upper case writing in Croatian orthography.

Complex names with all capitalized words. One group consists of names whose all parts, according to Croatian orthography, should be written with capital letters except for prepositions and conjunctions. These are:

- personal names and names of cities and villages: although proper nouns are tagged in HrWaC and in search we looked for common nouns, we obtained results with proper names, probably because they were not tagged regularly (e.g. *Osijek* ‘city in Croatia’) or they coincide with some form of common noun, like in: *Posljednjih mjeseci, naime, **Maksim Mrvica** prakticki nema vremena za odmor.* ‘In recent

³ In quoting examples, the whole sentence with the exact result that corresponds to the regular expression is given. The exact result is marked with bold letters in Croatian, but it is not marked in English translation (provided in single quotation marks) since the examples are not translated word-by-word. Mark (S1) means that example is the result of the Search 1, and (S2) that the example is the result of the Search 2.

⁴ In each search, a few sentences which made no sense to us were excluded manually, e.g. (S1) *...shvatio sam da se sve svodi na samo farmanje NPCa ili josh sladje ljudi...* ‘...I understood that everything comes to ?farmanje NPCa? or even better of people...’

- months, Maksim Mrvica has practically no time to rest.’ (*mrvica* means ‘crumb’, and *Mrvica* is a surname); *Dalnji plan je bio stići do Ploča*. ‘The further plan was to reach Ploče.’ (*ploče* means ‘panels’, and *Ploče* is a city)
- names of gods, their periphrastic names, and the names of other supreme religious persons: (S1) *Svaki oblik nijekanja života i samougušivanja strasti putem mučenja tijela Poslanik islama je osudio izrekama kao...* ‘Any form of denial of life and self-denial of passion through torture of the body The Prophet of Islam condemned with the sayings like...’; (S1) *Što ti imaš sa mnom, Isuse, Sine Boga Svevišnjega?* ‘What do you have to do with me, Jesus, the Son of the God Most High?’
 - names of states: (S1) *Velika Britanija evakuirat će u srijedu u Ujedinjene Arapske Emirate svo osoblje veleposlanstva u Teheranu*. ‘The United Kingdom will evacuate all embassy staff in Tehran on Wednesday into the United Arab Emirates.’

Names and complex names with capitalized word at the beginning. The other group obtained in our search are names and complex names in which only the first word is capitalized and also the word that itself is a name and a possessive adjective derived from the name. These are:

- geographic entities: (S2) *Što se događa na Afričkome rogu?* ‘What is going on on the Horn of Africa?’
- institutions, organizations, associations, factories, state and public services, banks, libraries, faculties and colleges, schools, companies, and other objects and their parts: (S2) *Općinu Gornji Kneginec predstavila je Turistička zajednica Općine Gornji Kneginec...* ‘Gornji Kneginec Municipality was presented by the Tourist Board of the Gornji Kneginec Municipality...’
- religious holidays, state holidays and memorials: (S2) *...organizirali su u nedjelju veliku gradsku biciklijadu u povodu obilježavanja Dana grada Zadra...* ‘...on Sunday, they organized a large city bike tour on the occasion of celebrating the Day of the city Zadar...’
- official texts, documents, laws, regulations, agreements: (S1) *...mora biti izravno propisano u Zakonu o medijima*. ‘...it must be directly specified in the Media law.’
- cultural, artistic, political, scientific and other social events, conferences, congresses, festivals, sports competitions...: (S2) *Napokon, Jug dobio Partizana u Beogradu te ušao u finale Lige prvaka*. ‘Finally, Jug beat Partizan in Belgrade and entered the Champions League final.’
- counties, administrative units: (S1) *... predstavljene su aktivnosti Krapinsko-zagorske županije i Grada Zaboka u mjesecu lipnju*. ‘...the activities of Krapina-Zagorje County and the City of Zabok were presented in June.’
- artistic, cultural and social groups: *...poslušnik kraljice pokušava Luciju uvjeriti u ljepote La La Landa, u kojem vječno ratuju Alfe, Bete i Game...* ‘...the queen's servant tries to convince Lucia in the beauty of La La Land, in which Alfas, Betas and Gamas are eternally at war...’

Some complex names are taken from other languages (mostly English): (S2) *Posljednja sesija skupa, nazvana "Buffer zone polities" II: Croatian Principality,*

sadržavala je četiri izlaganja. ‘The last session of the conference, called "Buffer Zone politics" II: Croatian Principality, consisted of four expositions.’

Mistakes. The search also gave us some complex names where the target word is correctly written, but other words are not: (S2) *Kao tajnik Hrvatskog Sabora Kulture obišao sam sva mjesta u Hrvatskoj gdje se njeguje kulturni amaterizam...* ‘As the secretary of the Croatian Parliament of Culture, I have visited all the places in Croatia where cultural amateurism is cultivated...’ Correct writing would be *Hrvatski sabor kulture*.

3.2 Uppercase letter, no name

In S1 we got 461 results which do not match our intention to find names consisting of common noun, and in S2 we got 631 of such results. We established several reasons why our searches extracted sequences with uppercase letters which are not complex names.

Problems with interpunction. There are three typical situations that resulted in uppercase letter which is not a part of complex name.

Paragraph mark (¶). In many examples, paragraph mark is followed by a word with uppercase letter. It probably means that in the original source, before processing it for the corpus, the text was structured in two lines, and the paragraph mark is probably the sign of a new line and, in accordance with that, the sign of a new sentence (e.g. (S1) *Čuvati med od visoke temperature ¶ Mišljenje da kristalizirani med nije prirodan posve je pogrešno.* ‘To keep honey from high temperature ¶ The opinion that crystalized honey is not natural, is completely wrong.’).

Lack of interpunction. The search gave us results where there is no interpunction at the end of the sentence and before the uppercase letter, which is probably, like with paragraph mark, the result of two lines in original binding in one line in corpus: (S2) *Borac za pravdu Uz članove obitelji vijence na njegov grob položili su i saborska zastupnica...* ‘A fighter for justice Along with family members, wreaths on his grave laid also a member of parliament...’

Another punctuation mark. In some examples, uppercase letter comes after punctuation mark other than the full stop, the exclamation mark or the question mark: (S2) *Razmislio je na trenutak i rekao: - **Zelim milijun** dolara svake godine do kraja moga života.* ‘He thought for a moment and said: - I want a million dollars every year to the end of my life.’ Unlike the full stop, the question mark or the exclamation mark, some other marks (like dashes, brackets, and colons) can be followed by uppercase or lowercase letter depending on the context, so this situation could not have been predicted and avoided with more precise search quest. Here we have also counted ordinal numbers, which in Croatian are marked with the full stop mark after numerals. If the

full stop is a part of ordinal number, as examples from HrWaC show, it is not treated as the full stop at the end of sentence (even when it is at the end). Cases like this should have been excluded because the regular expression said no full stop, but they still showed up as results of our search: (S1) *Članak 110. Izvjestitelji sredstava javnog priopćavanja imaju pravo pratiti rad Vijeća...* ‘Article 110th Public relations media reporters shall have the right to monitor the work of the Council...’

Text written in upper case. Sometimes, a whole part of the text can be written in upper case for stylistic reasons and this affects the results of the search, giving us examples which are not complex names like: (S2) *... sada zarađuju tu sumu u dolarima, eurima, funtama A I DUPLO SU MLADI TAKVI SU VANI I NEMA TEORIJE DA SE VRATE...* ‘... now they are earning that sum in dollars, euros, pounds AND THEY ARE TWO TIMES YOUNGER THEY ARE ABROAD AND THERE IS NO THEORY FOR THEM TO COME BACK...’ Abbreviations are also written in upper case, and search extracts them as capitalized words: (S2) *Snimanje filmova u HDTV kvaliteti* ‘Making films in HDTV quality’. Although some abbreviations are tagged as proper names (as pointed out in 2nd chapter), some of them showed up in our search for common nouns although, they should have been tagged as proper names: (S1) *... a da bi toplinska energija onda poskupjela za 25 posto, prenosi HRT* ‘... and then the heat energy would increase by 25 percent, reports HRT (Croatian Television)’.

Attribute + noun. Search 2 showed results in which obtained sequences are not complex names but a sequence of an attribute and a noun. Attribute can be; a) possessive adjective (which in Croatian is written with uppercase letter): (S2) *Božić je blagdan Kristova rođenja.* ‘Christmas is the feast of the Christ’s birth’; b) a pronoun capitalized as a sign of respect: (S2) *Čestitam Vašim čitateljima i Vama na otvaranju ove teme.* ‘I congratulate your readers and you on opening this theme’; c) a noun in function of an attribute (in Croatian, this is considered as a syntactic influence of English): (S2) *... najljepše od svega je da su sva tri Ferrari motora crkla u utrci.* ‘The most beautiful of all is that all three Ferrari engines crashed in the race.’

Influence of foreign languages. Some sequences can not be treated as complex names since capitalized word in them is taken from a foreign language (mostly English): (S2) *Pa program je predvidio Park-Ride sustav...* ‘Well, the program anticipated Park-Ride system...’

Mistakes. Some examples given by the search are capitalized and correctly extracted from HrWaC, but they are not in accordance with Croatian orthography – they are the result of author’s ignorance or, possibly, stylistic intention of giving the capitalized word a greater meaning: (S2) *...započinje suradnja s Aikido klubom u Mantovi.* ‘...the co-operation with the aikido club in Mantova begins’; (S1) *...lako bi se moglo dogoditi da Rusija postane posljednje utočište Bijele Rase.* ‘...it may well be that Russia

becomes the last retreat of the White Race.’ In Croatian orthography, there is no reason to write *white race* or *aikido* with uppercase letters.

4 Conclusion

In our research, we tried to extract complex names that consist of a common noun in Croatian corpus HrWaC. Since proper names form a specific group of words and they are mainly properly tagged in HrWaC, we have focused on common nouns, which can be used as regular nouns or they can be a part of complex names. Establishing the difference between complex names and the regular use of common nouns is important for lexicographic description. Names received in such a way could also serve as an additional extraction tool of entries for a *Dictionary of upper and lower initial case in Croatian*, which is being developed in the Institute of Croatian Language and Linguistics. In our search we got 539 names of 1000 examples in S1 (when we looked for capitalized common noun at the beginning of the name) and 369 of 1000 examples complex names in S2 (when we looked for sequence in which a capitalized attribute is on the first place and the common noun follows it). Some of the results are proper names, which should have been tagged as proper nouns and should not have come up as a result of our search for common nouns (probably they were not tagged regularly or they coincide with some form of common noun). We also obtained sequences with the uppercase letter that are not complex names, 461 of them in S1 and 631 in S2. One of the reasons for such a large number of “false” names can be linked to processing texts for corpus, namely inconsistencies of interpunction (paragraph mark, no interpunction...).

This search also emphasized the aspect of liability of sources in corpus. The texts collected for corpus come from public domain, they are often unedited and with many orthographic and grammatical errors. On the other hand, propositions for writing uppercase and lowercase letters in Croatian are very detailed and sometimes ask for extralinguistic knowledge. Besides, some rules have changed over the years (when parts of the earth, like *istok* ‘East’, refer to people and culture, they had to be written with uppercase letter – *Istok*, but according to new Croatian Orthography [4], they should be written in lowercase – *istok*). In order for texts from public domain to be a reliable source for linguistic purposes, their authors should be in step with orthographic changes and familiar with some general facts, which often is not the case.

Although it seems like search gave as many negative results, obtained names can contribute to the work on *Dictionary of upper and lower initial case in Croatian*, in checking the existing list and its enrichment, since some of obtained names were not previously included in it.

References

1. Ljubešić, N., Erjavec, T.: hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. *Text, Speech and Dialogue* 2011, 395–402, <http://nlp.ffzg.hr/data/publications/nljubesi/ljubesci11-hrwac.pdf> (2011).
2. Ljubešić, N., Klubička, F.: {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In: Bildhauer, F., Schäfer, R. (eds.) *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pp. 29–35. Association for Computational Linguistics, Gothenburg (2014).
3. Coates-Stephens, S.: *The Analysis and Acquisition of Proper Names for the Understanding of Free Text*, *Computers and the Humanities*, 26 (5/6), 441–456 (1992).
4. Jozić, Ž., Blagus Bartolec, G., Hudeček, L., Lewis, K., Mihaljević, M., Ramadanović, E., Birtić, M., Budja, J., Kovačević, B., Matas Ivanković, I., Milković, A., Miloš, I., Stojanov, T., Štrkalj Despot, K.: *Hrvatski pravopis*. Institut za hrvatski jezik i jezikoslovlje, Zagreb (2013).

Fixed Phrases in Language of International Law: A Problem of Translating Latin Formulaic Expressions into Farsi

Seyed Mohammad Hossein Mirzadeh

Islamic Azad University, Iran
interpret3@gmail.com

Abstract. For the past twenty years, “phraseology” has been considered a very important topic of study for various specialized languages. The linguistic view that used to see phraseology such as “idiom researches and lexicography classifying various kinds of idiomatic expressions” has changed meaningfully. Nowadays, thanks to these changes, the new view is focused on identifying and classifying phraseology as well as applying them to research in theory. That is why we would do well to try to define new horizons of phraseology in different specialized languages. The language of interest here is the prescriptive and descriptive language of international law instruments. We should consider this language as the normative language of judges, legislators, courts and international lawyers. These practitioners – who use specific types of phraseology and stable linguistic structures – should perhaps adhere to the use of a professional language that conforms to recognized standards of normative rules. This paper, therefore, tries to define the main relations between phraseology studies and IL Latin expressions and their systematic-semantic equivalences in languages with different roots like Farsi.

Keywords: Phraseology, International Law Instruments, Latin Phrases and Expressions

1 Introduction

Despite all efforts, yet there are too important factors which should be studied about language of International Law. If we accept or not, the IL also enjoys a specific form of language which is made by a specific class of users in the societies. The users of this type of language are judges, lawyers, interpreters, courts, UN commissions and so forth which have a primordial role for defining the correct way to describe denotation-connotation of legal norms and rules. However, some important questions which arises here are that if we accept that language of IL has its own characteristics, how can we analyze and perceive it? Is it possible just to understand this type of language only by international legal knowledge? What about the other languages? Accepting that not all IL language users and makers are English or French speakers, how can the courts and

international lawyers conclude documents that the other language users like Farsi or Arabic understand them well?

That's why we believe that we have to incorporate new data bases from the other sciences like linguistics, cognitive linguistics, and pragmatics and so forth into IL knowledge in order to be able to coding-decoding messages, statements and rules in international law. In this respect, one of the interdisciplinary studies which can help scholars to analyze correctly the sentences and expressions in international law could be phraseology or study of formulaic expressions. Enquiries about this type of knowledge show that it is necessary to evaluate IL formulaic expressions from this viewpoint specifically because of the existence of a huge number of Latin formulated expressions in that language. We believe that the study of phraseology and incorporating it into IL language, can help IL scholars to understand the real formulation-translation of these expressions used in international law.

We seriously believe that cognitive- corpus based linguistics could help International Lawyers to comprehend the main idea of expressions used in International Law Language part of which have been derived from Latin. In this aspect, a descriptive methodology can help reader or interlocutor of this paper to understand both, the main idea of this paper and know how to analyze this phenomena in his/her native language which in this specific case is Farsi since one of the biggest existing problems for Farsi speakers of IL texts is the huge vague or etymological- cognitive differences between Farsi and Latin.

For all above mentioned reasons, the aim of this paper is to evaluate the impact of some phraseological studies and enquiries in forming- translating International Law Latin formulaic expressions and the objective equivalences in the other languages like Farsi, a language that has nothing to do with Latin languages in any linguistic and extra linguistic aspect, responding to this important question that have systematic differences of Latin Legal language to do with semantic phase when translating IL Latin expression into Farsi or not? That is why we will try to show, at first hand, some useful IL Latin expressions and at second one, evaluate their systematic- translational equivalences in Farsi.

2 The Latin phraseological Units in IL texts

It seems to be a challenge analyzing and speaking about phraseology because the interlocutor of learner probably encounter it really confusing with the other aspects of cognitive linguistics, applied linguistics and discourse analysis [16]. According to Cowie, there is a lack of standardized terminology [6]. That is to say, there exist terminologies like collocations or idioms which referred for decades to the same idea of phraseological units. That's why we could declare that for having an uncertain terminology about the phenomena, so the final decision about what the identity is comes really difficult.

We believe that these stable and fixed phrases in IL have their own grammatical, discursive and semantic characteristics. For instance, when the ICJ or international lawyers use expressions of phraseological units like: *utti posidetti juris*, *jus gentium*, *ex aequo et bonno*, *jus cogens*, *erga omnes* and etc. there is going to emerge an idea of figurative language in particular interlocutors regarding the content of the discourse and

the usage of these types of legal expressions. “*jus cogens*”, being a type of phraseological fixed and stable unit which could be studied from different points of view, should be interpreted in a narrative way in order to connate in the strict manner :”promtury or imperative norms” or قواعد امره. An international judge or lawyer by this unit, has a clear intention which is to call the attention of specialized interlocutor: 1- this comes from a Latin legal system which has been used during years and it is working like an international norm which is applicable for all of the IL subjects 2- in a grammatical approach, it is formed by 2 parts (monems or signifiers) which connate an action or reaction: “these norms or *Jus* are to be considered important and binding that the violation of them could not be presumed.”

3 Translating some IL Latin formulaic units into Farsi

Translation of specialized texts, especially legal terms, is not as easy as some people think; it not only requires stylistic competence in the language, but it also requires knowledge of the inner meaning and connotation of words of the legal systems. That is why Farsi translators and then, interpreters of legal terms especially Latin ones, should be aware of both stylistic and phraseological features of these terms and semantic and connotative characteristics in order to get the right sense. For instance, there is no any literal meaning for *Ex aequo et bonno* term (phraseological unit) in Farsi but the translators are obliged to explain It in other words which could lead to descend the principal denotation of the term. For instance, the word *Jus* when is used with *ad bellum* has a specific connotation and equivalence in Farsi and when it is used with *Cogens*, has a stricter meaning regarding imperative norms of IL. The other useful example is when we use *Jus* with *Gentium*, when the meaning differ from the previous ones and it is transferred like “rights and demands of people”. The below Examples Show that it is almost impossible to preserve grammatical density, phraseological characteristics of Latin language and semantic properties while translating them into Farsi:

<i>Utti Possidetti Juris</i> : تبدیل مرزهای حین استعمار به مرزهای بعد از استقلال		
Eng.: As You Possess		
<i>Ex aequo et bonno</i> :	اصل صلاح و ثوابدید	Eng.: According to the Right and Good
<i>Lex fori</i> :	قانون مقر دادگاه	Eng.: the Local Law
<i>Jus Cogens</i> :	قواعد امره	Eng.: Imperative Norms
<i>Jus Gentium</i> :	حقوق مردمان	Eng.: Law of Nations

4 Concluding remarks

The objective of this paper was to propose the design a systematized view for encoding and describing phraseological information in IL language and to create a new aspect of fixed legal terms especially those which come from Latin.

Difference in language systems could lead differences in translational phase. For instance, *Jus* is sometimes translated like حقوق which is “law” in English and sometimes is translated like قواعد which is “norms or rules” in English. The systematic and semantic nature of *Jus* in Latin is the same but while translating it into Farsi, it changes from context to another context. That’s why it can be inferred that the “contextual criteria” of the target languages like Farsi has also to do with the exact meaning of IL Latin formulaic expressions.

As studied above, the systematic- etymologic differences between languages like Farsi and Latin especially in specialized areas like IL could result in different ways of translating expressions. Another clear example in this aspect is *Utti Possidetti Juris* which is a 3 words expression while its literal translation in Farsi does not exist so the Farsi lawyers translate it like: تبدیل مرزهای حین استعمار به مرزهای پس از استقلال. So it is translated in a completely contextual manner in order to show the exact connotative meaning.

The underlying idea is that when translating phraseological IL units from Latin languages into Farsi, it is particularly difficult to adequately convey an equivalence that connotes the same root, lexeme and sense of terms. Therefore, to understand the main ideas and connotations of these phraseological units, the Farsi interlocutors should use a triangular approach: 1) perception of denotative-connotative meaning, 2) analysis, and finally 3) expression of the main idea of the terms if there is not any literal equivalence, especially for units like: **ex aequo et bonno, utti posidetti juris** and the expressions with more than one word, signifier and connotation.

References

1. Cowie, A.P. The treatment of collocations and idioms in learners dictionaries, applied linguistics 2 (3): 223-235 (1981)
2. Cowie, A.P. “Introduction”. In Cowie, A.P. Phraseology. Theory, Analysis, and Applications, 1-20, Oxford: Clarendon Press (1998)
3. Moon, R. Textual aspects of fixed expressions in learner’s dictionaries. In Vocabulary and Applied Linguistics, P.J.L. Arnaud & Bejoint (eds), 13-27. London: Macmillan (1992)

On the Impact of (Il)literacy on L2 Italian Acquisition of Unaccompanied Foreign Minors

Castrenze Nigrelli

Università degli Studi di Palermo, viale delle Scienze, 90128, Palermo, Italy
castrenze.nigrelli@unipa.it

Abstract. The aim of the paper is to analyse the interlanguage of L2 Italian learners with the same L1 but different levels of education. The learners belong to the “unaccompanied foreign minors” category, whose linguistic profile is characterised by the frequent coexistence of a multilingual ability and a very low, or zero, level of education. Focusing on the acquisition of verb inflectional morphology and on phraseological units as well, the comparison of learners’ varieties aims to show that several differences depend on the only parameter that differentiates them, namely literacy vs. illiteracy in L1.

Keywords: Second Language Acquisition, L2 Italian, Unaccompanied Foreign Minors.

1 Introduction

1.1 Unaccompanied Foreign Minors: Multilingualism and Illiteracy

Minors who are not Italian citizens, nor asylum seekers, and are in Italy without adults who have legal responsibility belong to the category of “unaccompanied foreign minors” (henceforth UFM). The city of Palermo is a multilingual environment that hosts a considerable number of migrants from different places, and UFM make Italian classrooms linguistically mixed. UFM are a heterogeneous and peculiar socio-linguistic category (see [1]). They belong to very different cultures, and come from different multilingual areas (especially from Africa and Asia). UFM are characterised by an extraordinary multilingual ability and by a low level of education. Often, UFM have, in addition to their own L1, some competence in one or more languages spoken in the place of origin, as well as other languages they came into contact with during the long stops of their travels to Italy. At the same time, their level of education is mostly low, or very low (less than 5 years of school), or even non-existent. They are, ultimately, both multilingual and functionally or completely illiterate.

The above specificity makes these learners’ profile peculiar and unprecedented. In the absence of solid previous schooling, competence in languages (including L1) is very often limited to oral competence alone, and this has significant repercussions on second-language acquisition and, therefore, on social inclusion (see [2]-[3]).

UFMs have a certain ability to manage oral texts, but a low familiarity with written texts. This has inevitable consequences in terms of meta-textual and meta-linguistic reflection, which affect the process of acquiring an L2. The problem is even more serious for the completely illiterates UFMs. A competence, even partial, in writing and reading has an important impact on the mental processes of the individual (see [4]-[5]). Someone who knows how to write usually calls to mind the graphic images of the words and this makes him more able not only to memorize, but also to analyse, since they are more used to breaking down the phonic continuum into words (and the words, in turn, into smaller units). His brain is, therefore, more suited to metalinguistic reflection and to processes of abstraction, generalization and modelling, while that of the illiterate is anchored to mechanisms that are purely pragmatic and semantic (see [6]).

To evaluate the impact of (il)literacy on the acquisition of an L2, with particular reference to verbs and verbal categories in Italian, the paper analyses the interlanguage (see [7]) of two analogous UFM learners in the sociolinguistic profile, but, one is completely illiterate in his own L1, while the other has a low (but not very low) level of schooling. The corpus is based on their oral productions (interviews).

2 Methodology

2.1 Sample Choice and Learners' Profile

The study aims to verify to what extent any divergences in the learners' varieties are to be attributed to a given parameter, i.e. (il)literacy in L1, which it is therefore sought to isolate, by a suitable sample choice. The learners are in fact both UFMs who arrived in Italy a few months ago initially originating from Gambia, with Mandinka as L1 and English as L2 (official language in Gambia). Both learners are at the end of a course in L2 Italian for absolute beginners (about 40 hours). Sociolinguistically, the element of distinction between the two learners is (il)literacy in L1. The learner S.D. has been in Italy for 3 months, studied in Gambia for 9 years and is, therefore, literate in his L1 despite having a low level of schooling. The learner M.S. has been in Italy for 7 months, never went to school in his country of origin and he was illiterate when arrived in Italy. As shown by the interview, S.D. is more widely multilingual, since he seems to know Wolof as L2 and has a certain competence of Joola and Fula.

Mandinka is a language of the Mande group (Niger-Congo sub-family; Niger-Kordofanian family), spoken in a vast area of West Africa. In Mandinka, temporal, aspectual, and modal information is conveyed by affixes. Mandinka presents an unmarked syntactic order of the SOV type, and a variable order modifier/name following the scheme: Gen/N, Poss/N, N/A, DimN/NDim, NPlur (for further details, see [8]).

2.2 Method of Data Collection and Analysis

Data consists of oral texts produced by the two above learners through semi-structured face-to-face interviews (see Table 1, Appendix). The conversation started

from a list of questions prepared by the native interviewer and concerning aspects of the informant's personal and everyday life as well as his experiences in linguistic matters. To stimulate the informant to use certain linguistic structures and to better investigate the acquisition stage of the verbal categories, several specific questions were also chosen by the interviewer.

From the perspective of conversation analysis, the native-non-native interaction is considered as prototypically asymmetric [9]. The interviews were audio-video recorded and subsequently transcribed in accordance with the transcription method of conversation analysis. However, the presence of recording system has further inhibited informants. This contributed to raise Krashen's affective filter, partially compromising the spontaneity of the performance [10]. Furthermore, the way the interviewer deals with the situation could play a role due to the linguistic and emotional fragility of the informants. In some cases, the interviewer's reformulations resolves the interviewee's impasse too quickly and, if on the one hand this makes him feel more at ease, on the other hand it prevents the opportunity to find more data. Through a careful examination of learner's oral productions, the analysis aims to establish both the stage of acquisition they are in and the role played by (il)literacy, with reference to the acquisition of the verb inflectional morphology and to phraseology.

3 Analysis

3.1 Analysis of the Variety of M.S.

The presence of different verbal forms in M.S. learning variety shows that he somehow distinguishes the verb as an entity. Most verbs occur in a "base" form (see [11]): most of the time the learner overextends the use of the third person of the present indicative, other times the second person (probably basing on that of the input in the question), and in others the infinitive (e.g. *Io comincia; io vai a Mondello; non studiare Gambia*). Among the base forms, the use of the infinitive can usually depend on two factors: either the input to which the learner is most frequently exposed, especially if the natives with whom he is in contact use ultra-simplified varieties (foreigner talk), or the particular context of non-actuality, non-reality of the event (such as in this case). In the context of interaction, the phrase "non studiare Gambia" actually means "non ho studiato in Gambia" ("I did not study in Gambia") and therefore refers to an event that is perfective and non-actual: through overextension, the functional use of the infinitive form of the verb helps the learner to express a distinction which is aspectual (i.e. perfective vs. imperfective). The use of person markers in the present indicative is sporadic and the learner is still completely uncertain, as evidenced by the presence of several reformulations (e.g. "a mare: *giaoca: | gioco football*" or in "io non lo so", then reformulated as "non lo sa"). The correct use in "non lo so" could also depend on a memorization of said formula – generally, occurring frequently in the input – of which the learner is not fully aware. There are also other examples of formulas, e.g. "come stai?" or the request for help in the quite uncertain phrase "come mi tiam que?" (which would be "come si chiama questo?" "How do you say this?"). In several cases the learner simplifies the statement by omitting the copula, as in "Pa-

lermo buono, tutti buono, Mondello bravissimo”. In other cases, he tries to make up for it through extra-linguistic elements (gestures and mimicry), like when he says Casa nodding with his head to indicate “staying”, or through leaning on his L2 English, like when he says “to:: //” miming the action of swimming, which references back to the English infinitive form with the particle to.

The only isolated example of an aspectual marker is in the past participle of *capire* “to understand”, e.g. “io no capito italiano”. According to acquisition stages, said past participle marker -to is the first aspectual marker learnt, and, in turn, [auxiliary + past participle] constitutes an Italian tense conveying perfective aspect (i.e. *passato prossimo*), which is learnt successively. Furthermore, in the last example, the verb in question (i.e. *capire*) also represents a prototypical case where learners usually start to use the aspectual marker -to, due to actional features (*Aktionsart*) of the verb, namely telicity, which makes telic verbs semantically more compatible with expressing perfective aspect (see [12]-[13]).

3.2 Analysis of the Interlanguage of S.D.

The analysis of S.D. learning variety can identify some features common to the variety of M.S, despite there being several significant differences. Also S.D. seems to distinguish verbal entities within the sentence, although he is still lacking in terms of inflectional morphology. Even in his production there are several occurrences in which the verb occurs in the base form. However, the overextension concerns only the infinitive form, except for two occurrences of the third person, which are actually reformulated immediately (“io non è” > “non ho”). In fact, in “io / tre lingue / parlare tre lingue”, the use of the infinitive form overextends into present indicative contexts, while in the other three occurrences (“io andare scuola, giocare calcio, scuola giocare e basta”) seems to make up for the lack of the imperfect, in its value of imperfective past, to express habitual actions in the past (on verbal aspect in Italian, see, among others [14]). Also S.D. omits the copula, as in “Palermo:: bella, tutti:: bravo”, although less often than M.S.

S.D. is more at ease than M.S. in using of the inflectional markers of first person. In fact, the first person of the indicative present fully represents the most frequent verbal form in its variety, as it is evidenced by several examples: “io sono gambiano, ho novo ani, io ho / fratello: piccolo”, “mi piace, io mi piace capoeira, mi facio capoeira”.

3.3 Comparison of Learning Varieties

The characteristics of the learners’ interlanguages analysed above allow to consider their varieties as certainly basic varieties, as it was reasonable to predict. According to Klein & Perdue (see [11]), both learners are in fact in that non-inflectional initial phase of the second language acquisition process. Both seems to be aware of the entity of verb and distinguish it from noun. In some cases, they also seem to have some awareness of the argumental structure of the verbs they use. Regarding the expression of the verbal categories of tense and aspect, the functional role of both is made up exclusive-

ly for the indicative present or the infinitive form: both learners in fact present the typical base forms, overextending the infinitive or the second or third person of the indicative present to express the past as well as a perfective aspect. The only form of morphologically overt aspectual opposition is isolated: this is the past participle of *capire*, which M.S. uses only once. Both learners do not have the means to express the future and implement avoidance strategies (see questions that close interviews). They tend to omit the copula and do not yet have full awareness of the markers of first-person of the present indicative, although with significant differences between the two: S.D. has in fact several occurrences and seems clearly more at ease.

Available data already allows noting a certain gap between the performances and to consider that, at least to a certain extent, depending on the (il)literacy in L1 parameter. The schooled learner (S.D.), despite being in Italy for less time (3 months vs. 7) shows a greater skill in oral interaction, a greater communicative efficacy, greater accuracy and fluency. He understands the questions without the difficulties shown by M.S., he also responds more promptly, he takes fewer breaks, and without requesting any help. Moreover, he never produces bilingual statements, nor does it rely on English (second language for both learners) to make up for lexical or syntactical gaps, as M.S. does. Also, the dependence on the context and on common knowledge is minimal, if not null, with respect to the frequent use of the gestures or lexical strategies of M.S. Furthermore, in the acquisition of verbal morphology, a considerable gap - for accuracy and number of occurrences - is shown in the use of the markers of first-person that M.S. uses well only in some formulaic statements like *non lo so*. Finally, the use of the basic forms of S.D. (only the infinitive) is also limited to the narrative context that would have required the expression of the past (with a perfective and imperfective aspect).

4 Conclusions

From the strategies implemented by the two learners in the interviews and the characteristics of their interlanguage, it was possible to place them within the basic acquisition phase. Although the sociolinguistic profile and the learning context of the learners are close similar, the differences between their varieties are not without importance, though mostly blurred. To a certain extent, the differences found in the analysis of their productions seem to depend on that S.D. has a certain level of education and proficiency in writing skills, while M.S. is completely illiterate. Despite the same length of learning course (about 40 for both), the schooled learner had an exposure to the target language less than illiterate one (3 vs. 7 months). Besides his larger multilingual competence, the positive differences in the performance of the schooled learner can be reasonably ascribed to literacy, which substantially distinguished the sociolinguistic profiles.

5 Annex

Table 1. Interviews of M.S. and S.D.

I.1: <i>Ciao M.!</i>	M.S.2: <i>Ciao</i>	I.1: <i>Ciao S.!</i>	S.D.2: <i>Ciao</i>
I.3: <i>Ciao come stai?</i>	M.S.4: <i>Bene e tu?</i>	I.3: <i>Come stai?</i>	S.D.4: <i>Bene</i>
I.5: <i>Io tutto bene grazie</i>	M.S.6: <i>Sì</i>	I.5: <i>Allora / ehm:: da quanto tempo sei a Palermo?</i>	S.D.6: <i>Ehm:: due: tresetres+ [due? / ehm: a: trese</i>
I.7: <i>E:: ti faccio qualche domanda: / ehm:: da quanto tempo sei a Palermo?</i>	M.S.8: <i>Io?</i>	I.7: <i>Tre mesi?</i>	S.D.8: <i>Tre mesi sì ((nodding his head))</i>
I.9: <i>Sì</i>	M.S.10: <i>Forsa:: due tre ann+ ((making the number three with his fingers))</i>	I.9: <i>Ok / ehm:: ti piace questa città?</i>	S.D.10: <i>Sì sì mi piace</i>
I.11: <i>Tre an[ni]</i>	M.S.12: <i>[Sì] tre anni ((making the number three with his fingers)) ((nodding his head))</i>	I.11: <i>Perché?</i>	S.D.12: <i>E:: perché: Palermo:: bella / sì:: // e:: xx (smiling) se: e: xx // bellissimo</i>
I.13: <i>Ok // ehm:: ti piace Palermo?</i>	M.S.14: <i>Sì ((in a low voice)) ((nodding his head))</i>	I.13: <i>Uhm</i>	S.D.14: <i>Ehm:: buono</i>
I.15: <i>Perché? Che cosa ti piace di Palermo?</i>	M.S.16: <i>Palermo perché: Palermo buono / io / comincia / buono // tutti buono ((nodding his head))</i>	I.15: <i>Sì</i>	S.D.16: <i>Ok // tuti:: bravo</i>
I.17: <i>Mhm? Ti piacciono le persone?</i>	M.S.18: <i>Sì ((nodding his head)) ((in a low voice))</i>	I.17: <i>Ti piacciono le persone di questa città?</i>	S.D.18: <i>Sì sì me piace ((nodding his head))</i>
I.19: <i>a Palermo?</i>	M.S.20: <i>Sì ((nodding his head)) ((in a low voice))</i>	I.19: <i>Uhm ok / ehm:: e:: di quale nazionalità sei tu?</i>	S.D.20: <i>Io Gambia / Ga+ // io sono gambiano</i>
I.21: <i>E la città? Che cosa ti piace della città?</i>	M.S.22: <i>La sità // Σ no: non lo so ((shaking his head))</i>	I.21: <i>Ok ehm::: e che lingua:: parlavi in Gambia?</i>	S.D.22: <i>Mandinka</i>
I.23: <i>Mhm ho capito / e:: d da quale paese provieni? Qual è il tuo paese di origine?</i>	M.S.24: <i>Io? xx</i>	I.23: <i>Ok / a casa?</i>	S.D.24: <i>A casa sì</i>
I.25: <i>Gambia?</i>	M.S.26: <i>Sì Gambia Gambia</i>	I.25: <i>Ok e altre lingue? Che altre lingue conosci?</i>	S.D.26: <i>Cosa?</i>
I.27: <i>Quindi sei gambiano</i>	M.S.28: <i>Sì gambian</i>	I.27: <i>Che altre lingue conosci? Conosci altre lingue?</i>	S.D.28: <i>Io?</i>
I.29: <i>Ok ok // e::: in Gambia quale lingua parlavi a casa?</i>	M.S.30: <i>Mandinka.</i>	I.29: <i>Sì</i>	S.D.30: <i>Io Gambio:</i>
I.31: <i>Mandinka.</i>	M.S.32: <i>Sì ((nodding his head))</i>	I.31: <i>A parte mandinka, conosci altre lingue?</i>	S.D.32: <i>Mandinka, wolof =</i>
I.33: <i>Ok / e conosci anche altre lingue?</i>	M.S.34: <i>Ingl // scuola inglis+</i>	I.33: <i>Uhm</i>	S.D.34: <i>= Inglis =</i>
I.35: <i>Ah. Hai studiato a scuola?</i>	M.S.36: <i>Gambia ((making the "out" sign with his hand)) In Gambia?</i>	I.35: <i>Ah ah</i>	S.D.36: <i>= Joula =</i>
I.37: <i>In Gambia?</i>	M.S.38: <i>No no no ((shaking his head))</i>	I.37: <i>Sì</i>	S.D.38: <i>= Fula</i>
I.39: <i>Ah / [ok]</i>	M.S.40: <i>[non studiare] Gambia ((in a low voice))</i>	I.39: <i>Mi::</i>	S.D.40: <i>Cingue lingue</i>
I.41: <i>Non s+ non sei andato a scuola?</i>	M.S.42: <i>No no no ((shaking his head))</i>	I.41: <i>Tantissime lingue!</i>	S.D.42: <i>No io / tre lingue / parlare tre lingue bravo!</i>
I.43: <i>Ok / ehm: e invece::mhm:: da quanto tempo studi l'italiano?</i>	M.S.44: <i>Taliano ((nodding his head))</i>	I.43: <i>Ok ne parli tre</i>	S.D.44: <i>Sì</i>
I.45: <i>L'hai studiato qua?</i>	M.S.46: <i>Sì ((nodding his head))</i>	I.45: <i>Però / ne [conosci cinque]</i>	S.D.46: <i>[Cingue lingue] sì</i>
I.47: <i>o anche [prima]?</i>	M.S.48: <i>[No] qua/ Palermo qua ((making the "stay" sign with his head))</i>	I.47: <i>Ok / quali lingue parli?</i>	S.D.48: <i>Io: / mandinka =</i>
I.49: <i>Solo a Palermo?</i>	M.S.50: <i>Sì</i>	I.49: <i>Sì</i>	S.D.50: <i>= Wolof / inglese</i>
I.51: <i>Ok / e per quanto tempo? Quanti mesi?</i>	M.S.52: <i>In Palermo? // No [no] ((shaking his head))</i>	I.51: <i>Ok =</i>	S.D.52: <i>Sì ((a voce bassa))</i>
I.53: <i>[No] l'italiano: il corso di italiano</i>	M.S.54: <i>[Sì]</i>	I.53: <i>= Ok / ehm::: e cosa ti</i>	S.D.54: <i>Che cosa?</i>
I.55: <i>[Per quanti] mesi hai studiato?</i>	M.S.56: <i>Palermo qua?</i>		
I.57: <i>Sì</i>	M.S.58: <i>Nsa: // Σ no no: ((making a sign</i>		

1.59: Non:: non ti ricor[di:]	with his fingers)) ((shaking his head))	piace fare nel tempo libero?	
1.61: [ok] ehm:: volevo chiederti:: e nel tempo libero cosa ti piace fare?	M.S.60: [No] non [ti ri+] ((smiling)) M.S.62: /// Io no: /// ((making a sign with his fingers)) ((smiling)) capito itali[ano] ((making a sign with his fingers)) =	1.55: Nel tempo libero: / quando non sei a scuola	S.D.56: Io? Ehm::: ho novo ani novo ano
1.63: [ok]	M.S.64: = Bela [poco]	1.57: No / ti chiedo: e: quando non vieni a scuo:la::... = e	S.D.58: Inglese
1.65: [Si si] ti rifaccio la domanda?	M.S.66: ((nodding his head))	1.59: No no: e:: cosa cosa fai durante la giornata? cosa ti piace fare?	S.D.60: E::
1.67: Cosa ti piace quando sei libero? Quando hai tempo libero?	M.S.68: ncas+	1.61: Ehm::: vai a gioca:re:: o::...	S.D.62: Si si si ((smiling)) io: mi piace capoeira
1.69: E:: a casa:: o fuori::	M.S.70: Si [cas+]	1.63: Ah	S.D.64: Si si mi facio / mi facio capoeira
1.71: Quando non sei a scuola / cosa fai?	M.S.72: Casa ((making the "stay" sign with his head))	1.65: Ah bello!	S.D.66: Si
1.73: Sei Stai a casa?	M.S.74: ((nodding his head))	1.67: E:: e in Gambia: sei stato a scuola?	S.D.68: Si
1.75: Ok / ehm::: vai fuori anche? Oppure stai sempre a casa?	M.S.76: Fori / quan+ fori io vai a Mondello	1.69: Quanti:: quanti anni hai studiato?	S.D.70: xx nov novo ani ano
1.77: [Mhm:]	M.S.78: [Si] Mondello bravissimo! ((smiling))	1.71: Per nove anni?	S.D.72: Si
1.79: [Ah]	M.S.80: Si ((smiling))	1.73: Ok ehm::: raccontami: che cosa facevi in Gambia?	S.D.74: Io?
1.81: Bello / ti piace?	M.S.82: Si	1.75: Si / nella vita	S.D.76: Io andare scuola =
1.83: E che cosa fai a Mondello?	M.S.84: Perché: // caldissimo xx va a Mondello: =	1.77: Si	S.D.78: = E::: giocare calcio =
1.85: Si	M.S.86: = to: // ((miming the action of swimming)) [-come mi tiamque? -]	1.79: Si	S.D.80: = E che:: io non è non è: ho:: e: la+ li+ che cosa-lavoro in Gambia no: / scuola:: giocare: e basta
1.87: [vai a nuotare?] A mare?	M.S.88: A mare si	1.81: Ho capito =	S.D.82: Si
1.89: Vai a mare / [ho capito]	M.S.90: [Si] ((nodding his head)) / xx a mare: gioca: gioco football	1.83: = Ho capito / ehm::: quando sarai grande quando sarai più [grande?] =	S.D.84: [Io:]
1.91: Ah! [E::]	M.S.92: [Si]	1.85: = Cosa farai?	S.D.86: Io: ho / fratello: piccolo uno / grandi uno
1.93: Con::: giochi come?	M.S.94: Qua ((making the "out" sign with his hand)) // Palermo qua // Io comincia: /// // Io//	1.87: Si / E: vuoi andare / da loro?	S.D.88: Si si
1.95: Qua / vicino alla stazione?	M.S.96: Firi ((making a circle with his hands))	1.89: Ok / ok / S. e:: ti ringrazio / ciao.	S.D.90: Ciao / grasi
1.97: Al foro?	M.S.98: Foro si ((nodding his head))		
1.99: Foro italico?	M.S.100: Si		
1.101: Ah: e come giocate?	M.S.102: Tutti comi		
1.103: No: che / che gioco? A che gioco giocate?	M.S.104: Io? ((indicating himself))		
1.105: Si	M.S.106: *Goalkeeper ((miming a goalkeeper))		
1.107: Ah::!	M.S.108: [Si]		
1.109: [E:::] goalkeeper!	M.S.110: si [goalkeep+] ((smiling))		
1.111: [Fai] fai quindi giocate a pallone, a calcio?	M.S.112: Si si ((smiling)) ((nodding his head))		
1.113: E tu fai il portiere:	M.S.114: Si: ((smiling))		
1.115: Ok ok // Ehm::: un'altra cosa / e::: quando sarai più grande? =	M.S.116: Mhm		
1.117: = Cosa vorrai fare?	M.S.118: ///		

L.119: <i>Fra un anno: fra due anni: // che cosa ti piacerebbe fare? Cosa vorrai fare?</i>	M.S.120: <i>Sa due a+</i>
L.121: <i>Fra due anni:</i>	M.S.122: <i>Mhm</i>
L.123: <i>Quando sarai più grande: /</i>	M.S.124: <i>Mhm</i>
L.125: <i>Cosa::: ti piacerebbe fare?</i>	M.S.126: <i>// No io non lo so ((shaking his head))</i>
L.127: <i>Non lo sai?</i>	M.S.128: <i>No no non lo sa</i>
L.129: <i>Ok / va bene / ok M. grazie</i>	M.S.130: <i>Prego</i>

References

1. Barone, L.: L'accoglienza dei minori stranieri non accompagnati. Tra norma giuridica e agire sociale. Key Editore (2016).
2. Demetrio, D., Moroni, F.: Alfabetizzazione degli adulti: teoria, programmazione, metodi. Editrice sindacale italiana, Roma (1980).
3. Watson, J. A.: Cautionary tales of LESLLA Students in the High School Classroom. In: Vinogradov, P., Bigelow, M. (eds.), *Low Educated Second Language and Literacy Acquisition - Proceedings of the 7th Symposium*, Minneapolis, Minnesota, USA, pp. 203–234. (2011).
4. Luria, A. R.: *Cognitive development: Its cultural and social foundations*. Harvard University Press, Cambridge (1976).
5. Ong, W.: *Orality and Literacy*. Methuen, London-New York (1982).
6. Minuz, F.: *Italiano L2 e alfabetizzazione in età adulta*. Carocci, Roma (2005).
7. Selinker, L.: *Interlanguage*. *International Review of Applied Linguistics in Language Teaching* 10(3), 209–231 (1972).
8. Banfi, E., Grandi, N. (eds.): *Le lingue extraeuropee: Asia e Africa*. Carocci, Roma (2008).
9. Orletti, F.: *La conversazione diseguale: potere e interazione*. Carocci, Roma (2000).
10. Krashen, S.: *Second language acquisition and second language learning*. Pergamon, Oxford (1981).
11. Klein, W., Perdue, C.: *The Basic Variety (or: Could's Natural languages be much simpler?)*. *Second Language Research* 13, 301–347 (1997).
12. Vendler, Z.: *Linguistics in Philosophy*. Cornell University Press, Ithaca (1967).
13. Giacalone Ramat, A. (ed.): *Verso l'italiano. Percorsi e strategie di acquisizione*. Carocci, Roma (2003).
14. Bertinetto, P. M.: *Tempo, aspetto e azione nel verbo italiano: il sistema dell'indicativo*. Accademia Della Crusca, Firenze (1986).

Improving Textual Competence in a Second Language Initial Literacy Classroom

Castrenze Nigrelli

Università degli Studi di Palermo, viale delle Scienze, 90128 Palermo, Italy
castrenze.nigrelli@unipa.it

Abstract. The paper aims to illustrate some textual learning activities developed for an L2 Italian initial-literacy classroom, and, in particular, for illiterate plurilinguals (mostly unaccompanied foreign minors). The activities in question belong to an experimental proposal that consists of a specific textual teaching module integrated with a second-language initial literacy course employing the communicative/affective-humanistic teaching approach. Textual activities are normally proposed to intermediate and advanced level literacy learners in second-language classrooms, in order to fully develop reading and writing abilities (i.e. functional literacy). However, based on the importance of learners' plurilingualism, oral ability, and everyday needs, the specific textual activities proposed in the initial literacy classroom have produced significant results, also with important effects on learners' motivation.

Keywords: Illiteracy, Textual Competence, Second language learning.

1 Introduction: Learning Literacy in a Second-language Classroom and Textual Activities

1.1 Illiteracy, L2 Learning, and Integration in a Migration Context

Italy is very often the first country reached by a number of migrants fleeing from African and Asian countries via a grueling journey, in order to start a new life in Europe. With reference to migration, the problem of illiteracy is nowadays increasing and it needs to be challenged by societies. Most of these people, both adults and minors, are in fact low or very low-educated, that means functionally illiterate, or even completely illiterate.¹ The problem is particularly evident in reference to the category of the so-called “unaccompanied foreign minors” (henceforth UFM), that are minors without any Italian citizen or adult who is legally responsible for them. These minors have a peculiar sociolinguistic profile, since their mostly very low, or zero, education level goes together with a sometimes striking plurilingual competence, the two fea-

¹ For a classification of illiterates it is possible to refer to Minuz [1], whose classification is in line with the Canadian document *Canadian Language Benchmarks 2000: ESL for Literacy Learners* concerning illiterate or poorly educated learners of L2 English.

tures not necessary being linked to each other (see [2]; for more details about UFM, see, among others, [3]).

The main instrument for social integration is a certain competence in the language of the host country. Unfortunately, the linguistic competence of illiterates, which also includes their L1, is basically limited to orality. The lack of familiarity with written texts has strong repercussions on the lack of meta-textual and meta-linguistic reflection. This has in turn negative consequences on the second-language acquisition process (see [4]-[5]). Low-educated and illiterate learners have in fact a slower pace in learning a second language and a greater risk of fossilization. From this perspective, illiteracy represents a factor of exclusion (see [6]): in fact, the complexity of the present-day society, with its so-called urban linguistic landscape (see [7]), continuously requires decoding activity. On the contrary, learning to read and write changes the way the learner looks at the world around him and gives him a tool to actively participate in the social life of the community in which he is located, changing somehow the whole society as well [8]. However, learning literacy in a second language is a hard and complex objective, because the learner must not only learn another language, but also learn to read and write for the first time and to transform his knowledge into useful and expendable skills. Similarly, teaching literacy in a second language classroom is twofold, since it consists in teaching not just how to read and write, but also how to use these tools independently, therefore guiding learners to build a more solid relationship with the host country. As is well known, learners' needs and expectations are crucial factors for the success or failure of a teaching-learning pathway, as well as learners' motivation. In particular, the development of motivation in illiterate foreign learners strongly depends on the awareness that the effort required in the classroom has an actual usefulness in everyday situations.

1.2 A Model of Integration for Illiterate Italian L2 Learners

The University of Palermo has developed models of integration of all of the so-called "fragile" users (i.e. UFM, asylum seekers, women) in its own classes of L2 Italian, mixing them together with other learners who have an ordinary or high level of education (often university students from other countries) and with tutors alongside who can serve as a reinforcement for them along the way. Such language courses mostly follow the communicative teaching approach, joined by the affective-humanistic perspective, looking therefore at the learner and his specific communication needs, as well as emotional and psychological aspects linked to learning.

As regards to foreign learners who are illiterates, instead, they start by following a dedicated literacy pathway, to then be included in the ordinary classes at the end of this pathway. This literacy pathway is articulated in three levels, i.e. initial, intermediate, advanced, and the illiterates are assigned to the respective classes based on their level of mastery of reading-writing. In the multiethnic, multicultural, and multilingual microcosm of these mixed classes, language learning and social integration not only coexist simultaneously, but also favor each other. Furthermore, in this environment, plurilingualism and oral ability in the target language of fragile learners constitute crucial elements as precious resources for both the learning and the integration path-

way. A core issue and starting point of the literacy pathway is the oral competence of the learner's L2. In fact, if the learner has a good oral competence in the target language, it will be easier and faster to set up a read-write pathway starting from words or phrases to be analyzed on an alphabetical level. However, this does not always happen, in this case making it a priority to focus on the development of orality, of oral communicative competence, in terms of both reception and production, as reference to which the illiterate learner can cling to. In order to be able to propose an alphabetical analysis with ease, the word (or sentence) must indeed have a meaning and therefore there must always be a recognizable context to keep in mind, which acts as a frame of reference, as a linguistic container, for the learner. In this sense, keeping classroom activities always linked to the context of external reality is a fundamental step. In addition, the process of learning basic oral skills is much shorter than that concerning the primary skills of reading-writing. Therefore, it is right to provide the learner with a basic oral knowledge, together with the pathway of reading-writing in order to allow him to navigate himself in the real world by dealing with the communicative tasks of his daily life, and starting to have and use the first means for social inclusion and promotion. As for a syllabus for illiterates, since the CEFR does not take into account this category of users (and, in general, of the levels prior to the A1), the only points of reference are the syllabus of Borri et al. [9], a guide to a course from literacy to A1 level, as well as the aforementioned Canadian Language Benchmarks 2000: ESL for Literacy Learners.

Going further into the matter, the proposed literacy pathway is divided into two interconnected phases, i.e. the phase of orality and the literacy phase. The first phase is in turn divided into three sub-phases, i.e. presentation, manipulation, and production. In the first sub-phase, the teacher starts using objects or images to elicit, or possibly present, the most suitable word to describe the object or situation shown, that is contextualized with reference to the learners' everyday reality. In the second, the teacher proposes activities that lead the learner to become familiar with the alphabetical sequence of the word linked to the image or object (e.g. placing the images in the correct reference context): this brings the learner to master the lexical area relating to the words already learnt and to deepen their semantic understanding. In the third, the learner is called to use the words just learned within free or controlled communication routines, so reusing the vocabulary learned in a precise communicative context (e.g. doing a role play inside or a task outside the classroom, in the real context). Since the consolidation of a well-placed basic vocabulary within a recognizable context facilitates the reading-writing pathway, given the particular users, at least at the level of initial literacy it is often necessary to dedicate a considerable part of the lesson to orality. After this first phase of orality, the teacher moves on to the literacy phase, in which the learner is gradually put in contact with the written word, and the teacher guides him through an analysis pathway starting from a basic form. The method adopted depends on the basic form chosen (i.e. grapheme, syllable, word, sentence, text). The synthetic methods, which start from the smallest element (the letter) to reach the largest one (the sentence or the text), are no longer in use. The global method, which starts from the sentence and manages to identify the word, is in itself very valid, but requires that the learner possesses a high oral competence in the target lan-

guage. The method adopted in the present case is analytical-synthetic, which starts from the word, analyzed semantically, then moves on to the analysis of the syllable, chosen as the base form, and finally returns again to the word, now also structurally analyzed.

1.3 Textual competence and illiterate learners

Focus on initial literacy classes, mostly composed of UFM, they basically revolve around the word as a starting point for a first semantic approach and, only then, as a support for a more analytical reflection. Taking into account the difficulty of abstraction and categorization which characterizes the illiterates, the teacher will prudently try to guide learners towards the awareness that words are made up of smaller elements bearing meaning, without using abstract categories.

In some ways, the word may therefore seem to be the horizon within which this first level is defined. If this is true for the literacy phase, the same cannot be said about the phase of orality: in the third sub-phase (i.e. production), the learner begins to become familiar, albeit to a minimum extent and not in an analytical and fully aware manner, with the upper levels of the sentence and the text. In the perspective of Textual Linguistics, the textual competence, or the ability to know how to identify the information conveyed by the text, as well as the way in which it is conveyed, is part of the more general communicative competence. Textual competence concerns, therefore, the awareness of logical and formal organization of the text, thus implying the ability to make inferences and to grasp the elements of cohesion and coherence. In a learning pathway it is important to work on textual competence because through it cognitive mechanisms are developed. This has important implications both on the learning of a second language, and on the reinforcement of instrumental reading and writing skills. The texts also serve to always propose new contexts for use, which stimulate and motivate the learner if drawn from everyday experience. Moreover, outside the classroom, the urban context offers (when it does not impose) continuous textual stimuli. Within the literacy pathway described so far, working on texts is normally destined to the intermediate literacy level and, even more, to the advanced literacy level, whose main objective is the full functional literacy. However, textual activities have also been tested in the initial literacy class, as an integrated module, despite the fact the main objective of this course is essentially limited to instrumental literacy. The aim was to immediately accustom the learner to some types of text. These texts were chosen because they were suitable for his profile and his needs, as well as perfectly integrated with the contexts and vocabulary proposed in the classroom. Given the minimum level, the texts were mostly discontinuous. The main textual activities proposed in the classroom are illustrated below (see next Section) in order to give an idea of the positivity of the result.

2 Experimental Textual Activities in the Initial Literacy Classroom

To ensure that the textual activity is useful, it is almost always proposed around the end of the orality phase, together with the production sub-phase, when the teacher proposes some production activities, which would help the practice within communicative routines. It is usually a role play activity. The teacher mimics a situation and invites the learners to elicit some simple communicative strings, which can be hypothesized starting from the teacher's actions. Once they have reached a series of acceptable strings, the teacher has them fixed to the learners, who will then try to recreate the scene in pairs. Precisely within this activity, the teacher inserts the textual element, very often introducing authentic or adapted materials that arouse the curiosity and interest of the class. Examples are the receipt and the label of a shirt. These are two discontinuous texts, classifiable as descriptive/regulatory, very common in everyday life. The two texts are proposed respectively in two different role plays, one set in the supermarket and the other in a clothing store. The two guided oral productions, related to both role plays, are linked to the vocabulary of food and clothing, and, at the same time, the communicative strings suggested by the mimed situation are also based on the textual element, thanks to the presence in the classroom of the real object. The focus on the textual element is almost always presented just before moving on to the role play, or contextually to it, during the string elicitation phase. Starting from the real object, and with the support of the blackboard and/or a worksheet, the teacher proposes to focus the learners' attention on salient textual elements in relation to the type of text and its function. On the receipt (see Figure 1), for example, it will be highlighted the importance of the date, the indication of the price and the change, while, on the shirt label, the indication of size, color, price, and washing instructions. This type of activity arouses the interest and motivation of the learner, since it is clearly coherent with the real context, with frequent and plausible communicative situations and with his communication needs.

DOCUMENTO COMMERCIALE
di vendita o prestazione

DESCRIZIONE	IVA	Prezzo(€)
1 X CAFFE CALDO	10,00%	0,90
1 X 1/2 ACQUA	10,00%	1,00
TOTALE COMPLESSIVO		1,90
di cui IVA		0,17
Pagamento contante		1,90
Pagamento elettronico		0,00
Non riscosso		0,00
Resto		0,00
Importo pagato		1,90

Fig.1. Example of receipt.

In other cases, the textual activity is disconnected from a communicative situation like the one suggested by the role play, but this does not mean that it is disconnected from the real and daily context of the learners' needs. This is the case with the coffee machine, which presents a number of textual elements. The activity takes place outside the classroom. After having proposed the vocabulary of food and drink in the previous lessons, the teacher guides the students to the use of the coffee machine, drawing their attention to the text (also in this case discontinuous and descriptive/regulatory), and, in particular, to the correspondence between number and product, between button and function.

A further example of a textual proposal, in some respects different than the examples just illustrated, is the daily agenda. Activity on the agenda is particularly interesting, because it can go beyond the planned textual objectives and the general expectations of the teachers. Such activity takes place at the end of the lesson. The basic idea is to accustom the learners to compiling an empty handmade notebook, as if it is a daily diary, on which to write down, with the help of the teacher, the words learned under the formula "Oggi ho imparato..." ("Today I learned"). Before completing the agenda, the teacher manages a brief moment of elicitation in plenum, which is then followed by the transcription of each one individually. This activity has the textual function of making the learner approach a blank page, which he himself - in a guided

manner - would fill with various elements: the date at the top, the lines on which to write within the margins, and finally the words. It represents also a daily opportunity to train writing, with particular attention to isolating words, identifying their boundaries. Furthermore, day after day, the empty notebook takes the form of a real agenda, which, in the absence of a textbook, is configured as the object that most resembles it. This activity is very useful because it develops a series of indispensable elements for the success of a learning pathway. First of all, it helps to develop the learners' autonomy: over time, in fact, it is the students who decided to write what they want, including words learned elsewhere. In addition, through this activity the teacher educates them to care for the book and, in general, for the school equipment. The students also initiate an identification mechanism, since, at the suggestion of the teacher, each of them could personalize the cover with a drawing or a name. A further important aspect is the motivation that arises in the learner when focusing and putting down in writing what he has learned in class, taking note of a progressive advancement in his own knowledge and skill. Finally, always placed at the end of each lesson, the time to complete the agenda gradually becomes, over time, an essential moment of relaxation: this has a very positive emotional impact, because it creates a comforting routine, an indispensable element for this type of learner who is not used to being in a school environment.

References

1. Minuz, F.: Italiano L2 e alfabetizzazione in età adulta. Carocci, Roma (2005).
2. Tarone, B., Bigelow, M.: A Research agenda for second language acquisition. In: Vinogradov, P., Bigelow, M. (eds.) *Low Educated Second Language and Literacy Acquisition - Proceedings of the 7th Symposium*, Minneapolis, Minnesota, USA, pp. 5–26. (2011).
3. Barone, L.: *L'accoglienza dei minori stranieri non accompagnati. Tra norma giuridica e agire sociale*. Key Editore (2016).
4. Luria, A. R.: *Cognitive development: Its cultural and social foundations*. Harvard University Press, Cambridge (1976).
5. Ong, W.: *Orality and Literacy*. Methuen, London-New York (1982).
6. Demetrio, D., Moroni, F.: *Alfabetizzazione degli adulti: teoria, programmazione, metodi*. Editrice sindacale italiana, Roma (1980).
7. Landry, R., Bourhis, R. Y.: Linguistic Landscape and Ethnolinguistic Vitality: An Empirical Study. *Journal of Language and Social Psychology* 16(1), 23–49 (1997).
8. Watson, J. A.: Cautionary tales of LESLLA Students in the High School Classroom. In: Vinogradov, P., Bigelow, M. (eds.), *Low Educated Second Language and Literacy Acquisition - Proceedings of the 7th Symposium*, Minneapolis, Minnesota, USA, pp. 203–234. (2011).
9. Borri, E., Minuz, F., Rocca, L., Sola, C.: *Italiano L2 in contesti migratori. Sillabo e descrittori dall'alfabetizzazione all'A1*. Loescher, Torino (2014).

Multiword Terms and Machine Translation

Serge Potemkin

Lomonosov MSU, Moscow, Russia
prolexprim@gmail.com

Abstract. In this article we discuss, using morphological analysis and foreign equivalent selection, the following issues: the definition of syntax and the logical-semantic structure of terms using morphological analysis and foreign equivalent selection. Texts of articles in the “Voprosy Psichologii” (*Psychological Problems*) journal were used as the source corpus. Statistics of multiword terms in this corpus were calculated as well as the pointwise mutual information (PMI) between terms. We defined the distance between multiword terms in a multidimensional space, and the measure of terms proximity, as the minimal semantic distance between them. Then, we performed multidimensional scaling for visualising the terms space. Machine translation was used as a means to finding equivalents of scientific and technical terms in two arbitrary languages. With the help of forward and reverse translation of terms, using online machine translation tools, the meaningful equivalents were evaluated. We measured the proximity between terms as the Levenshtein distance between the original term and its direct and reverse translation, and tried to minimise this measure.

Keywords: Term Equivalent, Structure of Composite Terms, PMI, Machine Translation, Direct and Reverse Translation.

1 Introduction

The increasing volume of published scientific and technical information is doubling annually or faster, requiring the identification and standardisation of the vocabulary used in science and technology, to allow the readers to correctly understand the essence of the message. This observation fully applies to the terminology used by different authors in the same and related areas of science and technology, as well as in academic and popular literature. Perhaps an even more important issue for development is the need for an exchange of scientific information generated in different countries in different languages. Such an exchange is impossible without the accurate translation of scientific articles which contain terminology.

Terminology problems are also associated with the development of machine translation (MT) systems. Grammar of the scientific text is simple and its translation depends mainly on the correct translation of nominal constructions, primarily terminological words and phrases. At the moment "it is no longer in doubt that for the correct, scientifically substantiated solution of terminological problems, one should learn the terminology with the recognition of its nature and logical existence. That's why with-

in the framework of the above-mentioned, the problems of terminology should be explored by linguists and technologists" [1].

2 The concept of the Scientific Term

The term is a word, phrase, acronym, or other lexical unit, which designates the corresponding extra-linguistic concept in the real world. According to Mikhail Glushko "the term – is a word or phrase to express concepts and a notation of objects having, thanks to its rigorous and precise definitions, clear semantic boundaries within an appropriate classification system" [2]. Generally, the term must be unambiguous, i.e. correspond to only one particular entity. The uniqueness of the term, in contrast to the general-language word, does not depend solely on the surrounding context. Specifically the term can be used in isolation, its belonging to the specific terminology defines the uniqueness of the term within this terminology; in other areas the term may have a completely different denotation. Unfortunately, even in the same subject area, or the same knowledge field, the term may be defined differently by different authors. The object of our study of terminology is the vast area of "Psychology".

For example, consider the definitions given by different authors of the term "*photopsia*":

a) Photopsia, (from the Greek φωτός — light + ὄψις - vision) - subjective light phenomena (feeling), not having the nature of certain figures or objects. Usually these are flashing spots, sparks, and light zigzags etc.; photopsia is caused by the action of the mechanical or toxic stimulation of the visual analyser [3].

b) Photopsia. The emergence of moving shapes, dots, spots, etc., mostly luminous, shining in the field of view. Photopsia is observed in diseases of the retina, and elementary visual hallucinations as a psychopathological phenomenon [4].

The first definition does not contain the notion of a "psychopathic phenomenon", which is important in the field of psychology. Such examples are numerous. Often, even the same author will give several interpretations of the term which reflects the different use of it in the various sub-fields of science. The reader could hardly guess what kind of entity is being described by the term in the context.

Even greater difficulties arise in the process of translation of the term. For example, the term "*слияние*" has at least 2 English equivalents: [4]

СЛИЯНИЕ (symbiotic) (MERGING)

СЛИЯНИЕ (synthesizing) (FUSION)

3 Term definition problems

The use of compound terms - terminological expressions partly cope with such difficulties. In this context, an attempt was made to review the syntax and logical-semantic structure of terms. A very important class of multiword terms (MWT) is one where the meaning/semantics is not obvious from the composition of the meanings of the constituent words. e.g. in "*a double blind expertise*" the individual constituent words have no connotation relation to the actual meaning of the phrase, which is to ask experts to give suggestions about some matter. Contrast this with *blind man*

which has no other interpretation than the literal one obtained from the composition of the constituent words. These kinds of collocations are very common in human language due to the prolific metaphorical and figurative use of language. Handling these kinds of MWT is crucial for robust natural language processing [5].

4 Morphology of Compound Terms

Using the resource glossary [6] containing more than 2,500 entries, we performed an automatic morphological analysis of single-word and compound-word psychological terms. As a result of the morphological analysis each term is attributed to the following type of information:

Table 1. Morphological characteristics of the compound word term

№	word form	lemma	morphology	POS
1	адекватность (adequacy)	адекватность (adequacy)	ж=Ns;Asi;	n
2	ощущения (feelings)	ощущение (feeling)	с=Gс;Np;Api;	n
3	и (and)	и (and)	союз=0;	cnj
4	восприятия (of perception)	восприятие (perception)	с=Gс;Np;Api;	n

where ж - feminine noun; с – neuter noun; N – nominal case; s - singular; A - accusative case; i - the inanimate; n - noun; etc. (Notation corresponds to the electronic version of A.A.Zaliznyak's dictionary [7]).

Counting the morphological characters of all terms in the glossary gave the results shown in the following Table 2.

Table 2. Frequency of syntactic structures of terms depending on the number of constituent words.

POS + case	Term example	# of words	Frequency
nN	абазия (abasia)	1	0.313
aN nN	абсолютный порог (absolute threshold)	2	0.319
nN nG	автоматизация движений (movements automation)	2	0.131
nN nN	ведущая деятельность (leading activity)	2	0.013
nN aN	агнозия зрительная (visual agnosia)	2	0.004
numN nG	двенадцать шагов (twelve steps)	2	0.002
nN aG nG	амнезия раннего детства (early childhood amnesia)	3	0.063
aN nN nG	агрессивное поведение животных (aggressive behavior of animals)	3	0.035
aN aN nN	абсолютная слуховая чувствительность (absolute hearing sensitivity)	3	0.029
nN preL nL	либидо во фрейдизме (libido in Freudianism)	3	0.004
nN preG nG	запечатление у животных (imprinting in ani-	3	0.003

	mals)		
nN nG nG	нарушения восприятия времени (time perception disorders)	3	0.003

The frequency of the other grammatical structures is less than 0.002.

5 Logical and Semantic Structure of Terms

The study of the logical and semantic structure of a term was performed using Pointwise Mutual Information (PMI) metrics.

PMI is defined as:

$$PMI(wa, wb) = \ln(p(wa, wb)/p(wa)p(wb)).$$

Where wa, wb – terms; $p(wa, wb)$ – the probability of co-occurrence wa and wb ; $p(wa), p(wb)$ – probabilities of occurrence wa and wb respectively.

The distributional hypothesis in linguistics is: words that occur in similar contexts tend to have similar meanings [9]. This hypothesis is the justification for applying the PMI to measuring word similarity. A word may be represented by a vector in which the elements are derived from the occurrences of the word in various contexts. Proximate row vectors in the word–context matrix indicate similar word meanings.

We used machine-readable texts from the journal, “Voprosy Psichologii” (*Psychological Problems*) issued in the years between 1980-2010, [8] as the source corpus. It contains more than 13 billion words and 282,000 lemmas. A statistical diachronic study of this corpus was published in 2015 [10]. As its base, the experts in psychology chose 100 specific two-word terms as examples of terminology in this field. We performed the usage count for each of these terms and the PMI method for each pair of them. Some pairs with their counts and PMI are shown in Table 3. The table contains the top part of all pairs ordered according to PMI value. Good collocation pairs have a high PMI because the probability of co-occurrence is only slightly lower than the probability of occurrence of each word. These are:

клиническая психология (clinical psychology) <> консультативная психология (counseling psychology); PMI=3.6016

ценности жизни (values of life) <> экзистенциальный анализ (existential analysis); PMI=4.7447

гендерный анализ (gender analysis) <> женская психология (female psychology); PMI=4.7729

The PMI measure could be used for: ranking of web pages, rare term extraction from NL texts, sentiment classification, etc.

In the next stage we performed the search of proximity measure for two multi-word terms based on MTI values. Each term’s Ta is a vector in multidimensional space (in our case a 100-dimensional space) with components corresponding to the $MTI(Ta, Tbi)$ values where Tbi – the set of all the terms (in our case $i=\{1,2,\dots,100\}$).

Table 3. The top part of the list of distances between terms ordered in descending order

A-B distance	Term A	Term B
6.0681	методы исследования (re-	психологические исследования (psy-

	search methods)	chological research)
6.1343	индивидуальные особенности (individual characteristics)	индивидуальные различия (individual differences)

The distance measure could be used for information retrieval, terms clustering, multi-dimensional scaling, etc [11].

6 Terms Translation

Finally, we performed a study of translated terms using the online translator.

The methodology of this research was as follows the analysed term was subjected to machine translation with the help of an online translator¹. Then, the reverse translation was performed, and the Levenshtein distance² (LD) between the original Russian text and the text obtained as a result of direct – reverse translation was calculated³ [10]. If this distance is equal to zero, the term and its foreign equivalent are accepted. Otherwise we conclude that the machine translation system "does not know" the Russian term, and, accordingly, incorrectly selects its foreign equivalent. This result means that the online translator "knows" the Russian term and gives the adequate translation. Otherwise the Russian term should be changed.

Initial (Russian) = вера в справедливый мир

Russian to English = just-world hypothesis

Back English to Russian = вера в справедливый мир // Levenshtein distance = 0

An example of an incorrect translation of the term:

*запечатление у животных и *захватив животных // Levenshtein distance = 2*

While iteratively repeating the procedure of direct and reverse translation the process can sometimes come to zero LD. Then the source and the resulting terms often are synonymous.

An example of an iterative equivalents search:

Initial (Russian) = нарушения восприятия времени

R to E = disorders of perception of time

Back E to R = расстройство восприятия времени // LD = 1

R to E = time perception disorder

Back E to R = Время расстройство восприятия // LD = 3

R to E = Time perception disorder

Back E to R = расстройство восприятия времени // LD = 2

R to E = disorder of time perception

Back E to R = расстройство восприятия времени // LD = 0

One can conclude:

нарушения восприятия времени =sin= расстройство восприятия времени

¹ Google translator, available at <https://translate.google.ru/>, last accessed: 30.02.2019

² Levenshtein distance, available at https://en.wikipedia.org/wiki/Levenshtein_distance , last accessed: 30.02.2019

³ Direct and reverse translation available at <http://www.philol.msu.ru/~serge/Translation/form11.php> , last accessed: 30.02.2019

The resulting term, after iterations, may be the converse of the source one: *абсолютная слуховая чувствительность =conv= абсолютная чувствительность слуха* (*absolute hearing sensitivity =conv= absolute sensitivity of the hearing*); or the original term and the resulting term have a different word order (permutation): *амнезия раннего детства =perm= раннего детства амнезия* (*amnesia of early childhood = perm = early childhood amnesia*)
 Logical and semantic analysis using machine translation allows the selection of clusters of related terms, a partial order relation according to the "Levenstein distance."

7 Conclusion

We have performed an automatic morphological analysis of terms in the field of psychology on the basis of A.A. Zalizniak's grammar dictionary [7]. The statistics of syntactic structures of composite terms (phrases) is extracted for the later retrieval of such terms in the text. The foreign-language equivalents for the terms are searched by automated means through the procedure of multiple direct and reverse translations. The possibility of clustering a set of terms in a given subject area was mentioned. The results can be used to improve MT systems.

References

1. Gorelikova, S.N.: The nature of the term and some features of term formation in English [Priroda termina i nekotorye osobennosti terminoobrasovania v angliiskom iazyke] // Vestnik OGU №6, 129-136 (2002)
2. Glushko, M.M. et al.: Functional style of the popular language and methods of its study [Funkcionalnyi stil obschestvennogo iazyka i metody ego issledovania] Moscow. – pp 1–33 (1974)
3. Mescheriakov, B., Zinchenko, V.: The concise psychological dictionary [Bolshoi psikhologicheskii slovar], ACT Moscow, Prime-Evroznak, -pp.1 – 816 (2009)
4. Bleikher, V, Kruk, I: Explanatory dictionary of the psychological terms [Tolkovyi slovar psikhologicheskikh terminov] NPO MODEK, Voronezh, pp. 1 – 640 (1995)
5. Kunchukuttan, A.: Multiword Expression Recognition (2017) . – available at <https://pdfs.semanticscholar.org/3e3f/d0173dcb28aa1a11d5342da527a835235ae4.pdf> last accessed 15.02.2017
6. Barness, E.M., Bernard, D.F.: Psychoanalytic terms and concepts [Psykhoanaliticheskie terminy I poniatia], Class, M, pp.1 – 304 (2000)
7. Zalizniak, A.A.: Grammatical dictionary of Russian [Grammaticheskii slovar russkogo iazyka] (2017) available at: <http://www.speakrus.ru/dict/zdf-win.zip>, last accessed: 30.02.2019
8. Problems of psychology [Voprosy psikhologii] Scientific journal, Issues 1980-2010 years, Pedagogika, Moscow, (1980) available at <http://www.voppsy.ru/> last accessed 16.02.2019
9. Harris, Zellig: Distributional structure. Word 10(23). 146–162 (1954)
10. Potemkin, S.B., Hasin, L.A., Hasina, P.L., Schedrina, E.V.: Analysis of psychology development on the basis of terms frequency dynamics [Analiz tendencii razvitiia psikhologii na os-

- nove vyavlenia dinamiki chastoty ispolzovania psikhologicheskikh terminov], Problems of psychology [Voprosy psikhologii], Vol.6, 95-103 (2015)
11. Potemkin, S.B., Kedrova, G.E.: Exploring semantic orientation of adverbs, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2011” [Komputernaya Lingvistika i Intellektualnye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2011”], Bekasovo, pp.71–78 (2011)

Towards a Cross-linguistic Study of Phraseology across Specialized Genres

Ana Roldán-Riejos¹[0000-0002-9635-2814] and Łukasz Grabowski²[0000-0002-3968-9218]

¹ Universidad Politecnica de Madrid, Spain
ana.roldan.riejos@upm.es

² University of Opole, Poland; University of Ostrava, Czechia
lukasz@uni.opole.pl; Lukasz.Grabowski@osu.cz

Abstract. This poster paper aims to present an early-stage work of a group of researchers collaborating within the project EMPHRASE. The corpus-based cross-linguistic studies of a specialised phraseology across different linguistic registers, genres and domains of language use have not received sufficient attention yet (Buendía 2013, Aguado 2007, Ramisch 2015, Grabowski 2018), notably in terms of turning the results of largely descriptive studies into actionable knowledge. The project revolves around three main axes: 1) compiling and structuring an inventory of word combinations from different genres, disciplines and languages; 2) exploring and analysing cross-genre characteristics as well as typical features found in the phraseological repertoire (e.g. measuring the degree of frequency and use of phraseological patterns); 3) testing and fine-tuning methodologies for identification and analysis of recurrent multi-word items used in texts written in typologically different languages. The proposed study is concerned with lexical phrases commonly used in specialised/technical domains; we focus on word combinations which usually consist of two or more lexical items associated together in systematic ways, either by frequent use or by genre convention. Eventually, the analysis intends to integrate lexical, semantic and communicative aspects involving various European languages, i.e. English, Spanish, Polish and Russian.

Keywords: Genre-based study, Specialized phraseology, Corpus-based phraseology, Corpus analysis.

1 Introduction

In this poster we present an early-stage corpus-informed collaborative project on a specialized phraseology with the main purpose of developing accessible database with the most frequent and salient word combinations found in the written genres of product descriptions, patient information leaflets, summaries of product characteristics, technical reports and research articles, among others. The technical domains under study include pharmaceutical industry, sport sciences, construction engineering, computing engineering and maritime engineering in English, Spanish, Polish and Russian. To our knowledge, no cross-linguistic and cross-genre compilation of this type has been

previously undertaken. As this study is in a very preliminary stage, it is hoped that it will pave the way for more comprehensive and detailed research in the future.

1.1 Main aims

The aims of the project are threefold:

- a) To compile and analyse specialised expressions - largely in terms of use and discourse functions - of specialized phraseologies (e.g. collocations, recurrent n-grams, such as lexical bundles and phrase frames) in selected specialised text types and genres written in English, Spanish, Polish and Russian.
- b) To compare and contrast the results in order to explore cross-linguistic variation in terms of the use and functions of recurrent phraseologies across genres and text types as well as languages
- c) To fine-tune the methodology of identification and analysis of various types of recurrent multi-word units (contiguous and non-contiguous ones) when applied to texts produced in typologically different languages

2 Methodology

The study is largely based on, but not limited to, an electronic corpus of academic and professional genres compiled by the research group EMPHRASE based at Universidad Politecnica de Madrid over the last 10 years totalling over 250,000 lexical combinations, and the corpus is still work-in-progress. The text types and genres included in the corpus mainly consist of research journal articles, product descriptions, technical reports as well as texts compiled by individual authors and used in their research. The software used for identification of word combinations and their study are AntConc 3.5.7 (Anthony 2018), SkechEngine (Kilgarriff et al. 2014), Formulib (Forsyth 2015), among others.

To provide an example of a preliminary functional and linguistic analysis of a genre, we went through 4 sequential steps in the compilation and selection process; first of all, we used the Word List tool of the software to obtain the most frequent lexical words in each genre. For each list we decided to fix a threshold of frequency occurrence of 10 hits for the different corpora (initially English; later we will proceed with Spanish, Polish and Russian). Using Collocates tool, we searched for frequent collocates (to the right and to the left) of the obtained words. The program provides statistical measures of collocational strength, such as MI (Mutual information) or t-score. At this stage, we also elicited recurrent n-grams sorted by frequency using Clusters/N-Grams tool. Also, we explored in greater detail the following syntactic combinations: N+N; A+N: V+N. Finally, contextual factors were checked in borderline cases using Concordances KWIC (KeyWord in Context) tool, which provided co-text of lexical combinations (cf. Cuadrado et al. (2016) for a more detailed description of the procedure).

The next step was to record the obtained collocations in Excel files arranged by genre, domains and subdomains to be further compared cross-linguistically. This procedure also enables one to perfunctorily explore cross-genre differences, including register and other rhetorical features. Throughout the analysis, we looked at literal as well

as figurative uses of word combinations. For example, *resisting frames* from the construction engineering domain can be used literally or figuratively depending on context. Literally, *moment-resisting frames* designate rectilinear combinations of beams rigidly connected to columns so that they can bend and resist earthquakes. Notwithstanding this, contextual examples of “behavior of resisting frames”, and “sensitivity of resisting frames” were found, figuratively attributing animate properties to this engineering element (cf. Branci et al. 2016). In order to identify figurative phraseology, it is essential to examine the contextual clues in the concordances lines as well as to measure – in the future - inter-rater agreement (using raw inter-rater agreement, Cohen’s kappa etc.) so as to determine whether the collocation or multi-word item was indeed used figuratively.

3 Preliminary results

Table 1 shows the cross-linguistic criteria considered in genre analysis, including degrees of linguistic formality, use of figurative language, distinctive elements, politeness conventions and image use in texts written in different languages. Using these criteria we intend to conduct an initial qualitative analysis; we hypothesize that the results that may vary across languages.

Table 1. Criteria adopted for the analysis across languages

REGISTER	FIGURATIVE LANGUAGE USE	DISTINCTIVE ELEMENTS	POLITENESS CONVENTIONS	IMAGE USE
Formal	Metaphor (Mph)	Collocation class (e.g. N+N)	Hedges, Shields	Infographics
Informal	Metonymy (Mnm)/Other	Number of words (e.g. 2)	Passive voice, Modal verbs	Photographs

Table 2 presents the specific features to be analyzed in other genres under study.

Table 2. Features considered in the cross-genre analysis

Text type/genre	RESEARCH ARTICLES	PRODUCT DESCRIPTIONS	TECHNICAL REPORTS
Style	Formal	Informal	Formal/informal
Figurative use	Mph	Mnm	Mph
	Mnm		Mnm
Collocation class	N+N	A+N	N+N
		N+N	V+N
		V+N	A+N
		P+V	
		Adv+A+V	

Characteristic discourse fea- tures	Downtoners, Intensifiers Approximators Personal pro- nouns. Passive voice	Downtoners, Intensifiers Approximators Personal pronouns	Downtoners, Intensifiers Approximators Passive voice Modal Verbs
Visuals	Diagrams, ta- bles, charts	Photographs, draw- ings	Tables, diagrams, pho- tographs, drawings

All in all, we believe that similarities and differences between genres and languages at the level of salient specialized phraseology can be a useful resource for researchers and writers in a specialised discourse community. To that end, we plan to explore a sample of Spanish, Polish and Russian text types and genres in an attempt to reuse the proposed model cross-linguistically, although some early phraseological work – from a perspective of frequency-driven phraseology – on English, Polish and Russian patient information leaflets has been already conducted (Grabowski 2014, Grabowski 2018a, Grabowski, under review).

4 Conclusions and future work

(a) This work has considerable potential to contribute to cross-linguistic research on recurrent phraseologies in specialized text types and genres, and its results may, among others, help improve LSP vocabulary learning and fluency (cf. Boers & Lindstromberg 2008; Roldán-Riejos & Úbeda 2018a; 2018b), and provide potentially useful data for translators, terminologists, lexicographers and researchers from various discourse communities.

(b) The future study will offer an opportunity to test and fine-tune methodologies of identification and analysis of collocations and longer multi-word items when applied to data written in typologically different languages. It is essential since many approaches (e.g. lexical bundles (Biber et al. 1999), Pattern Grammar (Hunston & Francis 2000), phrase frames (Fletcher 2002)) have been developed and applied largely to English-language material. For example, it might be interesting to see how to apply the distributional criteria (frequency, distribution range, coverage, various collocational strength metrics) to identification of recurrent word combinations in texts written in English, Spanish, Polish and Russian. This will also provide an opportunity for reflection on how to measure the amount of formulaic language in texts written in typologically different languages.

References

1. Aguado de Cea, G. “A multiperspective approach to specialized phraseology: Internet as a reference corpus for phraseology”. In: M. Esteve & S. Posteguillo (Eds), *The Texture of*

- Internet: Netlinguistics in Progress*, pp. 182-207. Newcastle: Cambridge Scholars Publishing (2007).
2. Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. *The Longman Grammar of Spoken and Written English*. London: Longman (1999).
 3. Boers, F. & Lindstromberg, S. (Eds) *Cognitive Linguistic Approaches to Teaching Vocabulary and Phraseology*. Berlin/New York: Mouton de Gruyter (2008).
 4. Branci, T., Yahmi, D., Bouchair, A., Fournelley, E. "Evaluation of Behavior Factor for Steel Moment-Resisting Frames". *International Journal of Civil and Environmental Engineering* 10(3): 396-400 (2016).
 5. Buendía, M. *Phraseology in specialized language and its representation in environmental knowledge resources*. PhD dissertation. Universidad de Granada (2013). <http://hdl.handle.net/10481/29527>, last accessed 2018/10/16.
 6. Cuadrado, G., Argüelles, I., Durán, P., Gómez, M-J, Molina, S., Pierce, J., Robisco, M., Roldán, A. & Úbeda, P. *Bilingual Dictionary of Scientific and Technical Metaphors and Metonymies Spanish-English/English-Spanish*. London: Routledge (2016).
 7. Fletcher, W. "KfNgram". Annapolis: USNA (2002). <http://www.kwicfinder.com/kfNgram/kfNgramHelp.html>, last accessed 2019/05/20.
 8. Forsyth, R. "Formulib: Formulaic Language Software Library" (2015). <http://www.richard-sandesforsyth.net/zips/formulib.zip>, last accessed 2018/11/30.
 9. Grabowski, Ł. "On Lexical Bundles in Polish Patient Information Leaflets: A Corpus-Driven Study". *Studies in Polish Linguistics* 19(1): 21-43 (2014).
 10. Grabowski, Ł. "On identification of bilingual lexical bundles for translation purposes. The case of an English-Polish comparable corpus of patient information leaflets". In: R. Mitkov, J. Monti, G. Corpas Pastor and V. Seretan (Eds), *Multiword Units in Machine Translation and Translation Technology [Current Issues in Linguistic Theory 341]*, Amsterdam: John Benjamins, pp. 182-199 (2018).
 11. Grabowski, Ł. "Distinctive Lexical Patterns in Russian Patient Information Leaflets: A Corpus-Driven Study". *Russian Journal of Linguistics* 23(3) (in print)
 12. Hunston, S., & Francis, G. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins (2000).
 13. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. "The Sketch Engine: ten years on". *Lexicography* 1(1): 7-36 (2014).
 14. Ramisch, C. *Multiword Expressions Acquisition: A Generic and Open Framework*. New York: Springer (2015).
 15. Roldán-Riejós, A. & Úbeda, P. "El léxico de la ingeniería y su aprendizaje: estudio exploratorio". *EuroAmerican Journal of Applied Linguistics and Languages E-JournALL* 5(1): 60-80 (2018^a).
 16. Roldán-Riejós, A. & Úbeda, P. "Phraseological study and translation of technical metaphor in architecture and construction engineering" 13th *Teaching and Language Corpora Conference TaLC*. University of Cambridge (UK) (2018b).

Multi-word Units in Machine Translation: why the Tip of the Iceberg Remains Problematic – and a Tentative Corpus-driven Solution

Jean-Pierre Colson

Université catholique de Louvain
Louvain School of Translation and Interpreting,
place Cardinal Mercier 14, Louvain-la-Neuve, Belgium
jean-pierre.colson@uclouvain.be

Abstract. Neural machine translation (NMT) has recently made considerable progress in the improvement of the quality of the produced texts. Among the new features of NMT are the fluidity of the translations and their successful treatment of multi-word units. In this paper we first report the results of a global and automated evaluation of the percentage of phraseology in the translations produced by Google Translate and DeepL. A corpus-based approach makes it possible to estimate that both NMT systems succeed in producing an average percentage of phraseology that is quite natural and may sometimes even be higher than in natural language production by native speakers. Taking a closer look at some problematic cases, however, reveals that the phraseological value of NMT systems may be deceptive, as they are often unable to cope with contextual complexity and low-frequency idioms.

Keywords: Neural Machine Translation, Phraseology

1 Introduction: Lingering Doubts about Neural Machine Translation

Neural machine translation (NMT) is now generally considered as a major breakthrough in machine translation (MT). As pointed out by Loock [1], the growing success of NMT since 2015 is now such that in 2018, and for the first time, more than half of all translation companies and services in Europe reported using MT to some extent. In particular, the results obtained by DeepL with NMT have impressed the general public (as shown by the many newspaper articles on DeepL's web page), as well as language professionals and researchers.

According to DeepL's official [Press Release](#), their system outperforms all existing ones, both on the basis of human evaluation as against the Bleu score [2]. [Google Translate](#), however, has also been switching from statistical machine translation (SMT) to NMT and is reporting better and better results, with a much larger number of language pairs.

Apart from DeepL’s own references, a limited number of studies have been devoted to an in-depth evaluation of the system, and of NMT as a whole. An important theoretical remark should be made in that respect. As NMT and MT in general have been improving their results in recent years, they have attracted not only researchers in NLP, but also language professionals and translators in particular. However, communication between these two worlds is difficult, because the norms for quality checking differ very much between for instance researchers in NMT and professional translators. It is therefore crucial that the evaluation of the performance of NMT should be carried out in various ways. In the past, MT algorithms were mainly evaluated by means of automated metrics such as the BLEU score [2] or the Meteor score [3]. In recent years, those metrics have been completed by other techniques including human evaluation and the recourse to a ‘Challenge Set’, a selection of difficult linguistic constructions or structures submitted to the MT systems. Thus, Isabelle et al. [4] propose a set of 108 short English sentences that are used as a challenge set for measuring the efficiency of MT systems. The results were checked by 3 human evaluators. Interestingly, Pierre Isabelle published on the Web an [addendum](#) to the paper, in which the authors submitted the challenge set to the latest version of DeepL. It is striking to note that DeepL turned out to perform much better (with an overall correction score of 84 %) than the other systems.

This enthusiasm for the progress made by NMT in general, and DeepL in particular, is not really shared by Loock [1], who uses a methodology derived from Corpus-based Translation Studies [5]. The study compares a corpus of original French with a corpus of translations from English into French, carried out by 2 NMT systems: DeepL on the one hand, and the system used by the European Commission, [MT@EC/eTranslation](#) on the other. The results clearly show that, in spite of the apparent fluidity reached by NMT in the target texts, they systematically display a number of flaws as compared to original texts. Among the linguistic structures targeted by the study (such as the frequency of the French word for “thing”), both systems display an abnormal overrepresentation. This makes the end product, the target text produced by the two NMT systems, very poor from the point of view of translation professionals, and it therefore requires a lot of post-editing. In other words, the target texts produced by NMT may be artificial, because they miss some key features, especially idiomaticity, that make them natural in the eyes of native speakers.

As all evaluation methods only address part of this complex issue, our point of view is that a wide variety of approaches should be used in the study of the performance of NMT systems, which will also contribute to a better understanding of the future needs of post-editing. In this paper, two research questions address this problem from the point of view of multi-word units or phraseology [6]:

1. What is the rough proportion of phraseology in texts produced by NMT, as compared to original texts in the target language, belonging to the same register?
2. Does phraseology still pose a major problem to NMT, and if so why?

2 Phraseology and NMT: an Experiment

As regards the first research question mentioned at the end of the preceding section, it should first be pointed out that there is no standard measure of phraseology in a text. Phraseology as defined by Burger et al. (2007) [6] can be considered in the broad sense, including weakly idiomatic constructions (such as lexical collocations, e.g. *harsh criticism* and grammatical collocations, e.g. *again and again*) or in the strict sense (with very idiomatic constructions such as idioms, e.g. *spill the beans*). Formulaic language as defined by Wray [7] includes formulas in the broadest sense, many of which have a communicative function (e.g. *What time is it?*). Finally, construction grammar or CxG [8] [9] [10] considers that language as a whole consists of *constructions* in the sense of (partly) arbitrary pairings of form and meaning, at various degrees of abstraction and schematicity.

A construction may be a word in the traditional sense (e.g. *book*), a bound morpheme (*pre-*, *-ing*), an idiom (*spill the beans*, *take the rough with the smooth*), a partially filled idiom (*take X into account*), but also an abstract construction such as the ditransitive construction or the passive. At the intersection of phraseology (in the broad sense) and CxG, many discontinuous and partly idiomatic constructions (such as *the more... the more*, *for me to VERB CLAUSE BE the work of an instant / moment*) are particularly difficult to extract by means of automated methods.

Extracting phraseology (in the broad sense) from running text is a particularly daunting challenge, as demonstrated by the results of the [Parseme 2018 shared task](#) for the extraction of verbal multiword expressions. The best overall system (TRAVERSAL) taking part in the shared task reached, across the 19 languages, an F1-measure (token-based) of 59.67 percent, but a closer look at individual languages and specific categories of multi-word units reveals a number of weaknesses. Thus, for the English dataset (and in spite of the obviously immense linguistic resources for English), TRAVERSAL only reaches a general F1-score (token-based) of 30.15 percent (with recall as low as 20.33). Besides, the very central category of verbal idioms (as in *paint the town red*) receives with TRAVERSAL an even more disappointing F1-score of 4.22 (with precision at 41.67 but with recall at only 2.22). Of all the English verbal idioms that should have been extracted from the running texts, the best overall system was just able to extract correctly 2.22 percent of them. The very inspiring Parseme 2018 shared task illustrates again how difficult it is to establish a gold set of multi-word units (in this case by two human annotators), and also to derive very idiomatic structures from sparse data.

We have proposed [11] [12] the *cpr-score* for measuring the association strength between n-grams of size 2 to 12. It is an adaptation of metric clusters, based on the average distance between the component grams of an n-gram, measured in large corpora of at least 200,000 tokens. This simple metric makes it possible to extract phraseology in the broad sense, formulaic language or idiomatic constructions, with an acceptable precision and recall.

It is particularly difficult to measure precision and recall for a complete phraseology extraction method in a running text, because it is very hard to establish the gold set of constructions that have to be extracted. As explained in Colson [12], a reliable way of

implementing a general extraction method for phraseology is to test it against Chinese segmentation, as Chinese constructions make less strict differences between (Western) notions of words, collocations or idioms.

The precision and recall scores for segmenting Chinese with the *cpr-score* were measured in Colson [12] and reached an F-measure of 0.70 with one of the standard Chinese datasets. Recall is particularly high with the *cpr-score*: 0.749 for Chinese segmentation, which is almost the average degree of agreement for segmentation between Chinese native speakers (0.75).

The *cpr-score* has also been tested against the English dataset of the Parseme 2018 shared task for the extraction of verbal multiword expressions [13]. It should be noted that the gold set for this task was established by just 2 linguists and displayed a number of debatable cases. For precision, the results yielded by the *cpr-score* were less good than those produced by the best system for English: TRAPACC, based on neural networks [14], but the recall score (token-based) was much better with *cpr*: 0.5225, as opposed to 0.2898 with TRAPACC. These results indicate that there is still room for improvement in any extraction method, but they also highlight the limitation of deep learning (DL). Focusing on recurrent patterns in the training set, DL has a limited recall capacity, while a method such as *cpr*, based on huge linguistic data, does not require any training set and is able to reach high recall scores. The lower precision scores obtained with *cpr* at Parseme also derive from the fact that a much broader view of phraseology is taken, as opposed to the more restricted view of the annotators and the often debatable decisions they made. This partly explains why the *cpr-score* reaches much higher precision and recall scores for Chinese segmentation, as the average agreement among native speakers for segmentation is higher than for multiword units / phraseology / formulaic language / idiomatic constructions, all notions that are largely unknown to the average native speaker.

The *cpr-score* has been implemented in a web application: the *IdiomSearch* tool. As explained in Colson [11], the tool may always be improved, but it can be used for raising the phraseological awareness of language students. It also enables translators to get a general idea of the most frequent multiword units in the source text or in the target text. Experiments with *IdiomSearch* have indicated that this methodology makes it possible to extract phraseology in the broad sense and formulaic language, providing much richer results than those produced by manual extraction on the basis of reference books and stricter criteria. Thus, Dupal [15] showed that *IdiomSearch* confirmed results obtained manually for the difference between CLIL (*Content and Language Integrated Learning*) and non-CLIL students, while providing far more instances of phraseology and formulaic language.

In spite of its shortcomings as an ongoing project, the *IdiomSearch* tool, based on the *cpr-score*, offers, for any given text, a neutral and global view of the presence of phraseology in the broad sense / formulaic language / idiomatic constructions.

We have therefore used the *IdiomSearch* tool for finding an answer to the first research question mentioned above, namely the difference in phraseology between texts produced by NMT on the one hand and original texts on the other.

For our experiment, we checked by means of the *IdiomSearch* tool the average percentage of phraseology (in the broadest sense, including formulaic language and idiomatic constructions) in:

- 10 original articles from English newspapers, on the general topic of Brexit;
- 10 French newspaper articles on Brexit, translated into English by 2 NMT systems (Google Translate and DeepL).

The complete list of the chosen articles can be found in the Appendix at the end of this paper. The texts had to be of a non-specialised nature, and the length of each paper was about the same: 5,000 characters, including blanks, for the French articles. As translation from French into English results in a smaller number of words, the comparable original English articles had to be slightly shorter, and their length was therefore set at 4,500 characters; if necessary, longer texts were slightly shortened.

Table 1 below displays the average percentage of phraseology, as measured by *IdiomSearch*, for the original English newspaper articles. Table 2 shows the same percentage for the 10 French articles translated into English by Google Translate and by DeepL.

Table 1. Average percentage of phraseology in 10 English newspaper articles

English Newspaper Articles	
List	% Phraseology
Article 01	34.28
Article 02	32.22
Article 03	45.39
Article 04	41.41
Article 05	54.67
Article 06	46.65
Article 07	49.06
Article 08	43.73
Article 09	42.94
Article 10	55.23
Average	44.56

As shown by Table 1, the average percentage of phraseology, measured by *IdiomSearch*, varies a little from one article to the other, in spite of the fact that the number of characters per article was strictly the same (4,500) and the general topic identical. These figures suggest a phraseology range from 30 to 55 percent, with just 2 articles above 50 percent. We should take into account the fact that the *cpr-score* displays a margin of error, does not extract all phraseological units, and is based on the presence or absence of those units in corpora. We may then consider that these figures roughly confirm John Sinclair’s idiom principle [16], according to which about 50 percent of any text consists of phraseology in the broad sense.

Table 2. Average percentage of phraseology with Google Translate and DeepL (translations from French into English)

English Translations of French Newspaper Articles		
List	% Phraseology	
	Google Translate	DeepL
Article 01	48.46	47.84
Article 02	47.12	46.62
Article 03	48.07	45.32
Article 04	41.97	40.08
Article 05	55.06	53.73
Article 06	42.15	45.01
Article 07	48.64	46.59
Article 08	39.04	38.06
Article 09	41.10	43.96
Article 10	47.85	48.27
Average	45.95	45.55

The most striking result shown by a comparison between Table 1 and Table 2 is that there is (slightly) more phraseology in the 10 English translations of French articles, produced by both Google Translate and by DeepL, than in 10 original English newspaper articles on the same subject.

This interesting finding confirms that NMT pays a lot of attention to phraseology, as all recurrent sequences of n-grams in multilingual corpora are trained by the models. Table 2 further indicates that, contrary to what might have been expected, there is no significant difference in phraseology (as measured by our methodology) between the translations by Google on the one hand, and by DeepL on the other. Google Translate even seems to be slightly more phraseological than DeepL, with just 3 translations of DeepL being more phraseological than the corresponding translations by Google. The answer to our first research question, on the basis of this experiment, is pretty clear: there is (at least) as much phraseology in the English translations of French texts produced by both Google Translate and by DeepL, as in original English texts belonging to the same domain.

This is therefore an indication of the progress made by NMT with respect to the thorny issue of phraseology: translations produced by Google Translate and by DeepL are so fluid and natural, that the phraseological differences with original texts written by native speakers tend to disappear.

In seeking an answer to our second research question, however, we wish to take a closer look at some problematic cases.

3 Phraseology and NMT: a closer look at problematic examples

As illustrated by the preceding section, phraseology as a whole may be pretty rich in the translations produced by NMT systems, but this is no guarantee that those translations will in the first place fulfil their main role: conveying the correct meaning of the source text.

In order to address our second research question (*Does phraseology still pose a major problem to NMT, and if so why?*), we therefore adopted a more classical case-study methodology. There is indeed (yet) no reliable automated way of checking the semantic adequacy between a source text and a target text.

Barreiro et al. [17] showed that Google Translate was wrong in the translation of phraseology in about 40 percent of the cases. Since then, NMT was introduced into the system, and competing systems such as DeepL have been developed. However, translation practice reveals that there remain many errors in the texts produced by NMT. A survey of all cases of wrong translations, or an extensive overview of the performance of NMT across languages falls beyond the scope of the present contribution.

In addition to automated methodologies as illustrated in the preceding section, we agree with Look [1] that a careful manual analysis of examples that remain problematic for MT is of the essence. In this case, we have randomly selected from the news a number of examples of sentences with rich phraseology, which posed a real problem to both Google Translate and DeepL. The crux of the matter is to determine what are the common features between those very problematic sentences, and what conclusions could be drawn from them.

Example 1: *UK car industry in brace position ahead of Brexit deadline* (The Guardian, 9 August 2019).

Google Translation (English-French): *L'industrie automobile britannique en bonne voie pour la fin du Brexit.*

DeepL Translation (English-French): *L'industrie automobile britannique en position de force avant l'échéance de Brexit.*

Example 1 is an interesting case, because both NMT translations are rich in phraseology, as they contain two nice French idioms (*être en bonne voie*, *être en position de force*). However, both translations are a complete misinterpretation of the source text: the English sentence means that the UK car industry is fearing the worst (and placing itself in a defensive position, as people protecting themselves against an airplane crash). The two French translations, on the contrary, convey the opposite meaning: the UK car industry is right on track (*en bonne voie*) or in a position of power (*en position de force*). Clearly, the NMT systems were unable to train sufficient contexts of the phrase *in brace position*, so that they could not distinguish between positive and negative connotation.

Example 2: *How Trump's economic chickens are finally coming home to roost* (Forbes.com, 13 February 2019).

Google Translation (English-French): *Comment les poulets économiques de Trump sont enfin rentrés à la maison.*

DeepL Translation (English-French): *Comment les poulets économiques de Trump rentrent enfin au bercail.*

In Example 2, we notice, as in example 1, a rich phraseology in both French translations (especially for the very idiomatic expression *rentrer au bercail*, to return home). However, both NMT systems have completely missed the idiomatic meaning of the phrase *the chickens have come home to roost*: some day you have to pay for your past mistakes. The literal translation of this phrase does not make any sense in French. It should be noted that the canonical form of the phrase, *the chickens have come home to roost*, is known to [Linguee](#), the parallel corpora database underlying DeepL, and receives a number of correct French translations (*payer les pots cassés, refaire surface*). What is apparently problematic here is the number of figurative contexts, as well as the variants of the phrase (in this case, a modifying adjective before *chickens*, the adverb *finally* and the present progressive instead of present perfect).

Example 3: *Zwietracht bei der öffentlich-privaten Partnerschaft. Da haben wir gegenüber dem Ballungsraum keine Chance und fallen immer durch den Rost* (Parseme 2018 German dataset, id = . . newscrawl-19).

Google Translation (German-English): *Discord in the public-private partnership. Since we have no chance against the metropolitan area and always fall through the rust.*

DeepL Translation (German-English): *Disagreement in the public-private partnership. We have no chance against the conurbation and always fall through the rust.*

In Example 3, both NMT systems were unable to detect the German idiom *durch den Rost fallen* (be discarded), and therefore produced a wrong literal translation.

Example 4: *Max Buset dacht dat ik er het bijltje bij neer zou leggen”, zegt Eyskens daarover in zijn memoires. “Maar dat was buiten de waard gerekend! Na het herhaalde weigeren van de socialisten en de liberalen besloot ik zonder meer een minderheidsregering te vormen.”* (hln.be, 8 December 2018).

Google Translation (Dutch-English): *Max Buset thought that I would give up, "Eyskens says about this in his memoirs. "But that was not counted! After the repeated refusal of the Socialists and the Liberals, I decided to form a minority government."*

DeepL Translation (Dutch-English): *Max Buset thought I'd put my foot down," says Eyskens in his memoirs. "But that was out of the question! After the repeated refusals of the socialists and the liberals, I decided to form a minority government."*

In Example 4, the Dutch verbal idiom (*buiten de waard rekenen*, which means to make a miscalculation, to overlook sg) was wrongly translated by Google (by means of a literal verb) and also by DeepL. In the case of DeepL, the machine translation is again phraseological (*to be out of the question*) but the meaning is not correct: the sentence means that people who thought that minister Eyskens would call it a day were wrong in thinking so; the sentence does not mean that it was out of the question for him to resign. The Dutch idiom *buiten de waard rekenen* has been ignored (Google) or wrongly interpreted (DeepL) by the training systems of NMT.

Example 5: 相反地，他第一次在大学辍学是为了报读自己更有兴趣的室内设计，第二次辍学则是因为他成功为第一套房子做了室内设计，加上无法适应理工学院的学习环境，所以决定破釜沉舟放手一搏。[*Xiāngfǎn dì, tā dì yī cì zài dàxué chuòxué shì wèile bào dù zìjǐ gèng yǒu xìngqù de shì nèi shèjì, dì èr cì chuòxué zé shì yīnwèi tā chénggōng wèi dì yī tào fángzi zuòle shì nèi shèjì, jiā shàng wúfǎ shìyìng*

lǐgōng xuéyuàn de xuéxí huánjìng, suǒyǐ juédìng pòfúchénzhōu fàngshǒu yī bó.] (zaobao.com, 30 July 2019).

Google Translation: *On the contrary, he first dropped out of college to enroll in the interior design that he was more interested in. The second time he dropped out of school was because he successfully designed the interior of the first house, and could not adapt to the learning environment of the Polytechnic. So, I decided to let go and let go.*

In Example 5, the last sentence is wrongly translated by NMT (in this case, only Google because DeepL does not offer Chinese). In addition to the wrong pronoun (*I* instead of *he*), the Chinese phrase 破釜沉舟 [*pò fǔ chén zhōu*] (a case of *chéngyǔ*, a sort of idiomatic expression linked to Chinese culture), literally “break cauldrons sink boats”, means *to cross the Rubicon, to make a big decision*. The correct translation could have been “*so he decided to take a big step and call it a day*”.

It cannot be over-emphasized that, in cases such as Examples 1 to 5, NMT often creates the illusion of phraseologically correct translations, while being completely wrong. Far from being exceptions, such examples are easy to find in any domain, and in all languages. It may be pointed out that NMT is constantly being improved, and that the models will get even better in the future.

While this remains an open question, we would like to report the results obtained by our corpus-based approach in cases such as Examples 1-5. Contrary to what might be inferred from the poor results of NMT in those cases, the phraseological units used in the examples can be easily extracted from any large corpus by means of the *Idiom-Search* tool and the *cpr-score* mentioned in section 2, as shown in table 3.

Table 3. Association scores and frequency for the phrases in Examples 1-5

Association scores for phrases in Examples 1-5		
List	cpr	Freq
<i>brace position</i>	0.25	10
<i>the chickens have come home to roost</i>	0.73	8
<i>durch den Rost fallen</i>	0.81	29
<i>buiten de waard gerekend</i>	0.73	41
破釜沉舟 [<i>pò fǔ chén zhōu</i>]	1.00	45.55

As shown in table 3, the problematic phrases that were wrongly translated by the NMT systems under Examples 1-5 could be easily identified by the *cpr-score* and large corpora of about 1 billion tokens. The question then arises why such cases remain problematic to NMT.

In addition to the results obtained with the Parseme 2018 shared task mentioned in section 2, our data suggest that neural networks have difficulties in learning from training data multiword units with a low frequency. Future work might therefore consider a hybrid approach, in which the research carried out within the framework of computational and corpus-based phraseology is taken into account.

4 Conclusion

While current NMT systems reach an impressive percentage of phraseology in the translations produced, the tip of the phraseological iceberg remains problematic for them. Our examples show that complex contextual features are hard to cope with in the NMT models, and that very idiomatic but fairly infrequent combinations may fall through the cracks.

A combination of NMT and a corpus-based approach may in the future shed new light on the feasibility of solving most phraseological issues. In the meantime, the methodology presented here for checking the global phraseological percentage, as well as the phraseological character of specific combinations, may be used in the manual post-editing by professional translators.

References

1. Loock, R.: Traduction automatique et usage linguistique : une analyse de traductions anglais-français réunies en corpus. *Meta, Journal des traducteurs / Translators' Journal*, 63(3):786-806. (2018) <https://doi.org/10.7202/1060173ar>.
2. Papineni, K., Roukos, S., Ward, T., et al. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp 311-318 (2002)
3. Denkowski, M., Lavie, A.: Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, pp 376-380 (2014)
4. Isabelle, P., Cherry, C., Foster, G.: A Challenge Set Approach to Evaluating Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp 2486-2496 (2017)
5. Laviosa, S.: *Corpus-Based Translation Studies: Theory, Findings, Applications*. Rodopi, Amsterdam/New York (2002)
6. Burger, H., Dobrovolskij, D., Kühn, P., Norrick, N.: (eds.) *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research*. De Gruyter, Berlin / New York (2007)
7. Wray, A.: *Formulaic Language: Pushing the Boundaries*. Oxford University Press, Oxford (2008)
8. Croft, W.: *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford (2001)
9. Goldberg, A.: *Constructions at Work*. Oxford University Press, Oxford (2006)
10. Hoffmann, T., Trousdale, G.: (eds.) *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford/New York (2013)
11. Colson, J.-P.: The IdiomSearch Experiment: Extracting Phraseology from a Probabilistic Network of Constructions. In Mitkov R (ed.), *Computational and Corpus-based phraseology, Lecture Notes in Artificial Intelligence 10596*. Springer International Publishing, Cham, pp 16-28 (2017)
12. Colson, J.-P.: From Chinese Word Segmentation to Extraction of Constructions: Two Sides of the Same Algorithmic Coin. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), ACL*, pp 41-50 (2018)

13. Savary, A. Ramisch, C., Hwang, J.D., Schneider, N., Andresen, M., Pradhan, S., Petruck, M.R.L.: (eds.) Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018). ACL, Santa Fe (2018)
14. Stodden, R., QasemiZadeh, B., Kallmeyer, L.: TRAPACC and TRAPACCS at PARSEME Shared Task 2018: Neural Transition Tagging of Verbal Multiword Expressions. In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), ACL, pp 268-274 (2018)
15. Dupal, J.: Investigating the Phrasicon of CLIL and NON-CLIL Students : a corpus-based comparative analysis using IdiomSearch. Thesis, Université catholique de Louvain, Louvain-la-Neuve (2018)
16. Sinclair, J.: Corpus, concordance, collocation. Oxford University Press, Oxford (1991)
17. Barreiro, A., Monti, J., Batista, F., Orliac, B.: When multiword go bad in machine translation. In Proceedings of the workshop on multi-word units in machine translation and translation technologies, 14th Machine Translation Summit, Nice (2013)

Appendix: List of articles

French articles:

- Brexit : le chaos et des pénuries prévisibles au Royaume-Uni en cas de « no deal » (Le Monde, 19 August 2019)
- Brexit : Boris Johnson prépare l'UE et le Royaume-Uni à un « no deal » (Le Monde, 21 August 2019)
- Brexit, « gilets jaunes », sommet du G7..., Macron aborde les dossiers de la rentrée (Le Monde, 21 August 2019)
- Brexit : à Londres, la bataille du « no deal » a commencé (Le Monde, 28 August 2019)
- Suspension du Parlement britannique : le coup de poker de Boris Johnson (Libération, 28 August 2019)
- Brexit : l'« outrage constitutionnel » de Boris Johnson (Le Monde, 29 August 2019)
- Royaume-Uni : « La piste d'un gouvernement d'union nationale semble de moins en moins plausible » (Le Monde, 29 August 2019)
- En suspendant le Parlement britannique, Boris Johnson aggrave la crise du Brexit (Le Monde, 29 August 2019)
- Parlement suspendu : Johnson prend son temps (Libération, 29 August 2019)
- Les Pays-Bas tirent profit de « l'effet Brexit » (Le Monde, 30 August 2019)

English articles :

- Eurocrats know Boris Johnson well, making no-deal Brexit more likely (The Economist, 15 August 2019)
- What Britain's release of an Iranian tanker says about its post-Brexit foreign policy (The Economist, 22 August 2019)
- Britain to be 'energetic partner' after Brexit (The Guardian, 23 August 2019)
- Johnson Walks Tightrope at G7, as Trump Pledges 'Very Big Trade Deal' for U.K. (The New York Times, 25 August 2019)
- Jeremy Corbyn agrees to prioritise legislation to stop no-deal Brexit (The Guardian, 27 August 2019)

- Boris Johnson is trashing the democracy fought for with the blood of our ancestors (The Guardian, 28 August 2019)
- Johnson has wrongfooted opponents of no deal. An election surely looms (The Guardian, 28 August 2019)
- Boris Johnson suspends Parliament, causing uproar (The Economist, 29 August 2019)
- Has Brexit destroyed party loyalty? Ruth Davidson's resignation will show us (The Guardian, 29 August 2019)
- Boris Johnson's Parliament Suspension Prompts Growing Backlash (The New York Times, 29 August 2019)

Automatic Term Extraction from Turkish to English Medical Corpus

Gökhan Doğru¹

Universitat Autònoma de Barcelona
08193, Barcelona, Spain
gokhan.dogru@uab.cat

Abstract. This study aims to evaluate the state-of-the-art automatic and semi-automatic term extraction from a domain-specific bilingual corpus in Turkish - English language pair. Three different tools (a computer-assisted translation tool, a web-based corpus analysis toolkit, and a desktop corpus analysis tool) are used for extracting Turkish cardiology single-word and multi-word candidate terms with different parameters, and the results are compared in terms of number of candidate term counts. It has been observed that each tool responds to different needs of translators and comes with a limited number of customization options. It is concluded that while monolingual term extraction is useful for translators and terminologists, there is still no tool providing bilingual candidate term extraction for Turkish – English language pair.

1 Introduction

Effective terminology management is one of the pillars of accurate, fast and consistent translation, a fact illustrated by the inclusion of terminology as one of the core categories of the widely used Multidimensional Quality Metrics (MQM) for translation quality assessment. Among other things, with the ubiquitous use of computer assisted translation tools and the necessity to have more than one translator in the workflow of translation require translation companies and other translation stakeholders to plan the terminology to be used across the projects. This planning phase is crucial to avoid any future inconsistency or inaccuracy in the use of professional terminology of a specific domain. Depending on the availability of previously translated documents (in the form of parallel corpora or comparable corpora), different methods of terminology management strategies including manual and automatic term extraction are implemented. In this study we report our results regarding automatic (candidate) term extraction from a Turkish- English bilingual cardiology corpus, the tools and methods used to process the term candidates and select actual terms, and the problems we have encountered.

2 Terminology Management and Automatic Term Extraction

Specialized domain translators may participate in ongoing translation projects in different time frames. Namely, although they may participate in the beginning of a project, they may also join in the second or third year to an ongoing project, e.g. a Spanish to Turkish medical device user manual translation with content updates regularly. In such a case, they should be provided by bilingual terminology lists (or glossary databases) prepared by terminologists or previous translators to be used within the preferred CAT tool. Especially, if many translators are to collaborate within the same project, these lists are particularly important to provide consistency. When such lists are not provided, translators need to prepare them by themselves as fast as possible to make their translation flow more efficient or at least, the translation company shall prepare it for them (preferably, by a terminology expert). But as Heylen and Hertog[1] observes, “(...) in a rapidly changing world with an ever-growing technical vocabulary, the manual maintenance, or in the case of new technological fields, the manual exploration, indexation and description of a domain’s core vocabulary is a labor-intensive enterprise.” Hence, bilingual and monolingual automatic term extraction methods have been suggested to make this preparation phase more agile and cost effective. These methods may be statistical, linguistic or hybrid depending on the tools used. We can also assume that more and more machine learning methods will be used for this process, yet it is possible to qualify machine learning methods under the category of statistical methods. Nevertheless, it should be stated that the results of the automatic extraction methods (just like the output of machine translation) are far from being perfect and as Ahmad and Rogers[2] emphasizes, “Term extraction produces the raw material for terminology databases: this raw material has to be examined, tested and validated in some way before inclusion in a terminology database.” For this reason, the resulting lists of terms are called “candidate terms”, not “terms” per se. Therefore, the complete process of term extraction is still a semi-automatic process, which needs human intervention. Although nearly two decades have passed since the article by Ahmad and Rogers[2] and new milestones have been reached in terminology extraction in different language pairs, this observation for automatic term extraction is still valid. In this time framework, terminology databases have become an inseparable component of computer assisted translation tools and more and more CAT tools are including automatic term extraction feature to their software. Besides, new use cases have been introduced for terminology databases including use in automatic translation quality assurance as well as machine translation training.

3 Methods and Tools

There are different software tools that allow for automatic candidate term extraction with or without the possibility to later edit or validate the candidates

in-situ. We have used three different tools for term extraction based on statistical frequency: MemoQ, a proprietary CAT Tool with term extraction and editing feature; AntConc, a stand-alone freeware corpus analysis toolkit; and Sketch Engine, a web-based corpus query and management system. We have used these tools to process a corpus consisting of cardiology journal abstracts in Turkish and their translations into English. The results of each tool are reported in the following sections. Among these tools, MemoQ has been used in the corpus building phase as well since it also includes features both for text parsing, translation memory and terminology management features.

4 Bilingual Corpus Use for Term Extraction

The rise of automatic term extraction is highly tied to the developments in electronic corpus studies. The tools and procedures used in monolingual and bilingual corpora preparation have made it easier to prepare domain specific corpora from which terminology can be extracted with certain levels of success. In translation world, the most common type of corpora are translation memories which include source language strings and target language strings together with some metadata including, date, domain, translator etc. These translation memories are saved and interchanged across different CAT tools in translation memory exchange format (TMX). Hence, in order to decrease the requirement for translators to use different stand-alone tools to realize term extraction tasks and avoid different file format exchanges (import/export) between terminology tools and translation tools, there is now a tendency to integrate term extraction capabilities into CAT platforms. One good example is MemoQ, a desktop CAT tool which integrates an automatic term extraction feature into the workflow of the translator within a project. In our study, we have a Turkish to English cardiology corpus prepared in the form of a translation memory. Using this corpus, we examine automatic term extraction tools and methods and obtain Turkish to English cardiology terms to be used in statistical machine translation. However, as we have mentioned above, these terms can also be used in translation projects to make translations faster, more consistent and of higher quality.

5 Term Extraction from Turkish to English Medical Corpora

Translating terminology properly and consistently is very vital in medical texts. Hence, the availability of bilingual terminological database (shortly, “termbase”) is crucial in translation projects. Before describing our corpus, we have also investigated how we can ensure the quality or credibility of a termbase. Reynolds[3], in a professional translation context, divides terminology credibility into three: “a). High: “i). Terminology which the customer has specified should be used, ii). Terminology received from customer or acknowledged experts in the field terminology provided by the customer; b). Medium: i). Terminology verified by other

professional translators., ii). The more translators which agree that a particular term is correct, the more likely it is to be correct; Low: i). Terminology extracted using software tool, ii). Terminology received from other sources but not verified by the customer, industry experts or other translators”. Below it will be seen that the corpora that we have prepared is derived from a cardiology journal with abstracts in Turkish translated to English. Since these abstracts and their translations pass from a peer review (including experts of the same domain) before publishing, we assume that the corpus terminology has a high terminology credibility according to the classification of Reynolds[3]. Although we use software for extraction of candidate terms, we also validate them manually.

5.1 Corpus Preparation

We have used different tools and methods to create a domain specific bilingual corpus. We build it from the abstracts published in *Archives of the Turkish Society of Cardiology* which “is a peer-reviewed journal that covers all aspects of cardiovascular medicine. The journal publishes original clinical and experimental research articles, case reports, reviews and interesting images pertinent to cardiovascular diseases, as well as editorial comments, letters to the editor, news, guidelines, and abstracts presented at national cardiology meetings.”¹ Its topics include “coronary artery disease, valve diseases, arrhythmia’s, heart failure, hypertension, congenital heart diseases, cardiovascular surgery, basic science and imaging techniques.”² The journal’s online archive dates back to 1990 and the current issue is Volume: 47 Issue: 3 - April 2019. Most importantly for our purpose, nearly all the abstracts are translated into English. Considering that the journal keeps being published for nearly three decades and that it includes scientific articles from cardiology domain, we can argue that it covers a significant portion of the Turkish cardiology terminology and, through the translations of the abstracts, the English counterparts of the terms as perceived by Turkish cardiology specialists. One of our purposes has been to truly represent the termbase available in this journal archive about the abstracts and their translations. The abstracts have a consistent format. Original articles include subheadings of “objective”, “method”, “results”, “conclusions” and “keywords” (which is a valuable source for term extraction) while case report abstracts are shorter, have a keyword section and do not include subheading. Since the website has a well-structured HTML design, it has been possible to crawl all the abstracts (both in Turkish and English) in HTML format and save them locally for further processing. Then, we were able to convert the HTML files into plain text files using MemoQ’s regex text filter (one custom filter for Turkish and a custom filter for English), and align them to build a translation memory.

¹ *Archives of the Turkish Society of Cardiology website:*
<http://www.archivestsc.com/about-the-journal> (last access: 29.04.2019);

² Ibid.

The Turkish English Corpus in Numbers. The characteristics of our corpus are given in Table 1. Our corpus has 474,273 source language (SL) words and 542,783 target language (TL) words (Since Turkish is an agglutinative language with lots of suffixes added at the end of words, there is a 14,4 percent increase in the number of words in English). Ahmad and Rogers[2] suggest having a corpus of nearly 100,000 words “as a good starting point for corpus-based terminology management in a highly-specialized discipline” (p. 593) and later add that “As a rule of thumb, special-language corpora already start to become useful for key terms of the domain in the tens of thousands of words, rather than the millions of words required for general-language lexicography” (p. 594). Having 474,273 SL words and 57,368 unique SL words, our corpus seems to be satisfying the size criteria for terminology-oriented corpus. When we look at unique SL and TL word forms, we see a pattern similar to the total SL and TL word counts. While there are 57368 unique word forms in SL, there are 35.844 word-forms in TL³. For comparing how different corpus analysis tools show counts, we have also made a comparison of frequency and word counts in AntConc and Sketch Engine since how a word is defined can be different across tools and default tool settings.

Table 1. The profile of the corpus according to MemoQ.

Language pair	Turkish - English
Domain (field)	Medicine
Discipline	Internal medicine
Subdiscipline	Cardiology
UNESCO code	3205.01
Number of source words	474.273
Number of target words	542.783
Number of unique source words	57.368
Number of unique target words	35.844

A comparison of word and frequency counts is given below. It can be observed that each tool treats differently the concept of word; hence, each one yields different counts for both total word count and unique word count. Since results of one-word lists include around 40,000 words in each tool and there can be up to 5-word terms in this corpus, it is obvious that such a scenario is not terminologist-friendly and that it will be very time consuming to create a final terminology list. In the following sections, we will explain how we have constrained our term

³ Unique word (form) counts are realized in Memoq’s term extraction feature. Minimum frequency and maximum number of words per term are defined as “1” so that only unique words are extracted. For the initial analysis, any one or more letters are considered word forms. Of course, the term “term” is a different concept and it will be elaborated in the following sections.

Table 2. Comparison of word and frequency counts in 3 tools. *Word with number are ignored. **Non-words (“tokens which do not start with a letter of the alphabet.”)⁴

	MemoQ	AntConc	Sketch Engine
No. of source words	474273	489155	479,200
No. of target words	542783	557054	523,077
No. of unique source words	40656	42836	38800
No. of unique target words	20802*	18309	23326**

extraction parameters in each tool to create a noise-free (or with minimum noise possible) candidate term list.

Constraining Parameters for Automatic Term Extraction. The genre of our corpus has a unique advantage: by nature, scientific abstracts include a section called “keywords” where the author(s) add(s) the most relevant keywords in their study. In each abstract, these keywords section is provided in a sentence.

The screenshot shows the 'PARALLEL CONCORDANCE' interface in Sketch Engine. The search criteria are 'Medical Cardiology, English'. The interface displays a list of 46 keywords in English and their corresponding Turkish translations. The keywords are listed in two columns, with the English version on the left and the Turkish version on the right. The interface includes a search bar, a list of keywords, and a sidebar with navigation options.

English Keywords	Turkish Keywords
<=> Keywords: Criss-cross heart, dextrocardia, transposition of the great arteries </=>	<=> Nadir bir patoloji Anahtar Kelimeler: Criss-cross kalp, dektrokardi, büyük arterlerin transpozisyonu </=>
<=> Keywords: Arteriovenous fistula/diagnosis/radiography; coronary vessel anomalies/diagnosis/radiography; tomography, X-ray computed/methods </=>	<=> Anahtar Kelimeler: Arteriyovenöz fistül/tanı/radyografi; koroner damar anomalisi/tanı/radyografi; bilgisayarlı tomografi/ yöntem </=>
<=> Keywords: Arrhythmia, loss of consciousness, syncope, elderly. </=>	<=> Anahtar Kelimeler: Aritmi, bilinç kaybı, senkop, yaşlılık. </=>
<=> Keywords: WPW syndrome, algorithm, ECG, ablation </=>	<=> Anahtar Kelimeler: WPW sendromu, algoritim, EKG, ablasyon </=>
<=> Keywords: Adult, foramen ovale, patent/therapy, heart catheterization; heart septal defects, atrial/therapy; septal occluder device </=>	<=> Anahtar Kelimeler: Erişkin, foramen ovale açıklığı/tedavi, kalp kateterizasyonu; kalp septal defekti, atriyal/tedavi; septal tıkaçıcı cihaz </=>
<=> Keywords: Angioplasty, transluminal, percutaneous coronary; balloon dilatation/instrumentation/methods; catheterization/ instrumentation; coronary angiography; embolism/etiology/prevention & control; equipment design; filtration; myocardial infarction; stents </=>	<=> Anahtar Kelimeler: Anjiyoplasti, transluminal, perkütan koroner; balonla dilatasyon/enstrümantasyon/yöntem; kateterizasyon/enstrümantasyon; koroner anjiyografi; embolizm/etiyolojik önleme ve kontrol; ekipman tasarımı; filtreleme; miyokard infarküsü; stent </=>
<=> Keywords: Cardiovascular disease, elderly. </=>	<=> Anahtar Kelimeler: Kardiyovasküler hastalık, yaşlılık. </=>
<=> Keywords: Cardiovascular disease, comorbidities, elderly. </=>	<=> Anahtar Kelimeler: Kardiyovasküler hastalık, komorbiditeler, yaşlı hasta. </=>
<=> Keywords: Acute coronary syndrome, aged; coronary angiography; coronary artery disease; hemoglobins/metabolism; erythrocyte indices; inflammation; leukocytes. </=>	<=> Anahtar Kelimeler: Akut koroner sendrom, yaş; koroner anjiyografi; koroner arter hastalığı; hemoglobini/metabolizma; eritrosit indeksi; enflamasyon; lökosit. </=>
<=> Keywords: Child, cyanosis/etiology; graft occlusion, vascular/therapy; heart defects, congenital pulmonary circulation; stents </=>	<=> Anahtar Kelimeler: Çocuk, siyanoz/etiyoloji; greft tıkanması, vasküler/tedavi; kalp defekti, doğuştan; pulmoner dolaşım; stent </=>

Fig. 1. Parallel Concordance in Sketch Engine.

In Sketch Engine, we have been able to filter these sentences and sort them side by side in Turkish and English version (called parallel concordance). A sample of this process is shown in the Figure 1. Then we could export all these

⁴ https://www.sketchengine.eu/my_keywords/non-word (last access: 15.05.2019)

sentences to an excel file and then manually create a terminology list in cardiology domain. This bilingual list is very comprehensive because it includes most of the important terms inside the corpus. There are 2951 one-word and multi-word terms. And since these terms are not extracted but provided by cardiology experts, we can argue that they are reliable. We have used this manual list to compare the results of automatic extraction features of each tool.

6 Term Extraction from Turkish to English Medical Corpora

We have three different results for each tool. It should be stated that none of these tools allows for bilingual automatic extraction for Turkish and English language pair. Sketch Engine has this option for a few languages such as “Chinese, Czech, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish” as stated by Baisa[4] *et al.* but Turkish is not included. We believe that the addition of Turkish will make Turkish – English bilingual term extraction very fast and time-saving. In the sections below, we report our results for each tool. It will be observed that their different parameters and available features provide varying number of candidate one-word and multi-word candidate terms. It shall be emphasized that extracting more terms does not necessarily mean that the resulting list will be more useful for the translators since removing words that do not constitute terms can be time-consuming as well.

6.1 Term Extraction with MemoQ

MemoQ extracts multi-word and single-word terms monolingually based on frequency and confidence score and it is also possible to lookup the target terms through other active (reference) termbase lists. In our setting, we have the above-mentioned reference cardiology terminology to compare our results. Besides, we have realized that stop word lists are very important to avoid noise in term candidates and that MemoQ does not have a stop word list for Turkish. Stop word lists helps to filter out the words or groups of words that are frequent but are not actually terms. Hence, based on the initial candidate term results we have created a domain-specific stop word list. In other words, our stop word lists can be reused when extracting terms from medical corpora. We have made it openly available⁵.

Setting 1: Multiword terms: Maximum 5 words, minimum 5 times frequency, Single-word terms: Minimum character length 4, minimum 5 times frequency; words with number ignored, term lookup is not active; stop word list is not used.

Setup 2: Multiword terms: Maximum 5 words, minimum 5 times frequency, Single-word terms: Minimum character length 4, minimum 5 times frequency;

⁵ Stop Word List for Turkish language, <https://github.com/gokhandogru/stopwordstforturkish> (last access: 10/06/2019)

Table 3. MemoQ Candidate Term Extraction with Different Parameters.

Type of Extraction	Candidate Term Count	Full Match Term Lookup Count	Partial Match Term Lookup Count
Without Stop Word List	21557	334	5401
With a Stop Word List	13219	317	3631
Only Multi-Word Terms	5901	41	3088

words with number ignored, term lookup is active; stop word list is crafted and used.

Setup 3: Multiword terms: Maximum 5 words, minimum 5 times frequency, Single-word terms are ignored; words with number ignored, term lookup is active; stop word list is crafted and used.

6.2 Term Extraction with AntConc

AntConc allows for using a general monolingual corpus to compare with a domain specific corpus so that word(s) that occur more frequently in domain specific corpus and less frequently in the domain general corpus can be calculated as keywords or terms. We have used Turkish Wikipedia Corpus in OPUS Corpus as a reference corpus. This corpus has 4.7 million tokens. However, it has only been possible to extract single-word terms (“keywords”) in this setting. We have yielded 5701 (candidate) keywords. For multi-word terms, we have tried the “Clusters/N-gram” feature with a setting of minimum frequency of 5 for terms with the size of 2 – 5 words. The system has yielded 20821 multi-word terms. However, this setting does not include a comparison with the reference corpus. Hence, it is only frequency-based and very noisy.

6.3 Term Extraction with Sketch Engine

Sketch Engine differentiates keywords from terms. Keywords are “individual words (tokens) which appear more frequently in the focus corpus than in the reference corpus” while terms are “multi-word expressions which appear more frequently in the focus corpus than in the reference corpus and, additionally, match the typical format of terminology in the language.” Although this distinction is not that clear in Translation Studies and in Translation Technologies Studies, we will make our analysis with this distinction in mind. In the keyword extraction, Sketch Engine has yielded 7862 keywords. The reference corpus that they use is called Turkish Web 2012 and it includes more than 3 billion words. And the keywords are highly precise. As of June 2019, the Turkish multi-word term extraction has not been possible in Sketch Engine.

7 Results

Each tool has yielded a similar number of single-word term while the multi-word counts have differed from 0 to 20821 depending on the available configurations. On average, we have obtained 6960 one-word terms per tool. When it comes to Turkish multi-word terms, each tool has behaved differently. Firstly, Sketch Engine does not support multi-word term extraction for Turkish. We have experimented with the English corpus and the results have been very adequate. It will be very useful to have Turkish multi-word extraction option as well. While it is possible to extract multi-word terms with “Clusters/N-gram” feature, the results are too noisy because of the lack of a comparison with a reference corpus or linguistic normalization / filtering of the resulting noisy candidate multi-word terms. Excessive number of candidate terms is not going to be useful compared to manual term extraction. Lastly, Memoq has yielded a balanced number of multi-word terms after we have crafted a custom stop word list. The results given in the Memoq row in Table 4 reflects the extraction after the use of our stop word list. In all three tools, some of the Turkish terms are inflected, in other words, they include inflectional suffixes which shall be removed manually to lemmatize the terms before inclusion in the final term database. This remains as a problem to be solved in each tool. A lemmatization step in the automatic extraction stage can be a solution. The biggest productivity gain for translator can be achieved when bilingual candidate term extraction becomes possible. For now, none of these tools allow production-level Turkish-English or English-Turkish bilingual term extraction. Considering the increasing amount of translation between these two languages and the increasing need of collaborative translation, new techniques shall be developed to extract bilingual candidate terms.

Table 4. Monolingual one-word and multi-word term extraction in three tools.

	One-word terms	Multi-word terms
MemoQ	7318	5901
AntConc	5701	20821
Sketch Engine	7862	0

8 Conclusion

Turkish is a morphologically rich language with lots of suffixes in the end of the word roots, which has resulted in lots of noise in term extraction in all three tools that we have used since they all use conventional statistical methods. We think that the use of stop word lists for Turkish, growing use of deep machine learning (including neural networks) methods in term extraction as well as better strategies of lemmatization of Turkish words and multi-word units are going

to lead to better monolingual Turkish term extraction, which, in turn, will make bilingual term extraction possible for language pairs including Turkish. The new tools and techniques in natural language processing including neural machine translation provide an opportunity for bilingual term extraction and their integration into CAT tools can give translators an optimum work environment in terms of terminology.

Acknowledgements

This work has been funded and supported by the grant of AGAUR FI-2017.

References

1. Heylen, K., De Hertog, D.: Automatic Term Extraction. In: Kockaert, H. J., Steurs, F. (eds.) *Handbook of Terminology*, Vol. 1, pp. 203-221. John Benjamins Publishing Company, Amsterdam (2015).
2. Ahmad, K., Rogers, M.: Corpus Linguistics and Terminology Extraction. In: Wright, S. E., Budin, G. (eds.) *Handbook of Terminology Management*, pp. 725-760. John Benjamins Publishing Company, Amsterdam (2001).
3. Reynolds, P.: Machine translation, translation memory and terminology management. In: Kockaert, H. J., Steurs, F. (eds.) *Handbook of Terminology*, Vol.1, pp. 276-287. John Benjamins Publishing Company, Amsterdam (2015).
4. Baisa, V., Ulipová, B., Cukr, M.: Bilingual Terminology Extraction in Sketch Engine. In: Horák, A., Rychlý, P., Rambousek, A. (eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2015*, pp. 61-67. (2015)
5. Multidimensional Quality Metrics (MQM). <http://www.qt21.eu/mqm-definition/definition-2015-06-16.htmlquality-terms>, last accessed 2019/04/25.
6. MemoQ Homepage, <https://www.memoq.com>, last accessed 2019/04/25..
7. AntConc Homepage, <https://www.laurenceanthony.net/software/antconc/>, last accessed 2019/04/25.
8. Sketch Engine Homepage, <https://www.sketchengine.eu/>, last accessed 2019/04/25.
9. Archives of the Turkish Society of Cardiology Homepage:<http://www.archivestsc.com/about-the-journal>, last accessed 2019/04/25
10. Opus Corpus Homepage, <http://http://opus.nlpl.eu/>, last accessed 2019/04/25

Lexicographic Criteria for Selecting Multiword Units for MT Lexicons

Jack Halpern¹

The CJK Dictionary Institute, Inc.
34-14-2, Tohoku, Niiza-shi, Saitama, Japan
jack@cjki.org

Abstract. A basic assumption in bilingual lexicography and machine translation (MT) is that the linguistic units of one language correspond to those of another language. But even in close language pairs, such as Spanish and English, there are numerous exceptions, while in some language pairs, such as English and Japanese, cross-linguistic lexical anisomorphism is so great that it becomes literally impossible to map certain words and phrases across these languages. This is especially true of linguistic units that consists of multiple components, or *multiword units* (MWUs). The recognition and accurate translation of MWUs play a critical role in enhancing the quality of machine translation[9]. In spite of the recent advances in MT based on neural networks (NMT), MWUs still present major challenges to MT technology.

This paper discusses the fundamental principles for identifying and selecting MWUs for inclusion in bilingual dictionaries, both for humans and for MT systems (MT lexicons). It attempts to define the various subtypes of MWU based on lexicographic principles derived from extensive experience in bilingual lexicography, especially the compilation of a large-scale full-form lexicon for Spanish-English MT. It also introduces some large-scale resources designed to significantly enhance the translation accuracy of multiword proper nouns.

1 Introduction

1.1 The problem

The fundamental principle of selecting headwords in bilingual dictionaries is that *the words and phrases of one language can be mapped to those of another*. This is mostly true, but even in such relatively close language pairs as Spanish-English there are numerous exceptions. In some language pairs, such as English-Japanese or English-Chinese, lexical anisomorphism is so great that the principle of cross-lingual word/phrase correspondence often breaks down completely. That is, it becomes literally impossible to directly map certain linguistic units to those of the other language.

A string of words can be segmented into components in multiple ways. Linguists may disagree on how to combine these components to form meaningful

linguistic units. The challenge is to decide which combination of components qualifies as a lexical unit or dictionary headword. This paper attempts to define the various linguistic units that fall under the broad category of MWUs. These definitions are based on decades of experience in CJK lexicography, and on the compilation of a large-scale full-form Spanish-English lexicon (tens of millions of entries) for the Context-Based Machine Translation project[4] headed by Dr. Jaime Carbonell (MT expert and founder of the Language Technologies Institute at CMU).

1.2 Terminology

This paper defines and illustrates MWUs and their five subtypes, describing the criteria that qualify an MWU for inclusion as an entry in bilingual dictionaries. Intentionally, the terms "word" and "phrase" are avoided as formal categories because of their inherent ambiguity[7]. "Phrase," which can loosely refer to MWUs, is ambiguous and causes much confusion in language technology. It is often used in the sense of "any sequence of two or more words," or loosely in the sense of "compound word," without defining the relation between the components. However, for the sake of brevity, "word" below is used in the sense of *orthographic word*, defined as "an uninterrupted string of letters which is preceded by a blank space and followed either by a blank space or a punctuation mark"[10]. Two terms play a major role in defining MWU subtypes:

1. **Lexical unit** (or *lexical item*) is single free form ("word") or meaningful sequence of free forms or bound forms ("word elements") that constitute the basic elements of a language's lexicon. It is a distinctive unit of vocabulary that associates meaning with form. It is what the native speaker stores, or potentially stores, as a "word" or "phrase" in her internal lexicon, e.g. *house*, *in other words*, *take off*, *rain cats and dogs*, *unmarried*, *high school*, *headwaiter*. The near-synonym lexeme emphasizes all the members of an inflectional paradigm, rather than a specific wordform.

2. **Lexical status** refers to whether an MWU is a *meaningful* lexical unit (has a high degree of lexicalization); that is, whether it is (potentially) present in the internal lexicon of native speakers and functions as a meaningful syntactical/grammatical unit. On the whole, native speaker's intuitively feel that it is "a word or phrase" of their language. Thus *high school*, which is fully lexicalized, has lexical status, but *high building*, a free combination of words, does not.

1.3 MWU Subtypes

A **multiword unit** (MWU) is a combination of two or more words that commonly occur together. They may or may not function as a lexical unit, may or may not be semantically compositional, and may or may not have lexical status. This paper defines and analyzes five subtypes of MWUs.

1. **Multiword expression** (MWE): a lexical unit consisting of two or more words that together function as a single lexical unit.

2. **Free word combination (FWC)**: a meaningful free sequence of words that follow the rules of syntax but has no lexical status.

3. **Phrasal**: a recurrent meaningful free combination of words that has no lexical status in the source language but corresponds to a lexical unit in the target language.

4. **Collocation**: a recurrent combination of words co-occurring more often than by chance whose meaning is (mostly) compositional and transparent.

5. **Multiword proper noun**: a combination of two or more words that together function as a single proper noun.

Although the terms defined herein are based on morphological and lexicographic considerations, different linguists use these terms in somewhat different ways. It should be noted that the subtype categories defined, by their nature, are not necessarily rigorous, nor are they necessarily mutually exclusive.

2 Multiword Expressions

2.1 Definition

A **multiword expression (MWE)** is defined by linguists in different ways. Calzolari et al.[3] gives a general definition as “a sequence of words that acts as a single unit at some level of linguistic analysis.” In *Introduction to the special issue on multiword expressions*, Villavicencio et al.[10] define it as “an expression for which the syntactic or semantic properties of the whole expression cannot be derived from its parts,” while Sag et al.[8] define it “very roughly” as “idiosyncratic interpretations that cross word boundaries (or spaces).” Here it is defined as “a lexical unit consisting of two or more simplex words that together function as a single meaningful lexical unit.”

Table 1. Additional characteristics of MWEs.

a) <i>zona residencial</i>	residential zone (transparent compositional compound)
b) <i>dar a</i>	look out onto (opaque non-compositional phrasal verb)
c) <i>elefante blanco</i>	look out onto (opaque non-compositional phrasal verb)
d) <i>devanarse los sesos</i>	rack one’s brains over (idiomatic expression)
e) <i>matar dos pájaros de un tiro</i>	kill two birds with one stone (opaque compositional proverb)
f) <i>lo antes posible</i>	as soon as possible (locution)

MWEs have some additional characteristics: (a) they represent both content words and function words, (b) they have full lexical and lexicographic status,

(c) some are monolingually compositional but bilingually non-compositional, (d) they can range from semantically transparent to opaque, and (e) they have high semantic cohesiveness. Some examples (see Table 1).

2.2 Analysis

It is important to understand MWEs with some precision, and to distinguish them from FWCs and phrasets, as difficult as this may be in the case of borderline cases. MWEs are groups of words that co-occur more frequently than by chance, have a high semantic cohesiveness (attraction between components) and, *most importantly*, represent a concept, often a well established designatum. They are the core backbone of a language, what native speakers intuitively feel are “the words and phrases” of their language.

2.3 Typology

MWEs can be classified into eight (or more) subtypes. Though on the whole the subtypes are mutually exclusive, there is some overlap between them.

1. **Compound words** are combinations of two or more words (free morphemes) or word elements (bound morphemes) that together function as single lexical item, usually transparent, like *learner’s dictionary* (noun compound) or *take into account* (verbal compound). If they are opaque, they are normally called idioms. Some examples (see Table 2):

Table 2. Compound words.

<i>tinta china</i>	india ink
<i>parada general</i>	general strike
<i>lobo marino</i>	sea lion
<i>caja fuerte</i>	safe, strong box
<i>casa de campo</i>	field house
<i>papel cuadriculado</i>	graph paper

2. **Phrasal verbs** (*or verb particle constructions*) consist of a verb followed by one or more particles that together function as a lexical unit[10]. Some, like *acabar de* ‘just’ are idiomatic and opaque, while others have some transparent senses and some opaque or semi-opaque senses. For example, *fight on* ‘continue to fight’ is perhaps semi-opaque, but in the sense of ‘fight on the top of’, as in they *fought on the roof*, it is completely transparent. *Estar por* is opaque in the sense of ‘be on the verge of’ but transparent (compositional) in the sense of ‘be for’.

3. **Idioms**, and other **lexicalized phrases** like *fixed expressions* and *semi-fixed expressions* consist of word combinations whose overall meanings are typically not transparent from their components, e.g. *to rack one’s brains* over and

to *kick the bucket*. They are thus both opaque and non-compositional monolingually, but in some cases, like *elefante blanco* ‘white elephant’, they may be bilingually compositional.

4. **Proverbs** and similar sentential or semi-sentential constructions like adages, maxims and dicta express a general truth, belief or a moral. They are often idiomatic, opaque and non-compositional, e.g. *donde el Diablo perdió la camiseta* ‘the ends of the earth’.

5. **Collocations** are recurrent combinations of words co-occurring more often than by chance whose meaning are (mostly) compositional and transparent, e.g. *mal informado* ‘misinformed, incorrectly informed’ (see 5. **Collocations** below)

6. **Locutions** in the context of MWEs are grammatical collocations whose central component is a function word or adverb. An example of an adverbial locution is *lo antes posible* ‘as soon as possible’. Locutions are often idiomatic and non-compositional.

7. **Multiword proper nouns**, designate a person, place, company, organization, book titles and the like, e.g. *New York, George Washington, United Nations* (see 6. **Multiword proper nouns** below)

8. **Noncontiguous MWEs** are discontinuous lexical constructions that consist of fixed sequences of words interrupted by one or several gaps filled in by interchangeable words, *the more...the more*. Under this category can also be included MWEs like *be in control of*, which can be interrupted by lexical insertion, as in *be in complete control of*, or verbal phrases like *take off* in *he took his jacket off*. Non-contiguous MWEs are more challenging to identify and interpret than ordinary MWEs.

2.4 Inclusion criteria

Ideally, every type of MWE, especially non-compositional ones, should be included as headwords in both dictionaries for humans and in MT lexicons. Traditionally, dictionaries and MT lexicons have poor coverage of such MWEs as proverbs, locutions, and idiomatic expressions. It is self evident that if non-compositional MWEs are not included, or are not identified and interpreted correctly by some other means, translation accuracy will suffer.

3 Free Word Combination

3.1 Definition

A **free word combination** (FWC) is a *meaningful* free sequence of words that follow the rules of syntax but has no lexical status. FWCs have three characteristics: (1) they are potentially infinite in number, (2) they can be generated by native speakers spontaneously, and (3) they have no lexicographic or lexical status. Some examples include:

drink water
cerrar con las manos

cabrir un agujero
abrir la luz
 write a poem
 don't come home

FWCs are *meaningful* combinations of words (free word syntagmata), whereas meaningless combinations such as “went to New” as part of “went to New York” are ignored in linguistic analysis. FWCs are not lexical units in their own right but often appear in dictionaries in describing culture-bound terms and *untranslatable* words in place of a translational equivalent.

3.2 Analysis

It is important to note that such combinations as:

Table 3. Combinations. n.1.

<i>abrir un agujero</i>	dig a hole
<i>abrir un túnel</i>	dig a tunnel
<i>abrir la luz</i>	turn on the light
<i>abrir el agua</i>	turn on the water

may look like MWEs, perhaps because they are not based on the primary sense of *abrir* ‘to cause to open’. Nevertheless, they are indeed FWCs and have no lexical status, no more than combinations based on the primary senses of *abrir*, such as:

Table 4. Combinations n.2.

<i>abrir la puerta</i>	open the door
<i>abrir un hospital</i>	open a hospital
<i>abrir el baile</i>	begin the dance

Such FWCs are 100 percent transparent (compositional) and productive because *abrir* is a polysemous lexeme that has such senses as ‘cause to open’, ‘begin’ and ‘switch on’. The examples given are merely instances of how *abrir* is combined with direct objects.

It is important to understand this issue on the basis of objective linguistic factors, rather than subjective intuition, which is sometimes used by lexicographers in selecting dictionary entries or subentries. Below is a linguistic analysis

that demonstrates that *abrir la luz* and *abrir un agujero* are indeed FWCs, rather than MWEs.

Analyzing the semantic components of *abrir* ‘switch on’ and *abrir* ‘dig’ in relation to the free word syntagmata *abrir la luz* and *abrir un agujero*, what is required to explain the need for *la luz* and *un agujero* is not, as some lexicographers may be tempted to do, to consider them integral parts of lexicalized compound verbs, but to consider them to be semantic components consisting of an obligatory complementation of the verb by a noun phrase direct object with the selectional restriction that the complements are members of the semantic subdomain of utilities (gas, water, light...).

The fact that the senses in questions, i.e. ‘dig’ and ‘switch on’, are not central to the lexeme *abrir* is irrelevant. That is, *abrir* is a polysemic lexeme, and such productive senses as ‘switch on’ behave syntactically and grammatically exactly like the core meaning ‘cause to open’ in *abrir la puerta* ‘open the door.’ In other words, one must not be misled by the peripherality of the sense ‘switch on’, which may make *abrir la luz* look like a collocation or compound word, rather than the FWC that it actually is.

3.3 Inclusion criteria

Such frequently co-occurring syntactic constructions like *abrir la puerta* and *abrir el baile* must not be indiscriminately considered as MWEs, though they seem to behave like lexical units. They should *not* be included in dictionaries for humans, except as example sentences, or as part of occasional “descriptive equivalents” for difficult-to-translate headwords. If such FWCs were included, dictionaries would grow to astronomical proportions since it would allow billions of meaningful FWCs.

Let us take *abrir la puerta* as an example. Since the number of potential direct objects (*ventana, entrada, boca...*) is open ended (could be extremely large), it obviously makes no sense to list them exhaustively, especially not in dictionaries for humans. Any systematic attempt to do so would bloat the dictionary out of all proportion, since the potential number of FWC can be extremely large. That is, statistically significant co-occurrences of words combinations like FWCs are syntactic constructions that do not qualify as lexical units, not only because they are completely compositional, but also because they are often highly semantically productive. On the other hand, such FWCs could serve as useful example sentences in human dictionaries.

Though compositional FWCs, which are potentially infinite in number, need not (in fact cannot) be listed in dictionaries for humans, there is one exception. If a FWC has both non-compositional and compositional translation equivalents, for the sake of clarity both compositional and non-compositional) should be included. For example, *estar por* has the compositional (literal) equivalent ‘to be for’ (*estar* ‘to be’ + *por* ‘for’) and the non-compositional idiomatic sense of ‘to be on the verge of’.

Although FWCs such as *abrir la luz* and *abrir la puerta* are unnecessary in dictionaries for humans, they nevertheless can play a useful role in MT lexicons.

Theoretically, MT systems can correctly translate such FWCs as *abrir la luz* by word sense disambiguation (WSD) even if they are not in the lexicon. That is, once the system determines that the sense of *abrir* in this context is 'switch on', it can correctly translate to 'turn on the light'. Nevertheless, since memory is virtually unlimited, it makes sense to include some high-frequency FWCs in the MT lexicon explicitly because it greatly simplifies processing; that is, a simple lookup operation replaces the sophisticated semantic and contextual analysis that is required for WSD.

4 Phrasets

4.1 Definition

A phraset is a free, *meaningful* combination of words (FWC) that is recurrently used to express a concept that has no lexical status but corresponds to a lexical unit in another language, e.g. *cerrar con llave* corresponds to 'lock' and *ir en bicicleta* corresponds to 'cycle'.

Bentivogli and Pianta[1] [2] have discussed phrasets in detail in the context of WordNet. The term as used here has the following characteristics: (a) syntactically and grammatically they are indistinguishable from FWCs, (b) they have no lexical status but correspond to lexical units in another language, and (c) they are often used in bilingual dictionaries as a "descriptive equivalent" for lexical gaps, e.g. *cerrar con llave* as the equivalent of the lexical unit 'lock'.

4.2 Analysis

Lexical *anisomorphism*, a basic feature of language, refers to the lexical incompatibility between languages. One manifestation of this is the large number of *lexical gaps* in every language; that is, words that have no equivalents in the target language. Bilingual dictionaries overcome this by providing descriptive equivalents when possible, similar to a definition in monolingual dictionaries.

It is important to note that phrasets are not "lexical units" in the source language and are not normally listed as headwords in dictionaries (though they may appear as subentries or in example sentences) since their status in the language is essentially the same as FWCs. However, many phrasets do behave like lexical units; that is, they have semantic integrity and cohesiveness and express a concept compositionally for which the language lacks an established lexical unit.

It is only when viewed from the point of view of the *target* language that phrasets acquire a special status. For example, English-Spanish dictionaries translate *to cycle* by the phraset *ir en bicicleta*, so it gets its special status, if we can call it that, by virtue of that fact alone, not because native speakers consider it "special" in any way. This demonstrates an interesting and useful fact about phrasets: that they can be monolingually compositional and transparent (as *ir en bicicleta*), yet bilingually non-compositional or a simplex lexical units (asycle).

Distinguishing between FWCs and phrasets is, in principle, very difficult and often impossible since monolingually they behave identically. For example, though *write a poem* is an FWC in English, in Japanese there is a verb 作詩する *sakushi suru*, translated as ‘*write a poem*’ in English, so that *write a poem* can be classified as a phraset, rather than an FWC, from a Japanese point of view. For native speakers of English, *write a poem* has no special status – that is, it has exactly the same status as *write a letter*, *write a song*, *write a book* etc. – and consequently will not appear as a dictionary entry even in the most comprehensive monolingual English dictionaries, nor as a source headword in English-to-X bilingual dictionaries.

Why is this so? For native speakers, phrasets do not have a psychological reality as a combination of words that need to be treated as meaningful units; that is, they are completely transparent and compositional and thus are (probably) not registered in the internal mental lexicon of native speakers.

It should be noted that phrasets, just like FWCs and phrasal verbs (which are full-fledged lexical items), can be noncontiguous: that is, the phraset *cerrar con llave* can be interrupted by lexical insertion, as in *cerrar la puerta con llave*, adding to the difficulty of detecting them.

Phrasets are very useful for translating lexical gaps into the target language. Statistical techniques, such as extracting contiguous bigrams and trigrams of high occurrence and high semantic cohesiveness, have been used to detect them, but because monolingually their behavior is identical to that of FWCs, this is a difficult task. For example, *cerrar con llave* is identical in structure to *cerrar con las manos* – that is, they both have identical surface structures. One effective technique for detecting phrasets is to reverse the entries of bilingual dictionaries. For example, the entry *lock* in an English-Spanish dictionary will yield the phraset *cerrar con llave*. Another technique is using a database of bilingual aligned example sentences found in bilingual dictionaries, which are an excellent source of phrasets, or using sentence aligned parallel corpora.

4.3 Inclusion criteria

Most dictionaries for humans rarely, if ever, intentionally include phrasets as source language entries or subentries, since phrasets are semantically compositional and have no lexical status. On the other hand, MT lexicons, designed to achieve full reversibility and comprehensive coverage, listing phrasets is not only desirable but essential for achieving high translation accuracy. There is no question that including Spanish phrasets like *cerrar con llave* ‘to lock’ in a Spanish-English MT lexicon is essential since these cannot be compositionally translated into English (the literal translation ‘close with a key’ is unidiomatic and incorrect).

Another good example of a phraset is *ir en bicicleta*, which is equivalent to the English lexeme ‘to cycle’ and the FWC ‘ride a bicycle’. *Ir en bicicleta* is not normally listed as a source language entry in Spanish dictionaries; if it were, the question would be *where to draw the line?* That is, why not also include:

ir en coche
ir en carro
ir en monociclo
ir en avión
ir en patines
ir en globo

All of these have exactly the same status in Spanish, linguistically, lexicographically, and psychologically in the minds of native speakers and are thus are no different from *ir en bicicleta*.

It is worthwhile noting that translating source-language lexical units non-compositionally into target language phrasets is common, especially between highly anisomorphic language pairs like Japanese and English. This problem cannot be solved by haphazardly listing phrasets in dictionaries, but requires a comprehensive approach in which phrasets are collected systematically.

It is further worthwhile noting that grammatical anisomorphism combined with lexical anisomorphism are the reasons why languages have many “untranslatable” lexical units, some of which are not just “difficult to translate” but *in principle completely impossible to translate*. For example, *would’ve* as an isolated word is impossible to translate into many languages. Dictionaries *describe*, rather than translate, such words by using FWCs or phrasets. Because of the lexical, conceptual, and grammatical differences between languages, some lexical units in a language cannot, *in principle*, be translated, not because of lack of lexicographic skill.

5 Collocation

5.1 Definition

A **collocation** (or *institutionalized phrase*) is a recurrent combination of words that co-occur more often than by chance whose meaning is (mostly) compositional and transparent.

Table 5. Collocations

<i>bonita sorpresa</i>	nice surprise
<i>estar fascinado con</i>	be fascinated with
<i>tomar una decisión</i>	make a decision
<i>hacer una pausa</i>	take a break
<i>prestar atención</i>	pay attention
<i>hacer amor a/con</i>	make love to/with
<i>respecto a</i>	with regard to
<i>abandonarse a la des-</i> <i>peración</i>	to fall into despair

Collocations are a subtype of MWEs that have the following criteria: (a) they often cannot be translated literally, (b) they are (mostly) compositional and semantically transparent, (c) their components cannot be replaced without losing idiomaticity, and (d) they do not have full lexical status (see Table 5).

5.2 Analysis

A collocation is a group of words that co-occur more frequently than by chance. Collocations are a phenomenon that have linguistic status, and are also useful for statistical analysis and natural language processing. However, with the exception of specialized dictionaries of collocations and idiomatic phrases, they do not normally appear in dictionaries as main entries or subentries, but may appear in example sentences.

Collocations are difficult to define precisely. They are often discussed in contrast with FWCs on one end of the spectrum and with idiomatic expressions on the other. Whereas FWCs can be described in terms of general syntactic rules and semantic restrictions, idioms are fixed word combinations that are difficult or impossible to generalize. Collocations fall between these two extremes, though drawing a clear line between them is not always possible. Studying the above examples carefully should clarify how collocations differ from full-fledged lexical units.

Collocations don't have full lexical status because they don't normally represent concepts. For example, *nice surprise* is a conventional phrase corresponding to *bonita sorpresa* (*beautiful surprise* would be unidiomatic), but cannot be considered to be a full-fledged lexical unit in either language. That is, one can say that they are probably not registered in the brain's internal lexicon. They are more in the realm of usage conventions rather than full-fledged lexical units.

5.3 Inclusion criteria

Regardless of the theoretical distinction between collocations and full-fledged lexical units, because collocations are so common it is desirable to include them in bilingual dictionaries, not to speak of MT lexicons, in order to achieve higher translation quality.

Since collocations are compositional and semantically transparent, there is some chance that a human or MT system can translate them correctly by word-for-word substitution, but it is nevertheless desirable to list them explicitly in order to achieve better, unambiguous result.

It is highly desirable to make a systematic effort to collect collocations for including in comprehensive bilingual dictionaries as well as in MT lexicons. One technique for acquiring collocations is to extract them from aligned example databases based on paper dictionaries; another is to extract them from bilingual parallel corpora.

Table 6. Inaccurate translations of POIs.

Japanese	Google	Bing	Baidu	NICT	CJKI
海の中道線	Midair line of the sea	The middle line of the sea	The sea line	海の中道線	Umi-no-Nakamichi Line
三角線	Triangle	Triangular line	Misumi	Misumi Line	Misumi Line
十和田観光電鉄線	Triangle	Triangular line	Misumi	Misumi Line	Misumi Line
神津島空港	Towada Shimbun photoelectric wire	Towada Kanko railway line	Towada sight-seeing electric railway line	Towada Kankō Electric Railway Line	Towada Kanko Electric Railway Line
神津島空港	Kozu Island airport	God Tsushima Airport	Kozu Island Airport	Kōzushima Airport	Kozushima Airport
中部国際空港	Chubu International Airport	Chubu International Airport	Central Japan International Airport	Chubu International Airport	Chubu Centair International Airport
鬼の城公園	Demon Castle Park	Demon Castle Park	Demon Castle Park	Oni Castle Park	Oninojo Park

6 Multiword Proper Nouns

6.1 Definition

A *multiword proper noun* is a combination of two or more words that together function as a single proper noun. This includes place names such as *Republic of China*, companies and organizations such as *United Nations*, personal names such as *Shinzo Abe*, and points of interest (POI) such as *Narita International Airport*.

6.2 Analysis

The recognition and accurate translation of proper nouns, many of which are bilingually non-compositional, are a major issue in MT and other NLP applications. This is especially true for Chinese and Japanese, whose scripts present linguistic and algorithmic challenges not found in other languages. These difficulties are exacerbated by the lack of easily-available comprehensive lexical resources for proper nouns, especially POIs, resulting in a high rate of translation failure.

The CJK Dictionary Institute (CJKI), which specializes in CJK and Arabic computational lexicography, has been engaged in the construction of large-scale lexical resources that cover tens of millions personal names, place names, and POIs. These resources and methodology are described in Halpern[6] and Halpern[5]. Tests have shown that MT systems, including those using neural networks, often fail to accurately translate proper nouns, especially POIs. For example, a test to translate 75 Japanese POIs gave surprisingly poor results, as shown in Table 1. For example, 鬼の城公園 is translated by Baidu and Google word for word as 'Demon Castle Park', whereas the actual name of this park is 'Oninojo Park'.

6.3 Inclusion criteria

Dictionaries for humans normally do not include proper nouns, except possibly for very well know place names such as country names and famous people. Human users do not expect, and have no need for, comprehensive coverage of proper nouns. MT lexicons, on the other hand, should include as many proper nouns as possible. In fact, most MT systems perform poorly in translating proper nouns in general and multiword POIs in particular. To achieve higher translation accuracy, proper noun resources for MT must be greatly expanded.

7 Conclusions

This paper attempts to define the various types of MWUs, and to clarify the underlying linguistic concepts on the basis of linguistic and lexicographic principles while taking into consideration the needs of MT lexicons. It is clear that the accurate identification and translation of MWUs are critical to enhancing the translation accuracy of MT systems. It is hoped that the analysis given here will contribute to the improved identification of MWUs, based on (mostly) objective criteria, and that MT system developers will pay greater attention to the importance of large-scale lexicons with comprehensive coverage of MWUs, including proper nouns.

References

1. Bentivogli, L., Pianta, E.: Beyond lexical units: enriching wordnets with phrasets. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2 (EACL)., vol. 2, pp. 67-70. Association for Computational Linguistics, Stroudsburg, PA, USA (2003). DOI: <https://doi.org/10.3115/1067737.1067750>
2. Bentivogli, L., Pianta, E.: Extending WordNet with Syntagmatic Information. In: Sojka,P., Pala, K., Smrz, P., Fellbaum, C., Vossen, P. (eds.), Proceedings of the Second International WordNet Conference - GWC 2004, pp. 47.53. Masaryk University Brno, Czech Republic, (2004).

3. Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A.: Towards best practice for multiword expressions in computational lexicons. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), pp. 1934-1940. (2002) <http://www.lrec-conf.org/proceedings/lrec2002/pdf/259.pdf>.
4. Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T., Frey, J.: Context-based machine translation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas. (2006) <http://www.mt-archive.info/AMTA-2006-Carbonell.pdf>.
5. Halpern, J.: The Role of Lexical Resources in CJK Natural Language Processing. In: Proceedings of the Fifth Workshop on Chinese Language Processing, SIGHAN@COLING/ACL 2006, pp. 22-23. Association for Computational Linguistics 2006, Sydney, Australia (2006).
6. Halpern, J.: Very Large-scale Lexical Resources to Enhance Chinese and Japanese Machine Translation. In: TAUS Executive Forum Tokyo 2017. (2017).
7. Henderson, J. A.: What's in a Word?. The University of Edinburgh, Edinburgh (2007).
8. Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. F. (Ed.) Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), pp. 1-15. Springer-Verlag, Berlin, Heidelberg (2001).
9. Váradi, T.: Multiword units in an MT lexicon. In: Proceedings of the Workshop on Multi-word-expressions in a multilingual context. (2006) <https://www.aclweb.org/anthology/W06-24>.
10. Villavicencio, A., Bond, F., Korhonen, A., McCarthy, D.: Editorial: Introduction to the special issue on multiword expressions: Having a crack at a hard nut. in: *Comput. Speech Lang.* 19, 4 (October 2005), pp. 365-377. (2005) <http://dx.doi.org/10.1016/j.csl.2005.05.001>.

Multiword Expressions Under the Microscope

Aline Villavicencio¹

¹ Department of Computer Science, University of Sheffield (UK)

² Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
a.villavicencio@sheffield.ac.uk

Almost 2 decades have passed since the publication of the paper *Multiword Expressions: a pain in the neck for NLP*[15]. In this time there have been considerable advances in representing Multiword Expressions (MWEs) in various languages, resulting from several projects, events and initiatives devoted to them[13] [12][11][16]. Ranging from idioms (*make ends meet*), light verb constructions (*take a shower*) and verb particle constructions (*shake up*) to noun compounds (*loan shark*), MWEs have provided new challenges and opportunities for language processing[5]. Their integration in tasks and applications like parsing[9][6], information retrieval[1], machine translation[4] has brought improvements for language technology, providing a degree of precision, naturalness and fluency. Any amount of interest they have attracted is justified as they account for an important part of human languages with estimates that they appear in the mental lexicon of native speakers with the same order of magnitude as single words[10], and with about four MWEs being produced per minute of discourse[8], in all languages and domains from informal to technical contexts[3]. After all this time, should they still be considered as an *open problem*[17] and *hard going*[14], or is it all *plain sailing*[14]?

In this talk I present an overview of advances in the identification of multiword expressions, that often capitalize on the various degrees of idiosyncrasy they display, including lexical, syntactic, semantic and statistical[2][18]. I will concentrate on techniques for identifying their degree of idiomaticity and approximating their meaning, as their interpretation often needs more knowledge than can be gathered from their individual components and their combinations[7] (Fillmore, 1979) to differentiate combinations whose meaning can be (partly) inferred from their parts (as *apple juice: juice made of apples*) from those that cannot (as *dark horse: an unknown candidate who unexpectedly succeeds*).

Acknowledgements

This talk includes joint work with Carlos Ramisch, Marco Idiart, Silvio Cordeiro, Rodrigo Wilkens, Felipe Paula and Leonardo Zilio.

References

1. Costa Acosta, O., Villavicencio A., Pereira Moreira, V.: Identification and treatment of multiword expressions applied to information retrieval. In: Kordoni, V., Ramisch,

- C., Villavicencio, V. (eds.) Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, MWE@ACL 2011, pp. 101-109. Association for Computational Linguistics, Portland, Oregon, USA (2011).
2. Baldwin, T., Nam Kim, S.: Multiword expressions. In: Indurkha, N., Damerau, F. J. (eds), Handbook of Natural Language Processing, 2nd edn., pp. 267-292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA (2010).
 3. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E.: Longman Grammar of Spoken and Written English. 1st edn. Pearson Education Ltd, Harlow, Essex (1999).
 4. Carpuat, M., Diab, M. T.: Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, pp. 242-245. The Association for Computational Linguistics, Los Angeles, California, USA (2010).
 5. Constant, M., Eryğit, G., Monti, J., Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Multiword expression processing: A survey. *Computational Linguistics* **43**(4), 837–892 (2017).
 6. Constant, M., Nivre, J.: A transition-based system for joint lexical and syntactic analysis. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016. The Association for Computer Linguistics, Berlin, Germany (2016).
 7. Fillmore, C. J.: Innocence: A second idealization for linguistics. In: Annual Meeting of the Berkeley Linguistics Society (1979).
 8. Glucksberg, S.: Metaphors in conversation: How are they understood? Why are they used?. *Metaphor and Symbolic Activity* **4**(3), 125-143 (1989).
 9. Green, S., de Marneffe, M-C., Manning, C. D.: Parsing models for identifying multiword expressions. *Computational Linguistics* **39**(1), 195–227 (2013).
 10. Jackendoff, R.: Twistin' the night away. *Language* **73**, 534–559 (1997).
 11. Corpas Pastor, G., Colson, J-P.: Computational and Corpus-based Phraseology. John Benjamins, (2019).
 12. Ramisch, C., Cordeiro, S.R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljn, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Escartín, C. P., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., Welsh, A.: Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In: Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pp. 222–240. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018).
 13. Ramisch, C., Villavicencio, A.: Computational treatment of multiword expressions. In: Mitkov, R. (Ed.) The Oxford Handbook of Computational Linguistics. 2nd edn. Oxford University Press (2018).
 14. Rayson, P., Piato, S., Sharoff, S., Evert, S., Villada Moirón, B.: Multiword expressions: hard going or plain sailing?. *Language Resources and Evaluation* **44**(1-2), 1-5 (2010).
 15. Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for NLP. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, pp. 1-15. Springer-Verlag, Berlin, Heidelberg (2002).
 16. Savary, A., Parra Escartín, C., Bond, F., Mitrovic, J., Barbu Mititelu, V.: Proceedings of the Joint Workshop on Multiword Expressions and WordNet, MWE-

- WN@ACL 2019. Association for Computational Linguistics, Florence, Italy (2019), <https://www.aclweb.org/anthology/volumes/W19-51/>
17. Schone, P., Jurafsky, D.: Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2001. ACL, Pittsburgh, PA USA (2001).
 18. Villavicencio, A., Idiart, M.: Discovering multiword expressions. *Natural Language Engineering*, 1-19 (2019). <https://doi.org/10.1017/S1351324919000494>.

Author Index

Albano, Mariangela, 1
Andugar Andreu, Isabel, 11

Badalamenti, Rosa Leandra, 1
Baptista, Jorge, 70
Bautista Zambrana, María Rosario, 31
Bautista, Francisco, 19
Bevilacqua, Cleci Regina, 40
Blagus Bartolec, Goranka, 46, 106
Bonadonna, Maria Francesca, 53

Cataldo, Silvia, 61
Colson, Jean-Pierre, 145

Dogru, Gokhan, 157

Galvão, Ana, 70
Gobbo, Federico, 78
Grabowski, Lukasz, 140

Halpern, Jack, 167

Inoue, Ai, 86

Larsen-Walker, Melissa, 90
Luque Giraldez, Ángela, 99

Mamede, Nuno, 70
Matas Ivanković, Ivana, 46, 106
Mirzadeh, Seyed Mohammad Hossein, 114

Nigrelli, Castrenze, 118, 126

Potemkin, Serge, 133

Roldan-Riejos, Ana, 140

Seghiri, Míriam, 99

Villavicencio, Aline, 181

Zollo, Silvia Domenica, 53