draft-proceedings-europhras2022

Conference Paper · January 2022

CITATIONS
0

READS
97

2 authors, including:

Anita Braxatorisová
University of St. Cyril and Methodius of Trnava - Univerzita sv. Cyrila a Metoda
18 PUBLICATIONS **7** CITATIONS

SEE PROFILE

Computational and Corpus-based Phraseology

Proceedings of the International Conference
EUROPHRAS 2022

(short papers, posters and MUMTTT workshop contributions)

28-30 September, 2022
Malaga, Spain

ORGANISERS

SPONSORS

ISBN 978-954-452-080-9

2022. INCOMA Ltd. Shoumen, BULGARIA

This document is downloadable from `http://europhras.com/2022/`

Editors of the Proceedings

Gloria Corpas Pastor
Ruslan Mitkov
Maria Kunilovskaya
Rocío Caro Quintana

Organisers:

Europhras 2022 was jointly organised by the European Association for Phraseology (Europhras), the University of Malaga (Research Group in Lexicography and Translation), Spain, the University of Wolverhampton (Research Group in Computational Linguistics), United Kingdom, 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:

Margarita María Alonso Ramos, University of A Coruña, Spain
María Belén Alvarado Ortega, University of Alicante, Spain
Verginica Barbu Mititelu, Romanian Academy, Romania
Ignacio Bosque, Complutense University of Madrid, Spain
María Luisa Carrió-Pastor, Polytechnic University of Valencia, Spain
Anna Čermáková, University of Cambridge, United Kingdom
Parthena Charalampidou, Aristotle University of Thessaloniki, Greece
Ken Church, Baidu
Jean-Pierre Colson, Université Catholique de Louvain, Belgium
Dmitrij Dobrovolskij, Russian Language Institute, Russian Federation
Peter Ďurčo, University of St. Cyril and Methodius, Slovakia
Natalia Filatkina, University of Hamburg, Germany
Elizaveta Goncharova, National Research University, Artificial Intelligence Research Institute, AIRI
María Isabel González Rey, University of Santiago de Compostela, Spain
Stefan Gries, University of California, United States of America
Enrique Gutiérrez Rubio, Palacký University Olomouc, Czech Republic
Kleanthes K. Grohmann, University of Cyprus, Cyprus
Amal Haddad Haddad, University of Granada, Spain
Miloš Jakubíček, Sketch Engine
Eva Lucía Jiménez-Navarro, University of Cordoba, Spain
Cvetana Krstev, University of Belgrade, Servia
Natalie Kübler, Université Paris Cité, Grance
Maria Kunilovskaya, University of Wolverhampton, United Kingdom
Ljubica Leone, Lancaster University, United Kingdom
Óscar Loureda Lamas, Heidelberg University, Germany
Elvira Manero Richard, University of Murcia, Spain
Ramón Martí Solano, University of Limoges, France

María del Carmen Mellado Blanco, University of Santiago de Compostela, Spain
Flor Mena Martínez, University of Murcia, Spain
Pedro Mogorrón Huerta, University of Alicante, Spain
Johanna Monti, "L'Orientale" University of Naples, Italy
Esteban Tomás Montoro del Arco, University of Granada, Spain
Inés Olza Moreno, University of Navarra, Spain
Adriane Orenha Ottaiano, São Paulo State University, Brazil
Antonio Pamies Bertrán, University of Granada, Spain
Rozane Rebechi, Federal University of Rio Grande do Sul, Brazil
María Ángeles Recio Ariza, University of Salamanca, Spain
Ute Römer, Georgia State University, United States of America
Leonor Ruiz Gurillo, University of Alicante, Spain
Kathrin Steyer, University of Mannheim, Germany
Joanna Szerszunowicz, University of Bialystok, Poland
Yukio Tono, Tokyo University of Foreign Studies, Japan
Agnès Tutin, University of Grenoble Alpes, France
Aline Villavicencio, Federal University of Rio Grande do Sul, Brazil, and University of Sheffield, United Kingdom
Tom Wasow, Stanford University, United States of America
Eric Wehrli, University of Geneva, Switzerland
Michael Zock, Laboratoire d'Informatique Fondamentale de Marseille, France

Additional Reviewers:

Dayana Abuin Rios, University of Malaga, Spain
Rocío Caro Quintana, University of Wolverhampton, United Kingdom
Isabel Durán, University of Malaga, Spain
Richard Evans, University of Wolverhampton, United Kingdom
Emma Franklin, University of Wolverhampton, United Kingdom
Carlos Manuel Hidalgo Ternero, University of Malaga, Spain
Nieves Jiménez Carra, University of Malaga, Spain
Alfiya Khabibullina, University of Wolverhampton, United Kingdom
Lilit Kharatyan, University of Wolverhampton, United Kingdom
Ruslan Mitkov, University of Wolverhampton, United Kingdom
Daria Sokova, University of Wolverhampton, United Kingdom

Invited Speakers:

Aline Villavicencio, Federal University of Rio Grande do Sul, Brazil, and University of Sheffield, United Kingdom
Jean-Pierre Colson, Université Catholique de Louvain, Belgium
María del Carmen Mellado Blanco, University of Santiago de Compostela, Spain
Miloš Jakubíček, Sketch Engine

Organising Committee:

University of Malaga

Presentación Aguilera Crespillo
Marta Alcaide Martínez
Rosario Bautista Zambrana
Isabel Durán Muñoz
J. Alejandro Fernández Sola
Mahmoud Gaber
Rut Gutiérrez Florido
Carlos Manuel Hidalgo Ternero
Hanan Saleh Hussein
Adriana Iglesias Lara
Francisco Javier Lima Florido
Gema Lobillo Mora
Araceli Losey León
Jorge Lucas Pérez
Luis Carlos Marín Navarro
Desiré Martos García
Laura Noriega Santiáñez
Laura Parrilla Gómez
Míriam Pérez Carrasco
Encarnación Postigo Pinazo
María del Pilar Rodríguez Reina
Juan Antonio Sánchez Muñoz
Fernando Sánchez Rodas
Míriam Seghiri Domínguez
Cristina Toledo Báez

University of Wolverhampton

Dayana Abuin
Isuri Anuradha
Anastasia Bezobrazova
Rocío Caro Quintana
Ana Isabel Cespedosa Vázquez
Amal El Farhmat
Suman Hira
Alfiya Khabibullina
Lilit Kharatian
Maria Kunilovskaya
Gabriela Llull
Kamshat Saduakassova
Kanishka Silva
Daria Sokova

MUMTTT 2022 Workshop Chairs

Gloria Corpas Pastor, Universidad de Málaga, Spain
Ruslan Mitkov, University of Wolverhampton, United Kingdom

Johanna Monti, Università degli Studi di Napoli "L'Orientale" Italy
Maria Pia di Buono, Università degli Studi di Napoli "L'Orientale" Italy

MUMTTT 2022 Programme Committee

Giuseppe Attardi, University of Pisa
Verginica Barbu Mititelu, Romanian Academy Research Institute for Artificial Intelligence
Jean-Pierre Colson, Université catholique de Louvain
Anna Beatriz Dimas Furtado, University of Wolverhampton
Federico Gaspari, University for Foreigners "Dante Alighieri"
Amal Haddad Haddad, University of Granada
Philipp Koehn, The Johns Hopkins University
Judyta Mężyk, Paris-Est Créteil University and University of Silesia in Katowice
Pavel Pecina, Charles University
Éric Poirier, Université du Québec à Trois-Rivières
Carlos Ramisch, Aix Marseille University
Max Silberztein, Université de Franche-Comté
Kathrin Steyer, Institut für Deutsche Sprache, Mannheim
Beata Trawinski, Institut für Deutsche Sprache, Mannheim
Agnes Tutin, Université Grenoble Alpes

MUMTTT 2022 Organising Committee

Gennaro Nolano, Università degli Studi di Napoli "L'Orientale" Italy
Giulia Speranza, Università degli Studi di Napoli "L'Orientale" Italy
Khadija Ait ElFqih, Università degli Studi di Napoli "L'Orientale" Italy

Association for Computational Linguistics (Bulgaria)

Nikolai Nikolov

Table of Contents

Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2022) workshop contributions

Unidades fraseológicas verbales metafóricas con *testa* y *cabeza*: un análisis contrastivo italiano-español

Silvia Cataldo

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

Resumen. En esta contribución se presenta una parte de un estudio más amplio en el que se analiza una serie de unidades fraseológicas (UFs) metafóricas con función verbal italianas y españolas que contienen los sustantivos *testa* y *cabeza* con el objetivo de determinar analogías y diferencias interlingüísticas. En el análisis se recopilan las UFs, se clasifican las mismas en campos y subcampos semánticos, se identifican los campos y subcampos semánticos compartidos y las UFs con el mismo significado metafórico específico. Si bien la clasificación se basa en los significados metafóricos de las expresiones, a la hora de determinar las correspondencias interlingüísticas se considera también la componente literal: puesto que los significados metafóricos suelen nacer de imágenes concretas que sufren un proceso de abstracción y que dos lenguas pueden utilizar imágenes diferentes para referirse a un mismo significado o atribuir a una misma imagen significados distintos, posibles divergencias con respecto a la relación entre sentido literal y sentido metafórico podrían conllevar malentendidos o problemas de traducción. Todo esto evidencia también la dificultad de transferir ciertos contenidos de una lengua fuente a una lengua meta, pese a su gran afinidad, tratando de mantener inalterados el significado y la imagen evocada.

Palabras clave: Fraseología, Metáforas, Análisis Contrastivo.

1 Introducción

El cuerpo humano representa un dominio especialmente productivo en la formación de unidades fraseológicas (de aquí en adelante UFs o fraseologismos): los seres humanos comprenden una gran cantidad de conceptos abstractos en términos de cosas concretas familiares relacionadas con la experiencia física y su cuerpo es el instrumento mediante el cual son capaces de percibir y descubrir lo que está a su alrededor (*cf*. Kövecses 2002). Puesto que una porción considerable de UFs es metafórica, resulta razonable pensar que el cuerpo humano es un dominio muy explotado por la fraseología, lo cual permite adquirir mucho material para realizar un análisis contrastivo seleccionando tan solo una parte de él. La existencia de numerosos estudios dedicados a las UFs somáticas testimonia su enorme potencial para la investigación; además, su presencia constante en las comunicaciones humanas mantiene alto el interés por estudiarlas desde diferentes enfoques. Baran À Nkoum (2018), por ejemplo, se concentra en UFs españolas basadas en el dominio de la cabeza y las clasifica según la parte de la cabeza de la que se sirven

para luego identificar correspondencias en la lengua francesa; por otra parte, Navarro (2007) se concentra en el italiano y el español y construye un corpus de UFs en ambas lenguas estructurado según la parte del cuerpo que aparece en la expresión para determinar el grado de correspondencia interlingüística (total, parcial y niveles intermedios). En esta contribución, en cambio, se recopilan UFs italianas y españolas que contienen los sustantivos *testa* y *cabeza* y la clasificación se basa en la definición de campos y subcampos semánticos para cada lengua por separado para poder luego definir las correspondencias interlingüísticas; en la determinación de estas últimas, al mismo tiempo, se tiene en cuenta la relación entre el significado metafórico y el significado literal.

2 Marco teórico

Debido a que el significado de las UFs consideradas para este estudio no corresponde a su sentido literal y adquiere un carácter metafórico, las bases en las que se fundamenta el análisis conciernen a la fraseología por un lado y a las teorías relativas a la metáfora conceptual por otro. En cuanto a estas últimas, se hace referencia a la concepción de la metáfora como mecanismo cognitivo que se manifiesta en la lengua como resultado de transferencias mentales entre dominios más conocidos y normalmente concretos y dominios menos conocidos y en muchos casos abstractos, tal y como observan Black (1962), Lakoff y Johnson (1980) y Kövecses (2005; 2010a; 2010b; 2010c). Estas conceptualizaciones pueden ser convencionales o no en la mente (dependiendo de si son propias de una cultura o un grupo social en concreto o de difusión más amplia) y encontrarse lexicalizadas en la lengua (en forma de polisemias o UFs) o manifestarse como formulaciones libres, no fijadas a nivel léxico y/o sintáctico. El otro ámbito de investigación en el que se enmarca este trabajo concierne a la fraseología, cuyo objeto de estudio son las UFs. Estas representan expresiones estables en las lenguas que reúnen por lo menos algunas de las siguientes características descritas por Corpas Pastor (1996: 19-32): (i) frecuencia de uso y coaparición de los elementos que las constituyen; (ii) convencionalización/institucionalización, por haber llegado a ser combinaciones lingüísticas ya disponibles de las que los hablantes se sirven sin tener que producir sus propias formulaciones; (iii) estabilidad interna (que obstaculiza la manipulación de los componentes), externa (que obliga a utilizar la expresión en contextos preestablecidos) y semántica (que deriva de la adquisición de un nuevo significado compartido por todos los hablantes de una misma lengua); (iv) idiomaticidad, relativa al carácter metafórico de la UF, cuyo significado no representa la suma de los significados de cada uno de sus elementos; (v) variación, que permite aportar un número reducido de modificaciones a algunas UFs, las cuales están predeterminadas y fijadas en la lengua; (vi) gradación, que concierne al nivel de presencia o ausencia de las propiedades enumeradas.

Un trabajo de inspiración para definir un sistema de clasificación de los fraseologismos en este estudio ha sido el de Cataldo (2020), en el que se analiza a nivel cognitivo una serie de UFs italianas, españolas y alemanas referidas a la locura y se identifican las que comparten una imagen general y, al mismo tiempo, se diferencian por sus conceptualizaciones más específicas. Una metáfora general que asocia la locura a una carencia, por ejemplo, se especializa en la comprensión del loco en términos de máquina

a la que le falta algún componente (*faltarle un tornillo a alguien*) o en otros tipos de faltas (*faltarle a alguien caramelos en el frasco*). Este mecanismo cognitivo se podría aplicar a la dimensión del significado, puesto que es posible identificar, dentro de un grupo de UFs, una macroárea semántica que comprende distintos matices. En un principio análogo se basa el corpus de UFs elaborado por el grupo de investigación FRASYTRAM de la Universidad de Alicante, estructurado en campos y subcampos temáticos (*cf*. Gallego y Albaladejo 2016). Sin embargo, mientras que este corpus está organizado por significados, independientemente de aspectos léxicos, en la presente contribución los campos y subcampos semánticos están subordinados a la elección de un dominio específico, la cabeza. Esto se debe al objetivo del estudio, que no es crear una fuente lexicográfica, sino observar la complejidad de las relaciones interlingüísticas en el ámbito de las UFs, que oscilan entre analogías o disimetrías totales y afinidad únicamente formal o semántica.

3 Aspectos metodológicos y análisis

La primera fase del análisis consiste en la recopilación de UFs italianas y españolas con función verbal que contienen los sustantivos *testa* y *cabeza*. Para ello se han consultado el diccionario De Mauro y el de la Real Academia Española (DRAE), ambos en su versión en línea, los cuales presentan en forma de listado los fraseologismos cuyo término principal es *testa* y *cabeza*. De las aproximadamente 70 UFs recopiladas para cada lengua, se considera que las mencionadas a continuación pueden servir como muestra para este trabajo, pues ponen de manifiesto la variedad de relaciones interlingüísticas que ha resultado del análisis.

Italiano: *abbassare la testa*, *chinare la testa*, *curvare la testa*, *uscire con la testa rotta*, *alzare la testa*, *tenere testa*, *lavare la testa*, *mangiare in testa*, *mettere i piedi in testa*, *montare/montarsi la testa*, *prendere la testa*, *pagare un occhio della testa*.

Español: *bajar la cabeza*, *doblar la cabeza*, *levantar cabeza*, *hacer cabeza*, *dar en la cabeza a alguien*, *quebrantar a alguien la cabeza*, *henchir a alguien la cabeza de viento*, *llenar a alguien la cabeza*, *traer alguien sobre su cabeza a alguien o algo*, *poner alguien sobre la cabeza algo*, *dar alguien de cabeza*, *dolerle a alguien la cabeza*.

A partir de las UFs recopiladas, se han determinado para cada lengua varios campos y subcampos semánticos, que representan respectivamente los significados metafóricos generales y los específicos. Sucesivamente, se han identificado los campos y subcampos compartidos y se ha tratado de definir analogías y diferencias interlingüísticas con respecto a la relación que las UFs presentan entre sus sentidos literales y sus sentidos metafóricos. Las UFs totalmente análogas están alineadas, mientras que las que comparten solo el significado figurado se encuentran en el mismo bloque, pero en líneas distintas. Con un asterisco se señalan las que presentan una correspondencia únicamente literal entre el italiano y el español.

Italiano (diccionario De Mauro online)	**Español (DRAE online)**
CAMPO SEMÁNTICO: SUPERIORIDAD-INFERIORIDAD	

a. **SUBCAMPO SEMÁNTICO: SUMISIÓN, OBEDIENCIA**	
abbassare la testa: obbedire a una volontà superiore, rassegnarsi [obedecer a una voluntad superior, resignarse]	**bajar la cabeza**: coloq. Obedecer y ejecutar sin réplica lo que se manda. | coloq. Conformarse, tener paciencia cuando no hay otro remedio
chinare la testa: piegare il capo in segno di riverenza, sottomissione o saluto […] fig., rassegnarsi, sottomettersi [doblar la cabeza como signo de reverencia, sumisión o saludo [...] fig., resignarse, someterse]	
curvare la testa: ubbidire umilmente [obedecer humildemente]	**doblar la cabeza***: coloq. bajar la cabeza
b. **SUBCAMPO SEMÁNTICO: SUFRIR UNA DERROTA**	
uscire con la testa rotta: risultare perdente, avere la peggio in un contrasto, in una discussione, ecc. [resultar perdedor, perder en un contraste, en una discusión, etc.]	
c. **SUBCAMPO SEMÁNTICO: MOSTRARSE REBELDE U ORGULLOSO**	
alzare la testa: farsi valere; ribellarsi [hacerse valer, rebelarse]	**levantar cabeza:** coloq. salir de una situación desgraciada
tenere testa: opporsi e resistere validamente a qcn. o anche a qcs. senza lasciarsi sopraffare […] avere ragione di qcn., spuntarla con qcn. [oponerse y resistir válidamente a alguien o algo sin dejarse dominar [...] tener la razón de alguien, vencer a alguien [...]	
d. **SUBCAMPO SEMÁNTICO: REGAÑAR**	
lavare la testa: rimproverare severamente [regañar severamente]	
e. **SUBCAMPO SEMÁNTICO: PRIMAR, VENCER**	
mangiare in testa (i): fig., essere migliore [fig., ser mejor]	
	hacer cabeza: Ser el principal en un negocio o grupo de personas. desus. Hacer frente a los enemigos.
	dar en la cabeza a alguien: Frustrar sus designios, vencerlo.
f. **SUBCAMPO SEMÁNTICO: SUPERAR EN ALTURA**	
mangiare in testa (ii): di qcn., essere più alto, superare in altezza [de alguien, ser más alto, superar en altura]	
g. **SUBCAMPO SEMÁNTICO: SOMETER**	

mettere i piedi in testa: sopraffare [dominar]	
	quebrantar a alguien la cabeza: Humillar su soberbia, sujetarlo.
h. **SUBCAMPO SEMÁNTICO: EXALTACIÓN, SOBERBIA, ADULACIÓN**	
montare la testa: esaltare, far insuperbire, far credere a qcn. di possedere doti e qualità superiori a quelle effettive [exaltar, hacer ensoberbecer, hacerle creer a alguien que posee dotes y cualidades superiores a las reales] **montarsi la testa**: esaltarsi, insuperbirsi [exaltarse, ensoberbecerse]	
	henchir a alguien la cabeza de viento: coloq. Adularlo, lisonjearlo, llenarlo de vanidad
	llenar a alguien la cabeza de viento: coloq. henchir la cabeza de viento.
	traer alguien sobre su cabeza a alguien o algo: Hacer grandísima estimación de él o de ello
	poner alguien sobre la cabeza algo: Poner sobre su cabeza, en señal de respeto y reverencia, el documento que recibe, […] | Hacer grandísima estimación de ello
i. **SUBCAMPO SEMÁNTICO: MANDAR**	
prendere la testa: comandare, capitanare | estens., dirigere [mandar, capitanear | extensivo: dirigir]	
CAMPO SEMÁNTICO: GASTAR DINERO	
pagare un occhio della testa: strapagare [pagar mucho dinero]	
CAMPO SEMÁNTICO: PERDER RIQUEZA O AUTORIDAD	
	Dar alguien de cabeza*: coloq. Caer de su fortuna o autoridad
	Dolerle a alguien la cabeza: coloq. Estar próximo a caer de su privanza o autoridad

3.1 Resultados del análisis

Como ya se ha aclarado, este análisis representa una parte de un estudio más amplio que incluye alrededor de 70 UFs tanto para el italiano como para el español (a las que se añaden variantes formales en ambas lenguas). Sin embargo, los resultados de este

trabajo parcial reflejan los del análisis completo, dado que, desde una perspectiva contrastiva, ambos muestran una heterogeneidad de relaciones interlingüísticas, que se pueden resumir de la siguiente forma:

1. En lo que atañe al sentido metafórico:
   a. Con respecto al campo semántico:
      (1) Campo semántico presente tanto en italiano como en español
      (2) Campo semántico presente solo en italiano
      (3) Campo semántico presente solo en español
   b. Con respecto al subcampo semántico:
      (1) Subcampo semántico presente tanto en italiano como en español
      (2) Subcampo semántico presente solo en italiano
      (3) Subcampo semántico presente solo en español (caso no presente en esta contribución en concreto)
2. En lo que atañe a la relación entre sentidos literal y metafórico:
   a. Mismo significado literal y mismo significado metafórico (UFs alineadas)
   b. Mismo significado metafórico, significado literal distinto (UFs en un mismo bloque, no alineadas)
   c. Mismo significado literal, significado metafórico distinto (UFs en líneas o bloques diferentes, señaladas con un asterisco)

De las 13 UFs italianas (una de las cuales con dos sentidos metafóricos) y las 12 españolas examinadas: (a) 3 comparten su sentido literal y figurado[1]; (b) 12 italianas y 10 españolas comparten el mismo campo semántico[2]; (c) de estas, 4 italianas no tienen correspondencias en español con respecto al subcampo semántico[3]; (d) de las que no tienen correspondencias relativas al campo semántico, una es italiana y dos son españolas[4]. Por último, hay casos de afinidades únicamente literales: *doblar la cabeza* presenta una segunda acepción relativa a la muerte que no ha sido incluida en esta parte del análisis; análogamente, para *dar alguien de cabeza* existe una UF italiana (*dare di testa*) con el mismo sentido literal y un sentido metafórico diferente al que se indica en el DRAE, cuyo campo semántico ha quedado excluido de esta contribución. Es curiosa, además, la analogía parcial entre la expresión *pagare un occhio della testa* y su forma española más cercana en la que se pierde la referencia a la cabeza (*un ojo de la cara*). Sin embargo, es preciso señalar cierto carácter subjetivo del análisis, especialmente con respecto a la definición de las áreas semánticas, y su valor relativo, dado que podría sufrir cambios o resultar más completo si se integrara la consulta de las fuentes lexicográficas utilizadas con ulteriores recursos.

[1] *Abbassare la testa* con *bajar la cabeza*, *curvare la testa* con *doblar la cabeza* y *alzare la testa* con *levantar cabeza*

[2] Todas menos *pagare un occhio delle testa*, *dar alguien de cabeza* y *dolerle a alguien la cabeza.*

[3] *Uscire con la testa rotta*, *lavare la testa*, *mangiare in testa (ii)* y *prendere la testa*.

[4] *Pagare un occhio della testa*, *dar alguien de cabeza* y *dolerle a alguien la cabeza.*

4 Conclusiones

El análisis muestra que, pese a la gran afinidad entre el italiano y el español, las simetrías entre estas lenguas con respecto a las UFs examinadas suelen ser parciales: a una UF italiana corresponde a menudo una española con un significado figurado similar, pero que presenta una conceptualización distinta y, por lo tanto, otro sentido literal. Esto se debe a que determinadas asociaciones mentales son propias de una lengua y cultura y dos UFs pueden compartir su imagen general y diferenciarse en la específica (*cf*. Cataldo 2020). Análogamente, si se considera su significado, ciertas UFs italianas y españolas pueden pertenecer a una misma área semántica, pero presentar significados concretos distintos. Este trabajo, que se propone como una contribución más dentro del amplio universo de estudios contrastivos dedicados a las UFs somáticas, ofrece un posible modelo de análisis para investigaciones futuras basadas, por ejemplo, en la consulta de corpus, con otras combinaciones lingüísticas y otros tipos de UFs seleccionadas según criterios distintos, que puedan ser de interés didáctico y traductológico.

Referencias

1. Baran À Nkoum, P.: Estudio contrastivo español-francés de las locuciones verbales somáticas relativas a la cabeza. Tesis doctoral. Universidad Complutense de Madrid (2015).
2. Black, M.: Models and metaphors: Studies in language and philosophy. Cornell University Press, Ithaca (1962).
3. Cataldo, S.: Approccio cognitivo alla variazione fraseologica: alcune concettualizzazioni della pazzia in italiano e implicazioni per la loro traduzione in spagnolo e tedesco. En: Mogorrón Huerta, P. (ed.) Análisis multidisciplinar del fenómeno de la variación en traducción e interpretación. MonTI Special Issue 6, pp. 65-93 (2020).
4. Corpas Pastor, G.: Manual de fraseología española. Gredos, Madrid (1996).
5. Gallego Hernández, D., Albaladejo Martínez, J.A.: Clasificación temática de unidades fraseológicas sobre economía: un recurso para la acción docente. En: Tortosa Ybáñez, M.T., Grau Company, S., Álvarez Teruel, J.D. (eds.) XIV Jornadas de Redes de Investigación en Docencia Universitaria. Investigación, innovación y enseñanza universitaria: enfoques pluridisciplinares. Alicante: Universidad de Alicante, pp. 1342-1352 (2016).
6. Kövecses, Z.: Metaphor. A practical introduction. Oxford University Press, Oxford (2002).
7. Kövecses, Z.: Metaphor in culture: universality and variation. Cambridge University Press, Cambridge (2005).
8. Kövecses, Z.: A new look at metaphorical creativity in cognitive linguistics. Cognitive Linguistics 21(4), 663–697 (2010a).
9. Kövecses, Z.: Metaphor and Culture. Acta Universitatis Sapientiae, Philologica 2(2), pp. 197-220 (2010b).
10. Kövecses, Z.: Metaphor, Creativity, and Discourse. DELTA 26, pp. 719-738 (2010c).
11. Lakoff, G., M. Johnson: Metaphors we live by. The University of Chicago Press, Chicago/London (1980).
12. Navarro, C.: Fraseología contrastiva del Español y el Italiano (Análisis de un corpus bilingüe). Tonos digital: Revista de estudios filológicos 13 (2007).
13. Diccionario de la Real Academia (DRAE), http://dle.rae.es/, último acceso 09/04/2022.
14. Diccionario De Mauro, https://dizionario.internazionale.it/, último acceso 09/04/2022.

Long Word Sequences in the Discourse of Adventure Tourism

Eva Lucía Jiménez-Navarro[1][0000-0001-9377-6921] and Isabel Durán-Muñoz[2][0000-0002-6795-498X]

[1,2] Department of English and German Studies, Universidad de Córdoba, Córdoba, Spain
lucia.jimenez@uco.es

Abstract. Tourism discourse as a domain-specific discourse is characterized by a set of linguistic, pragmatic, and function features that make it different from other discourses and the general language. One of its essential elements is the usage of appealing, innovative, exotic-sounding words in order to attract potential tourists by "persuading, luring, wooing and seducing" [6]. In this context, formulaic language plays a key role. To date, research into chunks of language used in tourism have mostly focused on collocations [1, 8, 23], with a few works on longer sequences [11, 12, 13].
Bearing this in mind, this paper aims to contribute to the analysis of 4-word bundles in this domain, more specifically, in the segment of adventure tourism. To do so, a corpus-driven analysis was undertaken. As for our methodology, a specialized corpus containing English promotional texts was compiled. After that, the software Sketch Engine was used to extract a list of potential 4-word bundles. Next, manual verification was performed to ensure the validity of the units. Finally, the resulting list was classified according to their structural framework and their function in the text. The findings show that, in terms of the structure, the most typical sequences were verbal bundles; on the other hand, in terms of the function, a significant amount of the units was mainly used to address readers directly.

Keywords: Adventure Tourism, 4-Word Bundle, Function, Structure.

1 Long Word Sequences in Specialized Discourse

Traditionally, phraseological units have been categorized according to their degree of fixedness and compositionality [5, 14, 21]. Thus, collocations are found at the end of one continuum and idioms at the other end. It means that the former are less structurally fixed and more semantically transparent than the latter. However, another criterion commonly set to identify typical word sequences has been frequency of use. This has been possible thanks to corpus linguistics and automatic software that allows the exploration of corpora.

A typical focus of corpus and phraseological studies has been the specialized discourse. In this context, not only has the emphasis been placed on collocations, but

research has also delved into longer sequences of words. For instance, structures of 3, 4, and 5 words have been analyzed in the field of applied linguistics [17]; 4-grams have been explored in scientific research articles [19]; complex nominals have been covered in the specialized domain of the environment [4]. As to the discourse of tourism, recurrent lexical bundles and phrase frames have been examined in hotel websites [11, 12, 13], concluding that the flexible elements of these sequences are content words which fill the slot in frames such as *will be [required, charged] to* or *we are [happy, delighted] to*. Regarding the subdomain of adventure tourism, two-word combinations have been covered both in English and Spanish [8, 18, 20], but longer sequences have not been examined yet.

Having said that, the main aim of this study is to contribute to the linguistic description of this field by analyzing the usage of 4-word bundles focusing on two aspects, their structure and their function in the discourse, which is where the contribution of this paper lies in. These multi-word combinations can be defined as "sequences of [four] words that show a statistical tendency to co-occur" [2]. The underlying hypothesis is that the discourse of adventure tourism can display an extensive range of phraseological units which evidence its degree of specialization, to clarify, its being regarded as a specialized discourse. In order to test this hypothesis, two are the stated objectives: first, we will identify the structural frameworks of these sequences of words, and second, we will address their function in the text.

This paper is organized according to the following sections: Section 2 describes the methodology employed to achieve our objectives; Section 3 explains and discusses the main results obtained; Section 4 presents the conclusions drawn as well as some lines of further research.

2 Methodology

This section will explore the methodological steps followed in order to attain the objectives of this study, which are: (1) the compilation of a specialized corpus, (2) the extraction of 4-word bundles, (3) their structural classification, and (4) their functional categorization.

2.1 Compilation of ADVENCOR EN

The first step to perform a linguistic study is the compilation of a reliable corpus, given that "The results are only as good as the corpus" [24]. For this reason, this paper presents a corpus-driven analysis of 4-word sequences extracted from a specialized 1,005,480-word English corpus about adventure tourism, which was automatically compiled using Sketch Engine. The texts selected were originally written in English, contemporary, and recently published in electronic format by public or private institutions, registered tourist companies, or travel agencies from English-speaking countries all over the world, such as the United Kingdom, the United States, and Ireland. The texts included were full texts, since they represent the genre under study better than

samples of a certain length would [10]. Regarding the level of specialization, these promotional texts represent a specialized/non-specialized communicative situation (from expert to non-expert), for their primary purpose was to woo tourists interested in adventure tourism (in general) and adventure activities (in particular).

ADVENCOR EN has already proved to be representative of the domain of adventure tourism and shed new light on the linguistic description of this segment. For instance, the keyness of adjectives has been examined and it has been discovered that they can be descriptive (e.g., *aerial, complimentary*) and evaluative (e.g., *lovely, pleasant*), being their aim to persuade the reader by contributing to the creation of mental representations of destinations [7]. On the other hand, motion verbs have been analyzed from a lexico-semantic perspective and it has been found that they explain how knowledge is expressed in this tourism segment [9]. Last but not least, collocations of motion verbs have also been studied and the main findings have been that collocates represent semantic roles of the argument structures [8, 18, 20].

2.2 Extraction of 4-word bundles

The second step of this study was the extraction of 4-word bundles typical of our specialized corpus. At this point, the 'N-grams' function available at Sketch Engine was used. The reason for exploring 4-word sequences rather than 3-/5-word sequences is that the former often subsume 3-word sequences [22]; in addition to that, they are much more frequent than 5-word sequences, offering a clearer range of structures and functions [15]. A frequency threshold of 20 tokens per million words was set [16], which means that 4-word bundles occurring at least 20 times in ADVENCOR EN were retrieved. This step produced a list of 234 items with a total frequency of 8,236 tokens. Nevertheless, we had to manually weed out some troublesome chunks for the following reasons:

1. They belonged to the name of a document included in the corpus, for instance, *activity tourism in wales, paragliding and hang gliding*.
2. They had been wrongly annotated, such as *m ore likely to*.
3. They only occurred in one specific context, not being representative of the whole corpus, for example, *price is per adult, for gift certificate redemptions*.
4. They made no sense in this study, such as *m o u n, n ta i n, av i n g*.

After this manual work, 76 items were discarded, so the final list of 4-word bundles amounted to 158 sequences.

2.3 Structural categorization of 4-word bundles

The next step in this investigation was the categorization of the final list of the units according to their structure. For this task, we contemplated the following classes based on Biber *et al.*[1] [3]: (1) nominal bundles, whose head is a noun (e.g., *his bristly short*

[1] These are classes which could embrace sequences containing a number of words other than four; in fact, the examples provided are taken from the authors and do not specifically show 4-word bundles.

hair, the journey back); (2) verbal bundles, whose head is a verb (e.g., *was walking, can see*); (3) adjectival bundles, whose head is an adjective (e.g., *so lucky, subject to approval by*); (4) adverbial bundles, whose head is an adverb (e.g., *fortunately enough, hardly ever*); (5) prepositional bundles, whose head is a preposition (e.g., *to him, in a street*). Additionally, we considered two more classes, conjunctions and full phrases (when they were registered in a dictionary as such).

2.4 Functional categorization of 4-word bundles

The final step of our methodology was the categorization of the 4-word bundles according to their function in the text. Thus, three broad categories along with their own subcategories were considered [15]:

1. Research-oriented sequences, used to structure the information:
   a. Location, which indicate time and place (e.g., *at the same time*).
   b. Procedure, concerning methods and processes (e.g., *the role of the*).
   c. Quantification, related to quantities (e.g., *a wide range of*).
   d. Description, used to describe facts (e.g., *the structure of the*).
   e. Topic, connected to the field of research (e.g., *the currency board system*).
2. Text-oriented sequences, which concern the organization of the text and the meaning of its elements as a message or argument:
   a. Transition signals, establishing additive or contrastive links between elements (e.g., *in addition to the*).
   b. Resultative signals, which mark inferential or causative relations (e.g., *as a result of*).
   c. Structuring signals, defined as text-reflexive markers which organize stretches of discourse or direct reader elsewhere in text (e.g., *in the next section*).
   d. Framing signals, used to specify limiting conditions (e.g., *in the case of*).
3. Participant-oriented sequences, focused on the writer or the reader of the text:
   a. Stance features, which convey the writer's attitudes and evaluations (e.g., *are likely to be*).
   b. Engagement features, addressing readers directly (e.g., *it should be noted*).

3 Results and Discussion

As it has been previously mentioned, the final list of 4-word bundles amounted to 158 items. The most recurrent units were *one of the most* (253 tokens), *is one of the* (243 tokens), and *one of the best* (108 tokens). Some of the least recurrent units (i.e., occurring 20 times in ADVENCOR EN) were *at the bottom of, is famous for its, the great barrier reef*. The following subsections show the results obtained in this study in terms of the structural framework and function of the sequences selected.

3.1 Structural features of 4-word bundles in adventure tourism

The first specific objective outlined in this research was the structural classification of the 4-word bundles selected. Table 1 displays this classification and shows the different structures identified organized according to the number of items, along with their overall frequency in the corpus (i.e., the total number of tokens), the percentage they occupy, and some examples:

Table 1. Structural classification of the 4-word bundles selected

Structure	No. of sequences	Overall frequency	Percentage	Examples
Verbal bundle	61	2,168	38.6	*to book your trip, you are looking for*
Nominal bundle	48	2,128	30.4	*impact of outdoor activity, the heart of the*
Prepositional bundle	33	1,227	20.9	*in the middle of, for the first time*
Adverbial bundle	6	191	3.8	*off when you spend, all over the world*
Adjectival bundle	4	93	2.5	*likely to participate in, are more likely to*
Conjunction	3	92	1.9	*but not limited to, so that you can*
Full phrase	3	99	1.9	*as well as a, thank you so much*
Total	158	5,998	100	

As shown in Table 1, there is a big difference between the three most frequent structural categories (verbal, nominal, and prepositional bundles, whose representation is over 20%) and the four least recurrent categories (adverbial and adjectival bundles, conjunctions, and full phrases, whose recurrence is below 5%).

Regarding the most frequent category, verbal bundles, more than a third of the items (26 sequences) incorporate a subject pronoun into the sequence, such as *you are interested in* and *we look forward to*, which makes emphasis on the potential tourist as well as the adventure activity's provider. With respect to the nominal bundles, one of the most recurrent structures consists of a noun phrase plus a preposition, especially *of*, for instance, *the base of the, a full day of*; other prepositions are *to* (e.g., *a departure date to, the best way to*) and *in* (e.g., *via ferrata in the, a dip in the*). As to the most common prepositions introducing prepositional bundles, we found *at* (7 tokens, e.g., *at the foot of, at the bottom of*), *in* (6 tokens, e.g., *in the middle of, in the united states*), and *of* (5 tokens, e.g., *of the most beautiful, of the world's most*), among others.

3.2 Functions of 4-word bundles in adventure tourism

The second specific objective stated in this study was the classification of the 4-word bundles selected according to the function they perform in the text. Table 2 represents

this classification, showing the specific categories/subcategories identified, the number of sequences, their overall frequency, and the percentage they occupy in the corpus:

Table 2. Functional classification of the 4-word bundles selected

Category/Subcategory	No. of sequences	Overall frequency	Percentage
Research-oriented	**96**	**4,101**	**60.8**
1. Location	32	1,238	33.4
2. Procedure	0	0	0
3. Quantification	23	1,321	24
4. Description	18	499	18.6
5. Topic	23	1,043	24
Text-oriented	**8**	**259**	**5**
1. Transition signals	2	77	25
2. Resultative signals	1	39	12.5
3. Structuring signals	0	0	0
4. Framing signals	5	143	62.5
Participant-oriented	**54**	**1,638**	**34.2**
1. Stance features	18	488	33.3
2. Engagement features	36	1,150	66.7
Total	158	5,998	100

As it can be observed in Table 2, the "research-oriented" category contains more than half (60.8%) of the units analyzed. These items are classified into four distinct subcategories, being the largest one "location" (33.4%), which includes units referring to time and place, such as *at the end of, from the top of*. The second place is occupied by two subcategories, given that both "quantification" and "topic" incorporate 24% of the sequences, for instance, *one of the largest* and *there are plenty of* ("quantification"), *please select another departure* and *experience the thrill of* ("topic"). Finally, "description" includes 18.6% of the units, such as *speeds of up to, had a great time*. Regarding the "procedure" subcategory, no 4-word bundles were identified.

In the second place, the "participant-oriented" category contains over a third (34.2%) of the chunks selected. This category is divided into two subcategories: (1) "engagement features" represents more than half (66.7%) of the units, probably because they are used to address readers directly, for example, *you will need to* and *if you wish to*, which makes sense considering that ADVENCOR EN comprises tourism promotional texts; (2) "stance features" entail sequences used to voice the writers of the texts' opinions, and occupy 33.3% of the structures included in the "participant-oriented" category, such as *can't wait to, we look forward to*.

Last but not least, the "text-oriented" category represents only 5% of the 4-word sequences. Most of them (62.5%) are used to specify limiting conditions in the "framing signals" subcategory, for instance, *with the help of* and *including but not limited*. After that, "transition signals" occupy 25% of these units and are used to describe addition, such as *as well as the*. Finally, only one unit (12.5%) was found to show result: *as a result of*. No structuring signals were identified in the corpus.

On the other hand, Table 3 represents the relation between the structures and the functions performed by the 4-word bundles selected:

Table 3. Structural frameworks used in terms of the functional classification

Structure	Research-oriented		Participant-oriented		Text-oriented	
Nominal bundle	42	43.8%	6	11.1%	0	0
Prepositional bundle	26	27%	4	7.5%	3	37.5%
Verbal bundle	22	23%	38	70.4%	1	12.5%
Adverbial bundle	4	4.2%	1	1.8%	1	12.5%
Adjectival bundle	1	1%	3	5.6%	0	0
Conjunction	1	1%	1	1.8%	1	12.5%
Phrase	0	0	1	1.8%	2	25%
Total	96	100%	54	100%	8	100%

Table 3 shows that each broad functional category is mostly characterized by a different structural framework. To put it differently, nominal bundle (43.8%) is the most recurrent structure identified in "research-oriented" sequences (e.g., *the edge of the, queensland adventure activity standards*), the head describing location, the topic of the texts, quantities, among others. On the other hand, verbal bundle (70.4%) is the most typical structure of "participant-oriented" bundles (e.g., *if you have any, give us a call*), for the verbs help to engage the readers of the texts and render the writers' opinions. Finally, prepositional bundle (37.5%) is the most common structure found in "text-oriented" sequences (e.g., *as a result of, in the event of*), being useful to organize the text.

4 Conclusions and Further Research

The current investigation has explored the structural and functional features of 4-word bundles in the specialized discourse of adventure tourism. In total, 158 sequences were selected after their automatic extraction and manual verification.

As for our first objective, the most common structure was verbal bundle (38.6%). This result may be surprising, as it is not closely related to the findings revealed in the achievement of our second objective, that is, the functions performed by the bundles. To explain, the vast majority of items (60.8%) were included in the "research-oriented" category and subcategorized into "location", "quantification", "topic", and "description", and most of the structures in these groups were nominal (43.8%) and prepositional bundles (27%). Nevertheless, it must be highlighted that 34.2% of the units were classified as "participant-oriented" sequences, from which the largest amount referred to "engagement features" (66.7%) and were verbal bundles (70.4%). It means that the most recurrent structure does not represent the most typical function of the bundles. However, it makes sense considering that the texts of the corpus were promotional texts about adventure tourism which aimed to attract tourists, therefore, a wide range of the units address the readers directly. This fact also demonstrates the specificity of this domain, thus confirming our hypothesis.

All in all, the objectives of this study have been successfully achieved. Future research may focus on shorter/longer bundles and other languages, which may allow contrastive studies. Additionally, this methodology may be applied to other segments of the tourism discourse (e.g., eco-tourism, sun-and-beach tourism) or other specialized domains (e.g., the environment or the academic discourse).

References

1. Baynat Monreal, M. E.: El léxico de la gestión turística en lengua francesa en el Diccionario Multilingüe de Turismo: Análisis contrastivo con la lengua inglesa. Çédille, Revista de Estudios Franceses 13, 53–82 (2017).
2. Biber, D., Conrad, S.: Lexical bundles in conversation and academic prose. In: Hasselgård, H., Oksefjell, S. (eds.) Out of corpora: Studies in honour of Stig Johansson, pp. 181–189. Rodopi, Amsterdam/Atlanta, GA (1999).
3. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E.: The Longman grammar of spoken and written English. Longman, London (1999).
4. Cabezas-García, M., Faber, P.: Phraseology in specialized resources: An approach to complex nominals. Lexicography 5(1), 55–83 (2018).
5. Cowie, A. P.: The treatment of collocations and idioms in learners' dictionaries. Applied Linguistics 2(3), 223–235 (1981).
6. Dann, G.: The language of tourism. A sociolinguistic perspective. CAB International, Wallingford (1996).
7. Durán-Muñoz, I.: Adjectives and their keyness: A corpus-based analysis of tourism discourse in English. Corpora 14(3), 351–378 (2019).
8. Durán-Muñoz, I., Jiménez-Navarro, E. L.: Colocaciones verbales en el turismo de aventura: Estudio contrastivo inglés-español. In: Corpas Pastor, G., Bautista Zambrana, M. R., Hidalgo-Ternero, C. M. (eds.) Sistemas fraseológicos en contraste: Enfoques computacionales y de corpus, pp. 121–142. Comares, Granada (2021).
9. Durán-Muñoz, I., L'Homme, M.-C.: Diving into English motion verbs from a lexico-semantic approach. A corpus-based analysis of adventure tourism. Terminology 26(1), 33–59 (2020).
10. Flowerdew, L.: The argument for using English specialized corpora to understand academic and professional language. In: Connor, U., Upton, T. A. (eds.) Discourse in the professions. Perspectives from corpus linguistics, pp. 11–33. John Benjamins Publishing Company, Amsterdam/Philadelphia (2004).
11. Fuster-Márquez, M.: Lexical bundles and phrase frames in the language of hotel websites. English Text Construction 7(1), 84–121 (2014).
12. Fuster-Márquez, M.: The discourse of US hotel websites: Variation through the interruptibility of lexical bundles. In: Gotti, M., Maci, S., Sala, M. (eds.) Ways of seeing, ways of being: Representing the voices of tourism, pp. 401–420. Peter Lang, Bern/Berlin/Brussels/Frankfurt am Main/New York/Oxford/Wien (2017).
13. Fuster-Márquez, M., Pennock-Speck, B.: Target frames in British hotel websites. International Journal of English Studies 15(1), 51–69 (2015).
14. Howarth, P. A.: Phraseology in English academic writing. Some Implications for language learning and dictionary making. Niemeyer, Tübingen (1996).
15. Hyland, K.: Academic clusters: Text patterning in published and postgraduate writing. International Journal of Applied Linguistics 18(1), 41–62 (2008).
16. Jalali, Z. S., Moini, M. R., Arani, M. A.: Structural and functional analysis of lexical bundles in medical research articles: A corpus-based study. International Journal of Information Science and Management 13(1), 51–69 (2015).
17. Jalilifar, A., Ghoreishi, S. M.: From the perspective of: Functional analysis of formulaic sequences in Applied Linguistics research articles. International Journal of English Studies 18(2), 161–186 (2018).

18. Jiménez-Navarro, E. L.: Treatment and representation of verb collocations in the specialized language of adventure tourism. Doctoral dissertation (Universidad de Córdoba, Cordoba, Spain) (2020).
19. Jiménez-Navarro, E. L.: A corpus-based study of 4-grams in the research article genre. ELUA 38, 241–262 (2022).
20. Jiménez-Navarro, E. L., Durán-Muñoz, I.: Collocations of fictive motion verbs in adventure tourism: A corpus-based study of the English language. RESLA (2022/forthcoming).
21. Mel'čuk, I. A.: Phraseology in the language, in the dictionary, and in the computer. Yearbook of Phraseology 3(1), 31–56 (2012).
22. Pérez-Llantada, C.: Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. Journal of English for Academic Purposes 14, 84–94 (2014).
23. Piccioni, S., Pontrandolfo, G.: La construcción del espacio turístico a través de la fraseología metafórica. Linguistik Online 94(1/19), 137–153 (2019).
24. Sinclair, J.: Corpus, concordance, collocation. Oxford University Press, Oxford (1991).

Phraseoculture in the construction of the corpus of the DiCoP: The treatment of the phraseographic microstructure

Lian CHEN 陈恋[1][0000-1111-2222-3333]

[1] CY Cergy Paris University, France - LT2D laboratory (Jean Pruvost Centre)
INALCO, France - PLIDAM laboratory
loselychen@gmail.com

Abstract. This article illustrates the need to introduce phraseocultural information into specialized bilingual (or even multilingual) dictionaries, from the perspective of intercultural communication. In this sense, we present the Dictionary and Corpus of Phraseology (DiCoP) project, which is currently a computerized resource for French–Chinese and Chinese–French phraseology and will eventually be multilingual. This innovative project in the field of phraseomatics (informatics phraseology/phraseography) affects many linguistics and computer fields, such as contrastive phraseoculture, metaphraseography, phraseography, phraseotraductology, and natural language processing (NLP). In particular, we wish to demonstrate the importance of phraseoculture in computerized dictionary data and in phraseography.
This paper focuses on how the DiCoP electronic dictionary reflects the phraseocultural information of the phraseological units in their microstructure. It uses examples of such microstructure to illustrate their feasibility.

Keywords: digital phraseography, phraseoculture, corpus linguistics

1 Theoretical framework: Phraseoculturology and contrastive phraseology

A phraseological unit (PUs), or *shúyǔ* 熟语, is a "[…] polylexical sequence consisting of two or more categorically related, contiguous or non-contiguous graphic [lexies]" (Bolly 2011: 28). It is characterized linguistically by the following:

> (i) a certain degree of syntactic fixidness (blocking of transformational properties and unalterable constituent order); and/or (ii) a certain degree of semantic fixidness (at least partial non-compositionality); and/or (iii) a certain degree of lexical fixidness (paradigmatic restriction); and/or (iv) a constraint on use in a communication situation. (Bolly 2011: 28)

PUs are "the embodiment of a country's culture.[1]" The study of this specific idiomatic cultural phenomenon in phraseology is called "linguo-culturology" (Teliya 1996;

[1] 每一条熟语，都是民族文化的化身。(Yao 2013: 157).

Dobrovol'skij 1998; Szerszunowicz 2010, etc.), "culture bound" (Szerszunowicz 2009), or "*shúyǔwénhuàxué* 熟语文化学" (Wang 2012). Inspired by R. Galisson's (1988) "lexiculture" and to address this particular phenomenon of phraseology and comparative linguistics, we proposed the concept of "phraseoculturology" (L. Chen 2022b). This sub-discipline is characterized by a specific methodology and theory within the framework of the study of the national, cultural particularities of PUs. This new phraseocultural perspective necessitates a more attentive treatment of fixed expressions in bilingual dictionaries, which often neglect this aspect.

Applied contrastive phraseology responds to the interests and needs that arise when translating, learning a language, or developing a bi- or multilingual dictionary. The analysis and comparison of PUs from one language to another leads us to distinguish three types of relationships. The following examples concern French and Chinese idiomatic expressions (IEs[2]):

1) "**Perfect equivalence**[3]" signifies an identical nature, meaning, syntactic structure, and keyword(s) in two expressions from different languages. For example, the French IE *sans queue ni tête* (negation+tail+negation+head) has a perfect equivalent in the Chinese IE (or *chéngyǔ*[4]) *méitóu-méiwěi* 没头没尾 (negation+head+negation+tail[5]).

2) "**Partial equivalence**" occurs when the expressions have the same "fixed" nature (proverbs, IEs, etc.) in both or only one of the languages, without necessarily being IEs. For example, the expression *avoir la grosse tête* (have a big/swelled head) is the partial equivalent of *xīngāo-qìào* 心高气傲 (heart+high+air+proud); it is an IE in both languages, but the keywords differ. Similarly, the French proverb *bouche de miel, cœur de fiel* (mouth of honey, heart of gall) is the partial equivalent of the Chinese IE *fókǒu-shéxīn* 佛口蛇心 (Buddha+mouth+snake+heart), "words of a Buddha and heart of a snake."

3) Finally, the most difficult translation/transmission concerns "**non-equivalence**." Fixed expressions (FEs) belong to this category of figures of speech that are rarely translated without loss, or that may remain misunderstood outside the language and culture of origin. Non-equivalence contains two aspects: a) "**Semantic vacancy**,"

---

[2] An IE is a polylexical sequence, semantically non-compositional (its global meaning is not always deducible from the meaning of the elements that formally compose it), syntactically fixed or fixed by usage (it does not always submit to the combinatorial rules which govern free syntax), and loaded with cultural implicits (L. Chen 2021: 123).

[3] Here, the term quasi-equivalence would be more appropriate. It is difficult to find two units identical in all respects, because units are "bound to differ, at least slightly, in some parameters" (Fiedler 2007: 118).

[4] French IEs correspond to Chinese *chéngyǔ* (polylexical sequences, non-compositional, syntactic blocking); both are characterized by a high degree of fixedness. The *chéngyǔ* has a quadrisyllabic, fixed basic form (Sabban [1978] 1980; Henry 2016). It is often an expression from classical literature used in modern Chinese as a compound lexie. Its style is elegant and concise, and it frequently contains strong allusive content (Ma 1978; Shi 1979; Liu 1985; Wen 2006, etc.). Even if modern creations also exist, these expressions are generally issued from a tradition—often oral and popular in French, more often written and literary in Chinese—and evoke daily realities in a metaphorical, very pictorial, and picturesque way.

[5] In our examples, we reproduce the pinyin transcription (Latin alphabet transcription of Mandarin Chinese pronunciation), the sinograms accompanied by their word-for-word translation, a literal translation, their figurative or implicit meaning, and an equivalent in French if possible.

which has a specific cultural dimension that cannot be transposed from one language to another, is explained by the specificities of each culture: customs, historical references (e.g., *avoir les dents du bonheur*; to have lucky teeth), toponyms (e.g., *oies du Capitole*; geese of the Capitol), literary sources (e.g., *n'avoir pas froid aux yeux*; to not be faint hearted), etc. b) "**Lexical vacancy**," wherein the words are not translatable from one language to another or the concept does not exist: in one language, there is a lexical vacancy, while in the other, the word has a specific cultural connotation (e.g., *dúzhànáotóu* 独占鳌头 alone+occupy+Ao[6]+head; all alone on the Great Dragon Turtle's head/to get first place in a competition/to be the champion/to top the list).
J. Chen (2004) has noted that the phenomenon of "cultural non-equivalence" (*wénhuà bù děng zhí* 文化不等值) should be considered in the translation of PUs and that it creates an obvious difficulty in the development of a bilingual dictionary. Without corresponding lexies in the target language, it is difficult for the lexicographer to respect or reflect the culture of the source language.

2 Phraseoculture, poorly treated in specialized French–Chinese and Chinese–French bilingual dictionaries: A metaphraseographic approach

According to Diaz (1981: 75), IEs should be approached from four points of view, which show the insufficiencies of the treatment of IEs in the dictionary: at the language level[7], at the level of use[8], at the level of discourse[9], and at the cultural level. Unfortunately, "the dictionary rarely provides information about the connotation that the linguistic community attributes to an IE and that the learner will not be able to establish on his or her own" (*Ibid.)*. Some dictionaries pay special attention to "linguoculturological" (Teliya et al. 2006 [2014]; Bartmińsky 1996, 1999, 2012)[10] or "lexicultural" (Rey et Chantreau 2003) principles in the analysis of phraseologisms. However, these dictionaries with a partly cultural vocation are mainly monolingual projects and therefore do not allow the determination of cultural specificities.

We looked for FEs carrying strong allusions (history, fable, mythology, culture, etc.) in bilingual dictionaries of FEs (three specialized Chinese–French dictionaries[11] and three French–Chinese dictionaries[12]) and analyzed them with a metaphraseographic approach. We find that phraseoculture is not satisfactorily addressed. The example of a

[6] Áo: Legendary turtle with dragon body.

[7] The learner who does not master the entire system may have doubts and risk not using the IE wisely. These productions deviating from generally accepted usage constitute errors of language.

[8] Learners seek to know if expressions are usual and frequent. The indications they can find in the dictionary (pop., fam., etc.) are too vague and imprecise and only very relatively define the levels of language and the variations of register.

[9] The absence of context makes it impossible for learners to foresee which situations they will be able to use an IE in.

[10] Quoted by Pamies (2017).

[11] (Doan and Weng 1999; Beijing Publishing House 1980; Sun 2012)

[12] (Yue and Xiao 2000; Sun 2010 ; Cai 2014)

Chinese FE from a famous historical reference on page 747 of Sun's Chinese–French dictionary (2012) demonstrates this:

> 卧薪尝胆 (臥薪嘗膽) wò xīn cháng dǎn: Coucher sur de la paille et goûter du fiel/couver sa vengeance à travers de dures épreuves/boire le calice jusqu'à la lie pour se venger au moment venu/« L'homme a souffert! Il a bu tous les calices » (J. Anouilh)

The meaning of this Chinese expression is as follows :

> [to sleep on brushwood and taste gall in order to recall one's humiliations/to maintain one's resolve for revenge]

The current Chinese–French bilingual dictionaries include simplified characters, traditional characters, and Pinyin transcriptions, and they offer a literal translation, a free translation, and when possible, an idiomatic equivalent in French. However, sources and phraseocultural explanations are missing. The absence of these prevents users from gaining a sufficient understanding from the word-for-word (*wòxīn-chángdǎn* 卧薪尝胆; sleep+firewood+taste+gallbladder) and global translations. A short historical addition (see below for the DiCoP) for expressions with strong allusive content would facilitate the mastery of these expressions and reinforce the role of the dictionary in terms of cultural transmission.

Similarly, French–Chinese dictionaries only offer an idiomatic, if it exists, and an example(s) of use. We can take the obvious example of the *grue* (crane), whose image is very positive in China, in contrast to in the French culture, where this bird is very negatively connoted. Sun's dictionary (2010), on page 589, presents the following:

> grue n.f. faire le pied de grue loc. v. 鹄立，久等
>
> ① Et d'indiquer que **(sic)** des yeux les dizaines de Noires et de mulâtresses qui font, tout près **(sic)** le pied de grue en se pavanant lascivement. (Le Point, 18-03-1995) 还要指出，几十个黑女人或黑白混血女人就在近旁猥亵作态地等客呢。 (Sun 2010: 589)
>
> [crane (noun). cool one's heels
>
> Example : ① And to indicate that the eyes of the dozens of black women and mulatto girls who are waiting nearby, strutting lasciviously.]

As this example demonstrates, these works are not free from approximations, gaps, or even errors and inaccuracies in their translations ("sic"). In these dictionaries, the presence of *chéngyǔ* is limited to the compilation of glossaries, and specialized dictionaries only provide meanings in the form of paraphrases and some examples, with few etymologies in Chinese. Bilingual Chinese–French dictionaries still require much development.

Cai's (2014) dictionary does offer specific cultural explanations. The author provides fairly thorough explanations for each FE, especially for those with strong allusive content. It provides examples of the use of these expressions and evokes their origins and metaphorical meanings. However, many FEs that are rich in phraseoculture are not included; for example, *être dans les bras de Morphée* (in the arms of Morpheus*)* or *avoir des yeux de lynx* (have eyes like a hawk/be lynx-eyed), etc.

Moreover, their particular phraseocultural dimension is not always associated with an appropriate translation. Take the example of *avoir la grosse tête* (have a big or swollen head/be full of oneself) in Cai (2014) on page 36. In the absence of a more precise comparative study of the two cultures, the translator does not propose a satisfactory correspondent in Chinese, and on the other hand does not give any clarification on the

connotation of the word "*cœur*" (heart) in Chinese, often synonymous with the French word "*tête*" (head) : *xīngāo-qìào* 心高气傲 (heart+high+air+proud). Indeed, there is a notable difference between the "holistic" conception of the heart in Chinese, and the Western dichotomy between "cardiocentrism" and "cerebrocentrism" (Yu 2009; L. Chen 2022a).

It will probably be stated that a paper dictionary cannot explain everything, unless it is thousands of pages long. This is the pitfall of paper dictionaries, which are subject to physical volume constraints and must therefore make choices to be able to remain easily manipulated. Electronic dictionaries, with a system of links, can circumvent this problem, which represents a major advantage that it would be a pity not to exploit.

In bilingual dictionaries, if the lexicographer is content to communicate the meaning, ignoring the social and cultural dimension, then the translation and the transfer function are incomplete. It is in this context that we propose the Dictionary and Corpus of Phraseology (DiCoP) project[13].

3 Innovation of phraseographic microstructure in the DiCoP: The treatment of the contrastive phraseoculture of fixed expressions

3.1 Presentation of the DiCoP project

In line with the current electronic era, the DiCoP project aims at developing an electronic dictionary of multilingual phraseology (currently bilingual), based on a corpus of phraseological units (collocations, proverbs, IEs, puns, defrosting, etc.) and associated databases to determine their frequency of use (in newspapers, literary works, manuals, etc.) in practice and thus their vitality, to improve their automatic translation, and with the aim of giving easier access to the phraseological units.

Within the context of our thesis (L. Chen 2021) on French and Chinese IEs and their respective translations/equivalences, we have built a corpus of 2,400 entries related to the human body and animals. We are continuing this work and enriching the lexicographical corpus with fields such as gastronomy, numbers, and plants. In the digital era, we believe it crucial to integrate phraseoculture into this DiCoP, which is inspired by the lexicographical tradition of *Le Robert*. The DiCoP is aimed at students, teachers, translators, and professionals from all disciplines.

3.2 Computerized dictionary data and the design of an innovative microstructure in the DiCoP: The addition of contrastive phraseoculture

The DiCoP is currently constituted on the basis of two corpora, Chinese–French and French–Chinese. It pays particular attention to phraseoculture, especially for PUs of partial equivalence or non-equivalence. Take the example of grue (crane) in French–Chinese in the DiCoP.

[13] Site: phraseologia.com (development in progress).

Grue n.f. **Faire le pied de grue** « attendre debout » [exp.] v. 鹄立，久等

词源: 该表达式来源于17世纪的固定表达faire la jambe de grue和 faire de la grue (和16世纪的faire la grue)。在Faire de la grue中，grue这个名词有动词“等待”的含义 (如在诗人 Maurice Scève作品中), 但是Bonaventure des Périers认为，在16世纪的语境中faire de la grue中的grue一词也含有鸟类：鹤的比喻义。

比喻义: 在 16 世纪，鹤的比喻用途通常是贬义的（être grue 愚蠢，suivre la multitude comme les grues [Calvin]，s'en aller comme des grues [ibid.] 像鹤一样跟随众人，没有自己的想法。以及 Le Roux 在 1752 年给出的变体：être planté comme une grue，与 être planté comme un sot同义 ： 一动不动地站着，等待很久，久等。最后，“妓女”的意思来自faire le pied de grue，人行道上妓女等待客人的形象比喻。

实例 : ①Et d'indiquer des yeux les dizaines de Noires et de mûlatresses qui font, tout près, le pied de grue en se pavanant lascivement. (Le Point, 18-03-1995) 并使眼色表明附近几十个黑人妇女和黑白混血女人正在作出淫荡姿态等客上门呢。

[crane (noun). **cool one’s heels**

Source: this expression succeeds in the 17th century to “faire la jambe de grue” and to the form “faire de la grue” (“faire la grue”, 16th century). In this last expression, “grue” (crane) appears to be used as a verbal noun, derived from the verb "gruer" (to wait) in Maurice Scève. However, the context in Bonaventure des Périers shows that it also metaphorically refers to a bird.

Metaphor: The figurative uses of crane in the 16th century are generally pejorative (“to be a crane” means to be a “fool” and to follow the multitude like the cranes [Calvin]. In addition, the variant given by Le Roux in 1752, “to be planted like a crane,” also meant “to be planted like a fool.” Finally, the meaning of “prostitute” comes from “faire le pied de grue,” that is, “to wait for the customer on the sidewalk”).

Example: ① And to indicate that the eyes of the dozens of black women and mulatto girls who are waiting nearby, strutting lasciviously.]

For these language-specific expressions, a phraseocultural annotation is essential: the crane in French has a meaning of prostitute.

In Chinese–French, we propose a more exhaustive microstructure based on simplified and traditional characters, Pinyin, the word-by-word translation of monosyllabic or dissyllabic characters or lexies, and the figurative meanings, sources, and pragmatic uses of FEs according to the following model:

卧薪尝胆 [臥薪嘗膽] (coucher, bois de chauffage, goûter, vésicule biliaire)

Pinyin: wò xīn cháng dǎn

Traduction Littérale: coucher sur de la paille et goûter du fiel

Signification Implicite: couver sa vengeance à travers de dures épreuves/boire le calice jusqu’à la lie pour se venger au moment venu

Source: « 史记 Shiji » (Mémoires du Grand Historien ou Mémoires historiques) (109 - 91 av. J.-C.),

Histoire: L’épisode se déroule, en 494 av. J.-C. Vaincu par l’empereur de l’Etat de Wu (吴), l’empereur Gou-Jian de l’Etat de Yue (越) décide de prendre sa revanche. Afin de ne pas oublier l'opprobre qui couvre son pays déchu et de s'affermir dans sa résolution de se venger, il dort sur de la paille et goûte souvent la sécrétion d'une vésicule biliaire suspendue au mur de sa chambre. Cet exercice de mortification le rend plus fort et il finit par vaincre l'Etat de Wu.

Exemple d’emploi: 一时失败何足畏，若有卧薪尝胆的壮志，一定能反败为胜。Il n'y a rien à craindre d'une défaite, car avec de l'ambition et en surmontant le épreuves, on peut la transformer en victoire.

This idiomatic expression originated from the following history:

> The episode takes place in 494 BC. Defeated by the Emperor of the State of Wu (吴), Emperor Gou-Jian of the State of Yue (越) decides to take his revenge. In order to not forget the opprobrium that covers his fallen country and to strengthen his resolve to take revenge, he sleeps on firewood and often tastes the gallbladder hanging on the wall of his room. This exercise of mortification makes him stronger, and he eventually defeats the State of Wu.

4 Conclusion

Bilingual lexicographers (French–Chinese and Chinese–French) J. Huang and C. Chen (2003: 101) note that "dictionaries are the index of culture.[14]" They are the key to opening the door to the knowledge and culture contained in semantics and constitute important reference works for learning and mastering foreign languages. For J. Pruvost (2006), the role of the dictionary as a "tool of a language and a culture" need no longer be demonstrated. As T. Szende (2003) notes, "Establishing identity relations between the terms of two languages within the framework of a bilingual dictionary is as much a linguistic operation as a cultural one." (mentioned by M. Murano 2011 : 60) The bilingual dictionary "is revealed today in its phase of transformation as a tool for reflection on culture" (Celotti 2002: 464), and "can also serve as a bridge between cultures, or (as cultures are not monolithic) between two sets of cultural understandings" (Rodger 2006: 572).

Closely linked to a socioculture that differs from one country to another, phraseoculturology is an object of attention in linguistics insofar as it is a matter of considering "a particular and fundamental dimension of words which, unfortunately, is lacking in lexicography as well as in dictionary" (Pruvost 2005: 16). As it represents a major source of difficulties, it should occupy a more substantial place in the dictionary, which is "the preferred medium for compiling all lexical units […] as well as specific idiomatic expressions" (L. Chen 2021: 278). This DiCoP project is part of a contrastive French–Chinese perspective and is therefore related to lexicology, phraseography, phraseotranslation (according to overall meaning, context, and related traits rather than word by word), and the development of natural language processing (NLP).

It could also be the subject of further phraseodidactic exploitation (such as the introduction of phraseological units in learning earlier and more regularly according to their opacity, frequency in daily life, etc.). Moreover, the advent and subsequent generalization of computer tools has considerably modified the modes of consulting and designing dictionaries, offering possibilities for enriching content, as well as the additional development models of contributory, collaborative, and participatory. DiCoP intends to take advantage of open annotation and commentary functions to allow for an intercommunication between readers and lexicographers in an interactive practice.

[14] 辞书是文化的索引。

References

1. Bolly, C.: Phraseology and collocations. Corpus-based approach in French L1 and L2 [Phraséologie et collocations. Approche sur corpus en français L1 et L2]. Peter Lang, Bruxelles, New-York (2011).
2. Cai, H. : Explanatory dictionary of French expressions and phrases [Dictionnaire explicatif des expressions et locutions françaises]. The commercial press, Beijing (2014).
3. Celotti, N.: Culture in bilingual dictionaries: where, how, which? [La culture dans les dictionnaires bilingues : où, comment, laquelle ?]. Studies in applied linguistics (128), 455-466 (2002).
4. Chen, J.: The phenomena of cultural inequality in the translation of English and Chinese idioms [谈英汉熟语翻译中的文化不等值现象 Tán yīnghàn shúyǔ fānyì zhōng de wénhuà bù děng zhí xiànxiàng]. Education and Career (18), 17-19 (2004).
5. Chen, L.: Phraseoculturology: an indispensable modern sub-discipline of phraseology [Phraséoculturologie : une sous-discipline moderne indispensable de la phraséologie]. In Proceedings of 8th International French Linguistics [Congrès Mondial de Linguistique Française - CMLF], SHS Web of Conferences 138, 04011, pp. 1-18. Université d'Orléans (2022b).
6. Chen, L.: Contrastive phraseocultural analysis: representation and motivation of the heart in French and Chinese [Analyse phraséoculturelle contrastive : représentation et motivation du cœur en français et en chinois]. In Proceedings of 8th International French Linguistics [Congrès Mondial de Linguistique Française - CMLF], SHS Web of Conferences 138, 11001, pp. 1-18. Université d'Orléans (2022a).
7. Chen, L.: Comparative analysis of idiomatic expressions in Chinese and French relating to the human body and animals [Analyse comparative des expressions idiomatiques en chinois et en français relatives au corps humain et aux animaux]. Thesis in Language Sciences, Cergy Paris university (2021).
8. Diaz, O.: Acquisition of idiomatic expressions in a foreign language [Acquisition des expressions idiomatiques en langue étrangère]. Doctoral thesis, Sorbonne Nouvelle University - Paris 3, Paris (1981).
9. Chinese-French Dictionary of Phrases and Proverbs [Dictionnaire chinois-français des locutions et proverbes]. Beijing Publishing House, Beijing (1979), Hong Kong (1980).
10. Doan, P. & Weng, Z.: Dictionary of chéngyǔ: quadrisyllabic idioms of the Chinese language [Dictionnaire de chéngyǔ : idiotismes quadrisyllabiques de la langue chinoise]. You-Feng Library, Paris (1999).
11. Dobrovol'skij, D. O.: On Cultural Component in the Semantic Structure of Idioms. In: Ďurčo, P. (ed.): Phraseology and Paremiology. International Symposium Europhras 97, pp. 55-61. Bratislava: Akadémia (1998).
12. Fiedler, S.: English Phraseology, Tu ̈bingen (2007).
13. Galisson R.: Cultures and lexicultures. For a dictionnary approach of the shared culture. [Cultures et lexicultures. Pour une approche dictionnairique de la culture partagée]. Cahiers d'Études Hispaniques Médiévales (7), 325-341 (1988).
14. Henry, K.: Chinese chéngyǔ: characterization of non-standard phrasemes. [Les chéngyǔ du chinois : caractérisation de phrasèmes hors normes]. Yearbook of Phraseology (7), 99-126 (2016).
15. Huang, J. & Chen, C.: Introduction to Bilingual Lexicography (Revised version) [双语词典学导论Shuāngyǔ cídiǎn xué dǎolùn]. The commercial press, Beijing (2003).
16. Liu, J.: Chéngyǔ [成语 chéngyǔ]. The commercial press, Beijing ([1985] 2000).

17. Ma, G.: Chéngyǔ [成语 Chéngyǔ]. 2 edition. Inner Mongolia People's House, Hohhot (1978).
18. Murano, M.: The treatment of Fixed Sequences in French-Italian and Italian-French bilingual dictionaries [Le traitement des Séquences Figées dans les dictionnaires bilingues français-italien, italien-français]. Polimetrica, International Scientific Publisher, French (2011).
19. Pamies, A.: The Concept of Cultureme from a Lexicographical Point of View. De Gruyter, Open Linguistics (3), 100–114 (2017).
20. Pruvost, J.: French dictionaries, tools of language and culture [Les dictionnaires français, outils d'une langue et d'une culture]. Ophrys, Paris (2006).
21. Pruvost, J.: Some operational lexicographic concepts to be promoted on the threshold of the 21st century [Quelques concepts lexicographiques opératoires à promouvoir au seuil du XXIe siècle]. Studies in applied linguistics: Journal of didactology of languages-cultures and lexiculturology (137), 7-37 (2005).
22. Rey, A. et Chantreau, S.: Dictionnaire d'expressions et locutions.Le Robert, Paris. ([1997] 2003).
23. Rodger, L.: Beyond Butterscotch. The Place of Cultural Knowledge in the Bilingual Dictionary. In: Corino, E., Marello, C., Onesti, C.(eds.), Proceedings XII EURALEX International Congress, vol. 1, pp. 567– 573. Alessandria (2006).
24. Sabban, F.: The quadrisyllabic idiomatisms of modern Chinese [Les idiomatismes quadrisyllabiques du chinois moderne, 现代汉语四字格成语 Xiàndài hànyǔ sìzìgé chéngyǔ]. Thesis, EHESS Paris VII. ([1978] 1980)
25. Shi, S.: Study of chéngyǔ [汉语成语研究 Hànyǔ chéngyǔ yánjiū]. Sichuan People's Publishing House, China, Sichuan province. (1979).
26. Sun, Q.: New Chinese-French Dictionary of Phrases and Proverbs [Nouveau dictionnaire chinois-français des locutions et proverbes]. Xiamen University Press, China, Xiamen city (2012 [1999]).
27. Sun, Q.: New French-Chinese Dictionary of Phrases and Proverbs [Nouveau Dictionnaire Français-Chinois des Locutions et Proverbes]. Xiamen University Press, China, Xiamen city (2010).
28. Szerszunowicz, J.: On cultural connotations of idioms expressing language users collective memory in a comparative perspective. In: Korhonen, J. , Mieder, W., Rosa Piñel E. (eds) Phraseologie global – areal – regional. pp. 317-324. Tübingen: G. Narr Verlag (2010).
29. Teliya, V.: Russian phraseology. Semantic, pragmatic and linguo-culturological aspects. [Русская фразеология. Семантический, прагматический и лингво-культурологический аспекты]. Moscow: School Languages of Russian Culture. [Москва: Школа "Языки русской культуры] (1996).
30. Wen, D.: Chinese lexicon course [汉语词汇学教程 Hànyǔ cíhuì xué jiàochéng]. The commercial press, Beijing (2006).
31. Yao X.: Compendium of chinese phraseological units [熟语学纲要 Shúyǔ xué gāngyào]. Elephant Press, Zhengzhou (2013).
32. Yu, N.: The Chinese Heart in a Cognitive Perspective: Culture. Body. and Language. Mouton de Gruyter, Berlin (2009).
33. Yue, Y. & Xiao, Z.: French-Chinese Dictionary of Phrases and Proverbs [Dictionnaire Français-Chinois des Locutions et Proverbes]. Shanghai Translation Publishing House, Shanghai (1999 [2000]).

Frecuencia de uso de locuciones y paremias en el corpus Spanish Web 2018 (esTenTen18): implicaciones didácticas y lexicográficas

Enrique Gutiérrez Rubio[1][0000-0001-8877-4446]

[1] Palacký University Olomouc, Křížkovského 10, 779 00 Olomouc, Czech Republic
enrique.gutierrez@upol.cz

Resumen. En este trabajo se presentan la metodología y los resultados de un análisis de frecuencia de uso de 27 locuciones y 27 paremias españolas en un subcorpus de miles de millones de palabras que forma parte del "gigacorpus" esTenTen18. El objetivo es comprobar si, en los niveles formal y semiformal de la lengua, la presencia de paremias resulta tan minoritaria como en su nivel informal, hecho que ya quedó demostrado en un trabajo anterior realizado sobre la base de tres corpus orales del español. Los resultados presentados en este estudio demuestran una frecuencia de uso mucho mayor, aproximadamente 6,5 veces superior, de las locuciones respecto a las paremias. Así, los datos obtenidos revelan que sería recomendable replantearse la presencia de paremias en el aula de español como lengua extranjera (ELE). Además, se han documentado diferencias muy significativas en la frecuencia de uso de las distintas locuciones analizadas, de lo que también se infiere que resulta indispensable incluir la frecuencia de uso entre los criterios de selección de las UF incluidas en los manuales de ELE.

Palabras clave: Fraseología, gigacorpus, español como lengua extranjera.

1 Introducción[1]

A pesar de la revolución vivida en el campo de los estudios fraseológicos del español en las últimas tres décadas y que ha conllevado la proliferación de trabajos teóricos y aplicados, incluidas las primeras obras fraseográficas de cierta extensión y dirigidas por principios científicos, en la actualidad siguen echándose en falta trabajos que, de un modo sistemático y con base en los medios tecnológicos más actuales, investiguen aspectos tan básicos como el que trata este breve artículo: la frecuencia de uso de las unidades fraseológicas (UF).[2]

[1] La preparación y publicación de este trabajo ha sido posible gracias a la financiación del proyecto de investigación IGA_FF_2022_025 otorgado a la Universidad Palacký de Olomouc por el Ministerio de Educación, Juventud y Deporte de la Republica Checa.

[2] Seguimos, pues, la taxonomía de los tres trabajos que suelen tomarse como referencia en la historia de la fraseología española –Casares 1992 [1950], Zuluaga (1980) y Corpas Pastor (1996)–, que incluyen los refranes entre los objetos de estudio de esta disciplina lingüística.

Por otra parte, los materiales lexicográficos y didácticos, precisamente por la ausencia de datos precisos a este respecto, se han creado, y continúan haciéndolo, sobre la base de la intuición y el conocimiento personal de sus autores, lo que, evidentemente, no resulta una metodología rigurosa, capaz de garantizar la selección de las unidades más frecuentes. La falta de índices de frecuencia de las UF es algo que lamentan incluso los propios autores de materiales didácticos (cf. Penadés Martínez 1999: 26).

En uno de los pocos trabajos que tratan el tema de los criterios de selección de las UF para el aula de ELE, Velázquez Puerto (2018: 35 y ss.) –entre otras muchas recomendaciones, como el nivel del estudiante, el grado de idiomaticidad de las UF, la lengua materna del estudiante o su trasfondo cognoscitivo– se refiere explícitamente al tipo de UF. Esta autora defiende que habrían de ser las colocaciones y las locuciones los tipos que deberían introducirse primero en el aula de ELE (Velázquez Puerto 2018: 36). Así, se posiciona patentemente en contra de la opinión de Morvay (1997: 424), quien defiende que son precisamente las paremias el "material idóneo para iniciar a los estudiantes en el dominio de la Fraseología [, ya que al tratarse] de fenómenos muy similares en varias lenguas (europeas) permite apoyarse en la competencia lingüística que tienen los alumnos en su lengua nativa". La causa por la que Velázquez Puerto (2018: 36) se muestra reacia a la enseñanza de las paremias (o enunciados fraseológicos) reside exclusivamente en su elevado grado de idiomaticidad. Así, esta autora no entra a considerar el papel que ejerce la frecuencia de uso de esta clase de UF. El tema de la frecuencia respecto a la selección de UF para el aula de ELE (y para los diccionarios fraseológicos) sí es tratado por Penadés Martínez (1999: 24, 33), quien señala que "[...] siempre será más conveniente presentar a los que [aprenden una lengua] las locuciones más frecuentemente utilizadas para que las conozcan" (2015: 179). Un factor este que, de acuerdo con García-Page (1995: 160), habría de influir, además, sobre el "correcto empleo" de una UF dada por parte de los aprendientes de ELE.

Por otro lado, son numerosos los trabajos en que se ha discutido, sin aportar datos concluyentes, si las paremias son un recurso frecuente en el español actual o si, especialmente entre las jóvenes generaciones, su uso se halla en claro retroceso. En contra de las teorías más pesimistas de Combet (1971: 300), quien incluso considera que, en este grupo etario, el uso de refranes se asocia a un retraso cultural y a una inferioridad social –aunque más tarde (1996: 22) afirmará que los refranes no están desapareciendo, sino evolucionando–, Corpas Pastor (1996: 166) señala que "las paremias gozan de una excelente salud", si bien su aseveración se refiere a un corpus formado principalmente por prosa periodística y literaria. Por el contrario, autores como Ruiz Gurillo (1998: 47), Sancho Cremades (1999: 86-87) o Gutiérrez Rubio (2021: 92) consideran que su frecuencia en el español actual es muy baja, especialmente en el registro coloquial.

La existencia desde hace unos años de "gigacorpus" textuales compuestos por miles de millones de palabras permite que esta falta de datos objetivos pueda ser, al menos en parte, superada y que, por tanto, las futuras obras de fraseología aplicada tengan en cuenta la frecuencia de uso de las UF.

En este estudio de carácter cuantitativo vamos a tratar el caso específico de la diferente frecuencia de uso de las locuciones frente a las paremias (término empleado aquí como hiperónimo de todo tipo de UF como *aforismos, proverbios, sentencias*, etc., pero que, en la mayoría de los casos, equivale simplemente a *refranes*). La inspiración la

encontramos en una investigación previa en la que se constató la escasísima frecuencia relativa de uso de las paremias en un corpus oral compuesto preeminentemente de conversaciones altamente informales (corpus Val.Es.Co. y conversaciones extraídas del concurso *Gran Hermano*[3]), si bien aproximadamente un 19 % del material procedía de entrevistas de radio y televisión recogidas en el corpus CREA y, por tanto, pertenecientes a un registro más formal. Así, en un corpus oral de 53 050 palabras se documentaron apenas trece paremias, algunas de carácter dudoso, además. De hecho, seis de ellas no se recogían en ninguno de los seis diccionarios y glosarios especializados consultados (cfr. Gutiérrez Rubio 2021: 137-141). Se trata, en definitiva, de 0,25 paremias por cada mil palabras frente, a modo de ejemplo, a las 4,24 locuciones verbales (ocurrencias) documentadas también por mil palabras en el mismo corpus oral. En otras palabras, las conversaciones y entrevistas analizadas incluían aproximadamente 17 veces más locuciones verbales que paremias.

En este sentido, en un artículo dedicado específicamente al análisis de la fraseología en los materiales ELE, se demostró que en las 30 041 palabras analizadas –procedentes en este caso exclusivamente del concurso *Gran Hermano*– no se documentaba ni uno solo de los refranes recogidos en el manual *Hablar por los codos* (Vranic 2004) y en el glosario trilingüe *100 imágenes en la punta de la lengua* (Lamothe y Gingras 2009), dos de los escasos materiales dedicados exclusivamente a la enseñanza de UF. Esta escasísima frecuencia de uso de las paremias –demostrada para contextos altamente coloquializadores y hablantes de entre veinte y treinta años– nos llevó a afirmar que los autores de materiales didácticos especializados deberían replantearse la inclusión de este tipo de UF en sus manuales y, muy especialmente, en aquellos diseñados para mejorar las destrezas en aprendientes jóvenes (cfr. Gutiérrez Rubio 2020: 13).

Remarcamos el hecho de que las conversaciones que forman el corpus analizado cumplen de un modo *extremo* los cuatro principales rasgos coloquializadores propuestos por Briz Gómez (2010b: 126) y, pueden, por tanto, ser situadas en el extremo más coloquial del contínuum informal-formal. Lo que pretendemos en el estudio presentado en este artículo es comprobar si las conclusiones basadas en un breve corpus oral de conversaciones espontáneas resultan igualmente válidas para los datos obtenidos de un gigacorpus escrito y al que, aunque esté compuesto de un auténtico cajón de sastre de textos de las más diversas procedencias, se le presupone un carácter entre formal y semiformal.

2 Principios metodológicos

Como ya hemos adelantado en el apartado 1, la existencia de gigacorpus que recogen miles de millones de palabras permite afrontar la investigación lingüística desde una nueva dimensión metodológica cuantitativa que con los corpus que cabría denominar *tradicionales*, como los CREA y CORDE de la RAE, no era posible. Esta afirmación resulta especialmente válida para el estudio de la fraseología, ya que, como afirman

[3] Puede encontrarse una justificación relativa a la idoneidad de emplear las conversaciones mantenidas entre los concursantes de este *reality show* como fuente de material lingüístico espontáneo en Gutiérrez Rubio (2021: 233-237).

Corpas Pastor (2021: 4) o Penadés Martínez (2015: 76), las UF (y más específicamente las locuciones) suelen ser elementos de frecuencia de uso muy baja. Para este estudio emplearemos el material textual tomado de páginas webs en español y recogido en el Spanish Web 2018 dentro del corpus esTenTen18. Dado que estamos interesados exclusivamente en el español peninsular, hemos restringido nuestras búsquedas al subcorpus de páginas web con dominio ".es" (_European Spanish domain .es), que forma el 17,5 % del Spanish Web 2018 y recoge 3.421.734.353 formas *(tokens).*

Otro aspecto metodológico fundamental se refiere a qué UF buscar en el corpus. Para este propósito hemos partido del material recogido en uno de los pocos materiales ELE dedicados exclusivamente al proceso de enseñanza/aprendizaje de la fraseología: *Hablar por los codos: Frases para un español cotidiano* (Vranic 2004). Este manual recopila 175 "frases hechas" (en su inmensa mayoría locuciones verbales) y 45 refranes que, a tenor de la información de su contraportada, habrían de ser UF "de uso frecuente en el español cotidiano actual" (Vranic 2004: contraportada). Nada puede asegurarnos, sin embargo, que se trate realmente de UF de uso habitual en español actual, ya que, nuevamente, el material habría sido muy probablemente seleccionado por su autora de forma intuitiva. Sin embargo, damos por hecho que Vranic realizó una cuidadosa criba, acaso basada en otros manuales o glosarios, para priorizar UF de uso frecuente frente a otras arcaicas. Aun así, debemos reconocer que esta selección debió verse necesariamente influida por la imagen (subyacente o léxica) de las UF, dado que todas ellas vienen acompañadas por un dibujo que habría de ilustrar el uso o significado de las UF, así como, en muchos casos, por la explicación de su origen.

Una de las mayores dificultades metodológicas del análisis fraseológico a la hora de trabajar con gigacorpus de forma automática es realizar búsquedas que garanticen que los datos obtenidos se refieren exclusivamente a UF y no a combinaciones libres de palabras o a las formas no figurativas que frecuentemente se hallan tras las UF. Para minimizar estos riesgos, hemos tomado las siguientes decisiones metodológicas:

–Realizaremos búsquedas con el motor Concordance y el tipo de consulta *(query type)* "phrase", que encuentra combinaciones de palabras idénticas a las escritas en el buscador. De este modo, se reducirán las probabilidades de incluir en los resultados de la búsqueda usos no fraseológicos que coinciden formalmente con la UF en cuestión.

–Excluiremos de la búsqueda las formas verbales variables (conjugables), centrándonos en combinaciones de, al menos, tres palabras invariables seguidas. De ahí que UF como *tirar los tejos* o *tener manga ancha* no sean aptas para nuestra búsqueda, pero sí *hacer leña del árbol caído* o *al pie de la letra.*

–Descartaremos las UF cuya combinación de tres o más palabras invariables pueda formularse con cierta frecuencia en su forma no figurativa. De esta manera, la posibilidad de hallar una combinación libre de "peras al olmo" (de *pedir(le) peras al olmo*) resulta, *a priori* al menos, prácticamente imposible; todo lo contrario cabe afirmar de "en el bolsillo" o "de la lengua" respecto a las UF *meterse (a alguien) en el bolsillo* y *tirar (a alguien) de la lengua.*

– Por otra parte, se realizarán búsquedas del menor número posible de palabras invariables que cumplan los requisitos arriba señalados. De este modo, se pretende incluir en los resultados variantes acortadas e incluso creativas o erróneas de las UF. Esto resulta especialmente relevante en el caso de los refranes, para los que, precisamente por

la elevadísima fijación de su forma, resulta relativamente frecuente documentar el empleo de tan solo la primera parte de su estructura bimembre.

–Cuando se registren variantes –como en *quien / el que se pica ajos come*– daremos prioridad a las combinaciones de palabras invariables, "se pica ajos" en este supuesto concreto. En el caso de los refranes, como señalábamos más arriba, trataremos de realizar, siempre que sea posible, búsquedas centradas en su primer miembro.

–Para las búsquedas no emplearemos las formas recogidas en *Hablar por los codos,* sino las registradas en el diccionario en línea *DiLEA* (Penadés Martínez 2019) para las locuciones y el *Refranero multilingüe* del Centro Virtual Cervantes para las paremias.

–A pesar de que el análisis se base en un tratamiento automático de los datos, comprobaremos siempre las primeras veinte entradas y, en caso de documentar ejemplos que no se correspondan a UF, propondremos una nueva combinación de palabras para la búsqueda o, de ser esto imposible, eliminaremos la UF en cuestión de nuestro listado.

Una vez cumplidos los requisitos arriba expuestos, hemos creado un listado de treinta locuciones (verbales y adverbiales) y treinta paremias (en su mayoría, siempre de acuerdo con la clasificación del *Refranero multilingüe,* refranes, aunque se documentan también una locución proverbial, una frase proverbial y un proverbio). De esta primera lista, tres UF de cada grupo fueron eliminadas por mostrar en los primeros veinte resultados de la búsqueda entradas no fraseológicas. El listado definitivo ordenado alfabéticamente de las 54 UF es el siguiente:

Locuciones: (1) *aferrarse/agarrarse/cogerse (como) a un clavo ardiendo;* (2) *aguantar carros y carretas / pasar por carros y carretas;* (3) *al pie de la letra;* (4) *al pie del cañón;* (5) *alzarse/quedarse con el santo y la limosna;* (6) *atar los perros con longaniza(s);* (7) *caer/llover chuzos de punta;* (8) *coger el/al toro por los cuernos;* (9) *con pies de plomo;* (10) *con uñas y dientes;* (11) *dar gato por liebre;* (12) *darse con un canto en los dientes;* (13) *de la ceca a la meca;* (14) *empezar la casa por el tejado;* (15) *entre la espada y la pared;* (16) *entre Pinto y Valdemoro;* (17) *hacer leña del árbol caído;* (18) *irse/salir(se) por los cerros de Úbeda;* (19) *matar dos pájaros de un tiro;* (20) *pedir(le) peras al olmo;* (21) *poner el dedo en la llaga;* (22) *poner los dientes largos;* (23) *sacar las castañas del fuego;* (24) *tener el santo de cara;* (25) *tener el santo de espaldas;* (26) *tener la sartén por el mango;* (27) *tirar/echar la casa por la ventana.*

Paremias: (1) *a caballo regalado, no le mires el diente;* (2) *a cada cerdo le llega su San Martín;* (3) *a Dios rogando y con el mazo dando;* (4) *a enemigo que huye, puente de plata;* (5) *a falta de pan, buenas son tortas;* (6) *a la cama no te irás sin saber una cosa más;* (7) *a mal tiempo, buena cara;* (8) *ande yo caliente, y ríase la gente;* (9) *aunque la mona se vista de seda, mona se queda;* (10) *cada maestrillo tiene su librillo;* (11) *casa con dos puertas, mala es de guardar;* (12) *con pan y vino se anda el camino;* (13) *cuando las barbas de tu vecino veas pelar, pon las tuyas a remojar;* (14) *de noche, todos los gatos son pardos;* (15) *desnudar un santo para vestir otro;* (16) *dinero llama dinero;* (17) *el que come y canta algún sentido le falta;* (18) *mal de muchos, consuelo de tontos;* (19) *más vale pájaro en mano que ciento volando;* (20) *no se hizo la miel para la boca del asno;* (21) *quien a buen árbol se arrima buena sombra le cobija;* (22) *quien come la carne, que roa el hueso;* (23) *quien se pica ajos come;* (24) *un clavo saca otro clavo;* (25) *unos nacen con estrella y otros estrellados;* (26) *unos tienen la fama, y otros cardan la lana;* (27) *zapatero, a tus zapatos.*

3 Interpretación de los datos obtenidos

Los datos que se desprende del análisis de las 54 UF de acuerdo con los principios metodológicos presentados en la sección 2 pueden observarse en la Figura 1 (los números en el eje horizontal se corresponden a las UF listadas en el apartado anterior).

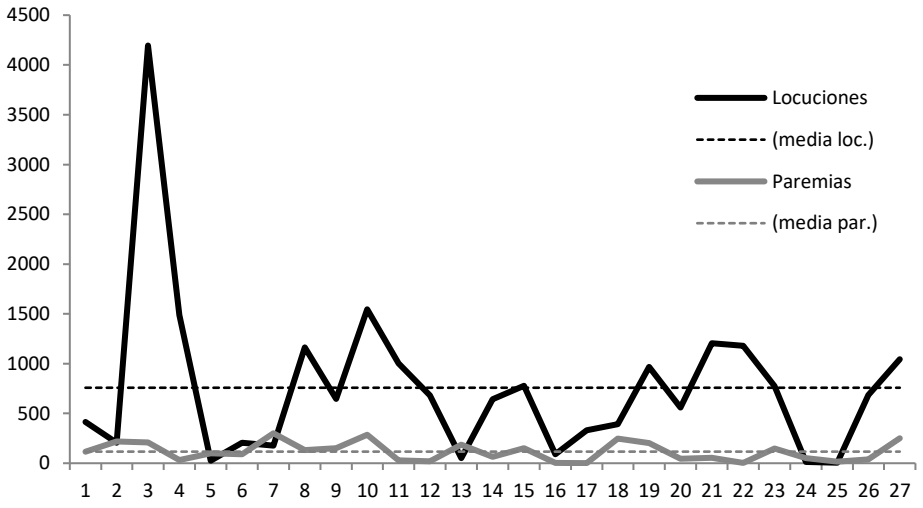


Fig. 1. Número de *tokens* por cada millón de formas documentados en el subcorpus "_European Spanish domain .es" del corpus "esTenTen18" para las 54 UF objeto de estudio.

De los datos presentados en la Figura 1 se pueden extraer dos relevantes conclusiones:

a) La media de *tokens* obtenidos del análisis de las locuciones es sustancialmente más elevada que la de las paremias; concretamente se han documentado 20.474 *tokens* (0,222 por millón de formas) en el caso de las locuciones frente a los 3.145 *tokens* (0,034 por millón de formas) en el de las paremias. En otras palabras, las locuciones en este corpus son de media aproximadamente 6,5 veces más frecuentes que las paremias.

b) La frecuencia de las distintas UF, tanto locuciones como paremias, destaca por su irregularidad. Así, la UF más común es la locución adverbial *al pie de la letra* con 4.195 *tokens*, lo que contrasta con la locución verbal *tener el santo de espaldas,* que registra apenas cinco entradas, es decir, su frecuencia en el subcorpus es 839 veces menor. En el caso de las paremias, para *a mal tiempo, buena cara* se documentan 301 *tokens* frente a la única entrada de *el que come y canta algún sentido le falta.*

Otros datos destacables se refieren a que las 54 UF, sin excepción, se documentan, al menos, una vez en el corpus, si bien una locución y tres paremias presentan cinco o menos *tokens*. Por otra parte, de los datos se deduce que las locuciones adverbiales tienden a emplearse con mayor frecuencia que las verbales, como indicaría el hecho de que las tres UF más frecuentes son precisamente tres locuciones adverbiales: *al pie de la letra* (4.195 *tokens*), *con uñas y dientes* (1.545) y *al pie del cañón* (1.489). De hecho, la media de *tokens* registrados en el corpus es de 1.336 para las locuciones adverbiales

frente a 593 para las verbales. No obstante, las cifras podrían depender del escaso número de locuciones adverbiales de nuestro inventario y no son, por tanto, concluyentes.

Por último, constatamos que no todas las locuciones son más frecuentes que las paremias: hay seis paremias[4] que –con un número de entradas que oscila entre 301 y 210– son más frecuentes que las ocho locuciones con valores de presencia más bajos.[5]

4 Discusión y conclusiones del análisis

El análisis realizado con base en un gigacorpus de miles de millones de palabras aporta información muy relevante sobre distintos aspectos del estudio lingüístico y, especialmente, sobre el uso *real* de las UF, incluida su frecuencia de uso. Los datos obtenidos al investigar la frecuencia de uso de 54 UF muestran que las locuciones son mucho más frecuentes (aprox. 6,5 veces) que las paremias. Sin embargo, dentro de cada grupo se han registrado sustanciales diferencias, habiéndose demostrado que hay paremias que presentan una frecuencia de uso mayor que algunas de las locuciones. Además, no debemos olvidar que estos datos se corresponden tan solo al español escrito y, por tanto, únicamente a sus registros formal y semiformal. Consideramos que, en el caso del registro oral informal, la frecuencia de uso podría variar considerablemente. No obstante, resulta difícil, si no directamente imposible, la creación de gigacorpus orales y, más aún, de aquellos que recojan material lingüístico perteneciente al eje informal.

Por otra parte, podemos ratificar que, como afirman algunos especialistas, las UF son elementos de frecuencia de uso muy baja, ya que tan solo una de las 54 UF analizadas se registra en más de un caso por millón de formas, mientras que dos locuciones y cuatro paremias muestran una frecuencia tan extremadamente baja que se redondea a cero *tokens* por millón de formas. Además, queda demostrado que ni siquiera en los textos que conforman el corpus (lejos del registro coloquial de las jóvenes generaciones) el uso de las paremias se puede considerar un elemento común, ya que incluso el refrán más frecuente se usa de media menos de una vez cada diez millones de formas.

Cerramos este trabajo afirmando que, ante la dificultad que entraña el aprendizaje de las paremias y su escasísima frecuencia de aparición en todo tipo de registros, creemos razonable minimizar su presencia en el aula de ELE e incluso excluirlas de los niveles más bajos de enseñanza. Respecto a las locuciones, consideramos indispensable realizar un riguroso cribado en que, junto a las recomendaciones de otros especialistas, se tome muy en cuenta la frecuencia de uso a la hora de seleccionar las UF como paso previo a la confección de materiales ELE. Queda pendiente, además, la creación de un ambicioso índice de frecuencia fraseológica que habría de servir de base para los futuros trabajos de fraseología aplicada y, muy especialmente, para la confección de materiales de ELE y la compilación de diccionarios, tanto fraseológicos como generales.

---

[4] Concretamente, *a mal tiempo, buena cara; cada maestrillo tiene su librillo; zapatero, a tus zapatos; mal de muchos, consuelo de tontos; a cada cerdo le llega su San Martín; a Dios rogando y con el mazo dando.*

[5] Concretamente, *aguantar carros y carretas / pasar por carros y carretas; atar los perros con longaniza(s); caer/llover chuzos de punta; entre Pinto y Valdemoro; de la ceca a la meca; alzarse/quedarse con el santo y la limosna; tener el santo de cara; tener el santo de espaldas.*

Referencias bibliográficas

1. Briz Gómez, A.: Lo coloquial y lo formal, el eje de la variedad lingüística. In: Castañer Martín, R. M., Lagüéns Gracia, V. (coord.) *De moneda nunca usada: Estudios dedicados a José Mª Enguita Utrilla,* pp. 125–133. Instituto Fernando El Católico, CSIC, Zaragoza (2010).
2. Casares, J.: *Introducción a la lexicografía moderna*. CSIC, Madrid (1992 [1950]).
3. Centro Virtual Cervantes: *Refranero multilingüe*, https://cvc.cervantes.es/lengua/refranero/, último acceso 2/5/2022.
4. Combet, L.: *Recherches sur le "refranero" castillan.* Belles Lettres, Paris (1971).
5. Combet, L.: Los refranes: origen, función y futuro. *Paremia* 5, 11–22 (1996).
6. Corpas Pastor, G.: *Manual de fraseología española*. Gredos, Madrid (1996).
7. Corpas Pastor, G.: Constructional idioms of 'insanity' in English and Spanish: A corpus-based study. *Lingua* 254, 1–20 (2021).
8. García-Page, M.: Problemas en el empleo de la fraseología española por hablantes extranjeros: la violación de restricciones. In: Grande Alija, F. J. *et al.* (coord.) *Actuales tendencias en la enseñanza del español como lengua extranjera II: actas del VI Congreso Internacional de ASELE (León, 5-7 de octubre de 1995)*, pp. 155–162. Universidad de León, León (1995).
9. Gutierrez Rubio, E.: Materiales didácticos y fraseología en el discurso oral coloquial. *Lingüística en la Red*, anexo monográfico XVII, 1–18 (2020).
10. Gutierrez Rubio, E.: *Fraseología española en el discurso oral.* Tirant lo Blanch*,* Valencia (2021).
11. Lamothe, R., Gingras, R.: *100 imágenes en la punta de la lengua* (2009), http://expressions.ccdmd.qc.ca/repertoire_es.php, último acceso 18/6/2020.
12. Morvay, K.: Aspectos lexicográficos y didácticos de la Paremiología y Fraseología. *Paremia* 6, 423–432 (1997).
13. Penadés Martínez, I.: *La enseñanza de las unidades fraseológicas*. Arco/Libros, Madrid (1999).
14. Penadés Martínez, I.: *Para un diccionario de locuciones. De la lingüística teórica a la fraseología práctica*. Universidad de Alcalá, Alcalá de Henares (2015).
15. Penadés Martínez, I.: *Diccionario de locuciones idiomáticas del español actual (DiLEA)* (2019), <http://www.diccionariodilea.es/diccionario>, último acceso 2/5/2022.
16. Ruiz Gurillo, L.: *Aspectos de fraseología teórica española*. Universitat de València, Valencia (1997).
17. Sancho Cremades, P.: *Introducció a la fraseologia, aplicació al valencià col·loquial*. Denes, Paiporta (Valencia) (1999).
18. Velázquez Puerto, K.: *La enseñanza-aprendizaje de fraseología en ELE*. Arco Libros – La Muralla, Madrid (2018).
19. Vranic, G.: *Hablar por los codos: Frases para un español cotidiano.* Edelsa, Madrid (2004).
20. Zuluaga, A.: *Introducción al estudio de las expresiones fijas*. Verlag Peter D. Lang, Frankfurt a.M./Bern (1980).

Una nota acerca de las dificultades de comprensión de unidades fraseológicas en los libros por parte de niños con Trastorno del Espectro Autista (TEA)

Valeria Kiselova Savrasova

Universidad de Málaga. Spain
vkiselova@uma.es, valeriakiselovasavrasova@gmail.com

Abstract. La lectura de los libros infantiles en los niños con Trastorno del Espectro Autista (TEA) presenta dificultades. Una de ellas y la más compleja es la incomprensión o la equivoca interpretación de las unidades fraseológicas de los textos. En esta nota se examinan algunos libros destinados a lectores de edades de 6 a 10 años, libros que leen también los escolares con necesidades educativas especiales que no tienen adaptación curricular. Aunque existe una gran variedad de libros con pictogramas para niños prelectores y primeros lectores, la inexistencia de textos adaptados para la mejora de la comprensión de estas personas con TEA de edad posterior abre una brecha en su desarrollo y en sus estudios.

Keywords: TEA · Trastorno del Espectro Autista · fraseología · necesidades educativas especiales · literatura infantil.

1 Introducción

La lectura de libros infantiles causa mucha dificultad a los niños con trastorno del espectro autista. Es escasa la presencia de libros adaptados para los alumnos de Primaria con TEA mientras si la hay para niños de edad prelectora. Una de las mayores dificultades para estos niños es la dificultad de la comprensión del lenguaje fraseológico y nuestro propósito es recalcar la necesidad de la creación de libros específicos para que la inclusión sea más fructífera.

En el currículo de la ESO se fomenta el desarrollo del hábito de lectura. Es uno de los objetivos de la educación primaria. "Conocer y utilizar de manera apropiada la lengua castellana y, si la hubiere, la lengua cooficial de la Comunidad Autónoma y desarrollar hábitos de lectura" (Real Decreto 126/2014, p. 7). Los alumnos leen extractos de obras literarias incluídas en sus libros de texto así como de otros libros, siendo estas lecturas algunas de ellas obligatorias y otras voluntarias. Para conseguir el desarrollo en las destrezas básicas del uso de la lengua en los alumnos, tales como entender, hablar, leer y escribir, se utiliza la lectura de estos textos haciendo hincapié en la comprensión y la reflexión sobre ellos por parte del alumnado. Se suelen alternar las lecturas obligatorias, las obras literarias importantes con las obras literarias que dependen de los gustos

personales y de la madurez cognitiva de los alumnos. Son importantes estos tres tipos diferentes de lectura (Real Decreto 126/2014, p. 26): lectura para buscar información, lectura para aprender la lengua materna y lectura por placer.

2 Libros adaptados para prelectores

Para los niños con trastorno del espectro autista, trastorno del lenguaje y otros diagnósticos que influyen en el desarrollo del lenguaje los recursos más adecuados son los cuentos adaptados con pictogramas, porque la comprensión con imágenes mejora notablemente la comprensión del texto en personas con TEA (Tamarit, De Dios, Domínguez y Escribano, 1990; Hodgdon, 1995; Grandin, 1995).

La adaptación de un cuento en un libro con pictogramas puede hacerse en el centro de educación primaria. Se usan los recursos de la biblioteca, que además de los libros, tiene otros elementos necesarios como el ordenador, el escáner, la fotocopiadora y la plastificadora para poder adaptar esas lecturas.

No obstante, los especialistas también disponen de cuentos ya adaptados en portales gratuitos en España, tales como: Pictocuentos, Aprendices visuales, diversos blogs de logopedas y profesionales de la enseñanza. Además, varias editoriales en España publican cuentos adaptados: Kalandraka (colección Makakiños), GEU, CEPE y Sallybooks cuando en la mayoría de los países europeos no hay editoriales que se dediquen a estas publicaciones especializadas.

Los lectores con TEA suelen empezar con lecturas de "historias sociales", sin embargo, la lectura adaptada de historias de fantasía les ayuda a mejorar el léxico. Consiguen ampliar su cultura general de manera adaptada y lúdica. Estos cuentos son un buen método para poder traducir en pictogramas todas las palabras que tienen el significado léxico. No se suele utilizar el lenguaje metafórico en estos libros, porque los niños que tienen el retraso en el desarrollo del lenguaje, como, por ejemplo, los niños con TEA, no comprenden las unidades fraseológicas con la misma facilidad que los niños de su edad (Nippold, 1991; Marshall & Kasirer, 2011; Vogindroukas y Zikopoulou, 2011; Yi, Hwang, Ah Lim, 2013; Walenski, Love, 2017; Tzuriel y Groman, 2017).

En los últimos años, algunas editoriales se preocupan por introducir en su catálogo libros que sirven para explicar el significado figurativo de las unidades fraseológicas. Las unidades fraseológicas

> son unidades léxicas formadas por más de dos palabras gráficas en su límite inferior, cuyo límite superior se sitúa en el nivel de la oración compuesta. Dichas unidades se caracterizan por su alta frecuencia de uso, y de coaparición de sus elementos integrantes; por su institucionalización, entendida en términos de fijación y especialización semántica; por su idiomaticidad y variación potenciales; así como por el grado en el cual se dan todos estos aspectos en los distintos tipos (Corpas Pastor, 1997, p. 20).

Así, la editorial GEU (Grupo Editorial Universitario) tiene publicados tres libros de la serie "Frases con doble sentido para niños con TGD y otras dificultades para interpretar el lenguaje", realizada por Inés Rubio Vega y Mónica Rubio Vega. En cada libro se explica el uso de 5 unidades fraseológicas con explicaciones verbales e imágenes. Se ha optado por locuciones verbales. En el primer libro se trabajan: *ponerse los ojos como platos, dar calabazas, ponerse como un tomate, estar como sardinas en lata, estar hecho un fideo*; en el segundo: *estoy como un flan, está chupado, vete a freír espárragos, me importa un pepino, estoy hecha una sopa*; y en el tercero: *está todo patas arriba, me he puesto las botas, voy volando, está delante de tus narices, he hecho la vista gorda.* Si en el primer libro se utilizan los verbos en infinitivo, en los otros dos los verbos están conjugados mayoritariamente en primera persona del singular en el tiempo presente. Los libros tienen pegatinas para que el trabajo sea más atractivo y ejercicios para comprobar la comprensión de las unidades fraseológicas.

Superhéroes literales de María Vega en colaboración con la fundación Quinta, publicado por la editorial Adarve en 2018 es un recopilatorio de unidades fraseológicas útiles para niños. Cada unidad fraseológica (hay más de 30) está acompañada de dos dibujos, en el primero está representado visualmente el significado literal y en el segundo, el significado figurado. La autora, madre de un niño con TEA, en el prólogo resalta la importancia de un apoyo visual a la hora de explicar las unidades fraseológicas a un niño con TEA:

> Viendo la imagen explicativa y a la vez divertida de expresiones raras, difíciles de entender por su doble sentido, nuestro hijo con T.E.A. ha sido capaz de captar su significado e integrarlas en su vocabulario, con lo que ha mejorado mucho su adaptación a las conversaciones incluso ha disminuido su frustración al comunicarse (Vega, 2018, p.11).

3 Libros para primeros lectores

En los libros para primeros lectores aparecen una mayor cantidad de unidades fraseológicas y no permiten una lectura autónoma y fluida a estos niños con necesidades educativas especiales. La comprensión se ve afectada debido a las siguientes características: la transparencia (Cacciari y Levorato, 1998), la familiaridad (Nippold y Rudzinski, 1993) y el contexto (Levorato y Cacciari, 1995). A las personas con TEA les resulta dificultoso interpretar el significado de la misma oración en contextos diferentes, porque "presentan problemas para asignar significados figurados, deshacer ambigüedades o para aprender acepciones múltiples para los mismos términos" (Pérez Juliá y Martos Pérez, 2002, p 74).

Para poner de manifiesto esas dificultades podemos incluir aquellos casos que permiten una doble interpretación y tienen el significado denotativo figurativo.

Veamos algunos ejemplos de locuciones que aparecen en algunos libros infantiles destinados a lectores de 7 a 10 años. En *Los despistes de Spider* de Alan Durant, lectura recomendada a partir de 8 años, entre otras aparecen numerosas

locuciones relacionadas con las emociones: *fruncir el ceño, encogerse de hombros, tener estómago revuelto, dar saltos de alegría, bajar la cabeza, quedarse mudo, estar loco de contento, helar la sangre a alguien, con aire preocupado, estar paralizado, estar con la boca abierta, tener un nudo en la garganta, ardérsele los ojos a alguien, estar perplejo*, etc.

En el texto se presentan muchas locuciones, y para un niño que no entiende su significado figurativo, la lectura es de mucha dificultad. Algunas veces en un solo párrafo puede aparecer más de una locución. Por ejemplo: "Tenía mucho que decir, pero se había quedado en blanco. Permaneció con la boca abierta y paralizado. [?] Tenía un nudo en la garganta y le ardían los ojos" (Durant, 2008, pps.57-58). Si un niño desconoce estas expresiones, no puede seguir la historia porque no comprende la trama y suele interpretar las unidades fraseológicas de manera literal.

Al realizar la lectura de la serie popular sobre Pupi, escrita por María Menéndez-Ponte, recomendada para primeros lectores, también encontramos una gran cantidad de unidades fraseológicas que tienen doble interpretación. En Pupi y los piratas encontramos entre otras las siguientes unidades fraseológicas: *traer mala suerte, ser una pesada, no dar su brazo a torcer, cabeza de chorlito, flaco como un palo de escoba, paralizado por el pánico, contener el aliento, inflarse de satisfacción, ponerse hueco, en picado, hacer caso, llevar a cabo, hacer la pelota.* Algunas de ellas están relacionadas con las descripciones de los personajes del libro, no llaman atención por su relación con las emociones, hay una gran variedad de las unidades fraseológicas con intenciones comunicativas diversas.

En *La lista de cumpleaños* de Anna Manso, lectura recomendada de 7 a 8 años, también encontramos una gran variedad de unidades fraseológicas. Procedemos a mencionar algunas que consideramos que pueden resultar difíciles a niños con TEA por su posible doble interpretación: *escaparse la risa, la cabeza está que echa humo, nadie dice ni mu, hacer mucha gracia, hincar el diente, entrar el tembleque, sudar como pollo, gastar una broma, freír espárragos, ahora que caigo.*

En *El caso del cementerio embrujado* de la serie Buscapistas de la autora Teresa Blanch entre otras encontramos las siguientes UF: *hacer cola, pasar lista, tomar asiento, echar un vistazo, estar claro, ser un caso, asomar las narices, dar importancia, con los ojos como platos, de par en par, boca de lobo, pegar las narices, dar un susto de muerte, poner los pelos de punta, por si las moscas, con el miedo en el cuerpo, ponerse manos a la obra, estar muerto del miedo, oler mal algo, salir pitando, en menos que canta un gallo, estar rojo como un pimiento, darse cuenta de algo, hacer caso.*

Los escritores de literatura infantil utilizan las unidades fraseológicas para aumentar la expresividad, darle un toque humorístico. Algunas unidades fraseológicas en sus implicaturas incluyen un contenido humorístico (*sudar como un pollo, nadie dice ni mu, pegar las narices*), otras tienen esta interpretación gracias al contexto en el que aparecen.

Los escritores a la hora de crear su texto eligen utilizar o no las unidades fraseológicas. Gonzalo Moure, autor de literatura infantil y juvenil, en el manual

de la escritura *Por qué llora la maestra. Carta larga para ti, que quieres escribir* aconseja evitar el uso de las unidades fraseológicas en cualquier texto literario:

> En ocasiones, cuando releo lo que he escrito, parece que lo he hecho comprando un surtido de expresiones en todo a cien. Son las frases hechas. Cualquiera: "no dio su brazo a torcer", o "estaba muerto de cansancio". Hay infinidad, miles de ellas y, cuando las usamos, nos parece que son efectivas; casi nos parece que las acabamos de inventar, cuando en realidad ya las hemos leído muchas veces. Pero son como bastones, como muletas de nuestro texto. Es que es lo que son. Hacen que la lectura de lo que has escrito suene banal, vulgar. Lee tu texto y, si hay alguna frase así, cámbiala. A veces con sencillez: "No cedió", o "estaba muy cansado", sin ir más lejos. Pero otras veces, ojalá que muchas, encontrarás maneras nuevas de expresarlo, y el texto ganará si tu forma de decirlo es original, siempre que no sea rebuscada. Por ejemplo, imaginemos que has escrito "me sentía entre la espada y la pared". Detienes el tiempo, buscas y, cuando lo encuentras, lo dices a tu manera. Un hallazgo, tu hallazgo. [...] Eso es lo que medirá tu verdadero talento (Moure, 2021, pps. 56-57).

No podemos discutir acerca de las soluciones que tomen los escritores a la hora de crear sus obras, sin embargo, podemos imaginar qué dificultades en su comprensión pueden tener los niños con TEA si leen textos con una gran cantidad de recursos en los que se emplea el lenguaje figurativo. De los libros escogimos mayoritariamente las locuciones nominales y locuciones verbales por su posible doble interpretación y la dificultad para niños con TEA debido al significado denotativo figurativo que contienen.

4 Serie de lectura fácil de la editorial SM

La editorial SM tiene una colección de Lectura Fácil en la cual hace las siguientes adaptaciones: cambios en la tipografía (letras más grandes, márgenes más amplios), ilustraciones más sencillas y el léxico más sencillo. La colección la constituyen 14 libros divididos por edades:

- 6-7 años: *Siete vidas* de Ana Guerrero y Andrés Guerrero, *El fantasma de la casa de al lado* de Iñaki R. Díaz, *Cómo consolar a una ardilla* de Begoña Oro Pradera,
- 7-8 años: *Pirata Plin, Pirata Plan* de Paloma Sánchez Ibarzábal, *La lista de cumpleaños* de Anna Manso Munné, *Los sueños de Aurelia* de Eduard Márquez Tañá,
- 8-10 años: *Siete reporteros y un periódico* de Pilar Lozano Carbayo, *Mi nombre es Skywalker* de Agustín Fernández Paz, *El club de los raros de Jordi Sierra i Fabra*,

- 10-12 años: *El robo del siglo* de Ana Alonso, *El país de los relojes* de Ana Alonso, Abdel de Enrique Páez,

Por un lado, la existencia de estos libros permite al profesorado tener un recurso excelente para que todos los niños lean las mismas obras. Sin embargo, uno de los rasgos destacados de esta serie es la supresión del lenguaje figurativo y metafórico (según la propia editorial), lo que le permite al niño tener la lectura fácil e independiente por un lado. Pero por el otro le perjudica en el aprendizaje de las metáforas y las unidades fraseológicas, porque se omiten. Veamos un extracto en la edición original: "[..] empezó a ponerse verde, azul, violeta, más bloqueado que un camello en el Polo Norte" (Sierra I Fabra, 2015, p.7.) y su adaptación a lectura fácil: "Su cara se puso de muchos colores: verde, azul, violeta?" (Sierra i Fabra, 2017, p. 12).

Sería un recurso más útil si se añadieran páginas con las explicaciones de las mismas, favoreciendo el aprendizaje de los niños que tienen problemas en su comprensión.

5 Conclusión

Existe un gran número de lecturas adaptadas para niños prelectores con TEA, pero no para niños con TEA en la educación primaria. Es necesaria la creación de esos libros adaptados en esa etapa, donde esté explicado el significado figurativo de las unidades fraseológicas de manera debida para que sean más fáciles de comprender por los niños con TEA.

El aprendizaje de las unidades fraseológicas acompañado del contexto en las obras literarias podría convertirse en un recurso excelente para estos niños con discapacidad.

Bibliografía de libros infantiles.

Blanch, T. (2013). *El caso del cementerio embrujado. Los buscapistas -4.* Editori digital titivillus.

Durant A. (2008). *Los despistes de Spider.* Alfaguara.

Manso Munné, A. (2010). *La lista de cumpleaños.* Ediciones SM.

Menéndez-Ponte, M. (2013 [2009]). *El cumpleaños de Pupi.* Ediciones SM.

Moure, G. (2021). *Por qué llora la maestra. Carta larga para ti, que quieres escribir.* Kalandraka.

Sierra I Fabra, J. (2015). *El club de los raros.* Ediciones SM.

Sierra I Fabra, J. (2017). *El club de los raros. Adaptación a lectura fácil.* Ediciones SM.

Rubio, I. (2014). *Frases con doble sentido para niños con TGD y otras dificultades para interpretar el lenguaje. Cuaderno 1.*GEU.

Vega Torres, M. (2018). *Superhéroes literales.* Adarve.

References

1. Cacciari, C. y Levorato, M. C. (1998). The effect of semantic analysability of idioms inmetalinguistic tasks. *Metaphor and Symbol*, 13, 159-77.
2. Corpas Pastor, G., y Ezquerra, M. A. (1997). *Manual de fraseología española.* Gredos.
3. Hodgdon, L. Q. (1995). Solving social-behavioral problems through the use of visually supported communication. In Quill, K. A. (Ed.), *Teaching children with autism: Strategies to enhance communication and socialization (pp. 265-285).*Delmar.
4. Grandin, T. (1995). "How people with autism think". In E. Schopler & G.B. Mesibov (Eds.). *Learning and cognition in autism.* Plenum Press, 137-156.
5. Real Decreto 126/2014, de 28 de febrero, por el que se establece el currículo básico de la Educación Primaria. Boletín Oficial del Estado. 1 de marzo de 2014, núm. 52, pp. 3-45.
6. Levorato, M.C., y Cacciari, C. (1995). The effects of different tasks on the comprehension andproduction of idioms in children. *Journal of Experimental Child Psychology.* 60 (2), 261-283.
7. Mashal, N., & Kasirer, A. (2012). Principal component analysis study of visual and verbal metaphoric comprehension in children with autism and learning disabilities. *Research in developmental disabilities,* 33(1), 274-282.
8. Nippold, M. A. (1991). Evaluating and enhancing idiom comprehension in language-disordered students. *Language, Speech, and Hearing Services in Schools,* 22(3), 100-106.
9. Nippold, M.A. y Rudzinski, M. (1993): Familiarity and transparency in idiom explanation: Adevelopmental study of children and adolescents. *Journal of Speech and Hearing Research,* 36,728-37.
10. Tamarit, J. D., De Dios, J., Domínguez, S., y Escribano, L. (1990). Proyecto de Estructuración Ambiental en el aula de Niños Autistas. PEANA.
11. Vogindroukas, Ioannis, & Zikopoulou, Olga. (2011). Idiom understanding in people with Asperger syndrome/high functioning autism. *Revista da Sociedade Brasileira de Fonoaudiologia,* 16(4), 390-395.
12. Tzuriel, D., & Groman, T. (2017). Dynamic assessment of figurative language of children in the autistic spectrum: The relation to some cognitive and language aspects. *Journal of Cognitive Education and Psychology,* 16(1), 38-63.
13. Yi, D., Hwang, M., Lim, J. A., Yi, D., Hwang, M., & Lim, J. A. (2013). Proverb comprehension in children with Asperger's disorder: The role of transparency. *Communication Sciences & Disorders,* 18(3), 288-296.
14. Walenski, M., & Love, T. (2017). The real-time comprehension of idioms by typical children, children with specific language impairment and children with autism. *Journal of speech pathology & therapy, 3(1).*

Estructuración de locuciones verbales por campos semánticos y su aplicación didáctica

Tatiana Denisenko

Universidad de Cádiz, P.º de Carlos III, 28, 11003 Cádiz, España
denisenkots@gmail.com

Abstract. Si bien en el campo de la fraseología teórica aparecen más trabajos dedicados a la delimitación de los conceptos entre las unidades fraseológicas, al análisis de las relaciones paradigmáticas y sintagmáticas, su clasificación, así como las obras lexicográficas de distinta índole, desafortunadamente en el ámbito de la lingüística aplicada y la enseñanza de ELE se observan ciertas deficiencias respecto a la aplicabilidad de dichas aportaciones en la enseñanza de las locuciones. El objetivo de este trabajo es presentar una propuesta de estructuración de las locuciones del español por campos semánticos con el fin de plantear un modelo práctico que sirva de apoyo tanto para los profesores de ELE a la hora de llevar al aula la enseñanza de las unidades fraseológicas, como para los alumnos que pueden recurrir a las listas de locuciones para consultarlas. En concreto, nos hemos centrado en las locuciones verbales que expresan gustos, deseos y sentimientos, las cuales han sido extraídas del *Diccionario de locuciones verbales para la enseñanza del español* de I. Penadés Martínez y el *Plan Curricular del Instituto Cervantes* para niveles de referencia para el español como lengua extranjera. Asimismo, hemos examinado las relaciones paradigmáticas de hiponimia-hiperonimia, sinonimia y antonimia que guardan las locuciones entre sí, hecho que facilitaría la incorporación de las locuciones en el proceso de enseñanza-aprendizaje de ELE.

Keywords: Locuciones verbales, Campo semántico, Relaciones paradigmáticas, Enseñanza ELE.

1 Introducción

La teoría del campo es considerada como la gran revolución de la semántica moderna, ya que la lengua se presenta como sistema a la hora de investigar el vocabulario. El vocabulario se concibe como una totalidad semántica articulada en cuyo interior cada unidad léxica, y en nuestro trabajo, las unidades fraseológicas, están bajo la dependencia de las otras. De esta manera, las unidades léxicas pueden formar campos léxicos y

campos semánticos[1] que abarcan y expresan una visión del mundo y que se organizan por significado en subsistemas de diferente extensión y complejidad.

Si bien existe un número considerable de estudios donde se aborda el tema de la enseñanza de las locuciones verbales en español como lengua extranjera (ELE), la gran mayoría de las investigaciones se concentran en la búsqueda de equivalentes en el idioma natal del estudiante o, en el caso de la traducción, en el idioma de destino. Por ello, el presente trabajo tiene como objetivo presentar una propuesta de estructuración de las locuciones del español por campos semánticos con el fin de crear una herramienta de consulta tanto para docentes, como aprendientes de ELE. Con dicha herramienta se pretende facilitar la búsqueda de locuciones por su significado, su incorporación en el proceso de enseñanza-aprendizaje empezando desde los niveles iniciales y, a partir de ahí, formación de series mnemotécnicas, que serían de gran utilidad para la memorización y la retención en la memoria de las unidades fraseológicas que presentan un alto grado de fijación e idiomaticidad.

Del mismo modo, se pretende analizar en ejemplos concretos las relaciones paradigmáticas de hiperonimia-hiponimia, sinonimia y antonimia entre las locuciones desde el punto de vista didáctico teniendo en cuenta su tradicional tratamiento en los manuales de ELE y actividades diseñadas para enseñar unidades léxicas y fraseológicas.

Aplicando la definición del campo léxico o campo semántico de Coseriu (1981: 146 y 210) a nuestro trabajo, consideramos que un campo semántico representa una estructura paradigmática constituida por unidades fraseológicas, que resulta de la distribución por el contenido semántico y de la oposición por medio de rasgos distintivos. En cada campo semántico se puede observar que las locuciones agrupadas bajo el mismo concepto, o el significado común, se relacionan por un rasgo distintivo común, es decir, en los enunciados se comportan de manera análoga desde el punto de vista léxico-gramatical.

En ese sentido, la organización de las locuciones en campos semánticos, aplicada en el presente trabajo, se produciría según el contenido semántico de estas e incluiría relaciones distintivas paradigmáticas: sinonimia, antonimia e hiperonimia-hiponimia.

2 Metodología

Para elaborar una propuesta de estructuración de locuciones por campos semánticos, en nuestro trabajo nos hemos centrado en las locuciones verbales tomadas del *Diccionario de locuciones verbales para la enseñanza del español* de I. Penadés Martínez (2002) y el *Plan Curricular del Instituto Cervantes para niveles de referencia para el español* (2006). Como punto de partida, fue establecido un macrocampo denominado "Expresar gustos, deseos y sentimientos", a partir del cual se han ido formando campos semánticos, o microcampos, supeditados al macrocampo elegido, cada uno de los cuales engloba locuciones con ciertos rasgos semánticos distintivos. Las fuentes indicadas han

[1] La diferencia entre ambos campos reside en que campo léxico se refiere a las unidades léxicas simples, en tanto que campo semántico comprende unidades léxicas simples, unidades léxicas complejas y unidades fraseológicas.

sido elegidas principalmente por la razón de ser ya orientadas a la enseñanza del español como lengua extranjera y de presentar obras de referencia para este mismo objetivo.

En total, hemos delimitado 27 campos semánticos, adaptando las nociones del inventario *Funciones* del *Plan Curricular del Instituto Cervantes*. Estas funciones se definen como "*funciones de la lengua*, entendidas como el tipo de cosas que la gente puede hacer mediante el uso de la lengua" y que en otro sentido pueden ser interpretadas como estrategias pragmáticas de la lengua (Instituto Cervantes, 2006). Asimismo, hemos agrupado dichos campos semánticos, o microcampos, en cuatro bloques, lo que se puede observar en la siguiente Tabla 1.

Tabla 1. Campos semánticos agrupados en bloques.

Macrocampo: Expresar gustos, deseos y sentimientos		
Bloque I	Expresar gustos e intereses Expresar aversión	Expresar indiferencia o ausencia de preferencia
Bloque II	Expresar deseos	Expresar planes e intenciones
Bloque III (sentimientos y emociones)	Expresar alegría y satisfacción Expresar afecto Expresar empatía Expresar enfado e indignación Expresar hartazgo Expresar tristeza y aflicción Expresar nerviosismo Expresar miedo, ansiedad y preocupación Expresar esperanza Expresar arrepentimiento Expresar placer y diversión	Expresar aburrimiento Expresar alivio Expresar decepción Expresar resignación Expresar sorpresa y extrañeza Expresar admiración y orgullo Expresar vergüenza Expresar sufrimiento Expresar molestia, disgusto y desagrado Expresar fuertes emociones y sentimientos
Bloque IV	Expresar sensaciones físicas	

Nuestro corpus está constituido por un conjunto de 443 locuciones a partir del significado de las cuales hemos ido delimitando campos semánticos. En cuanto a la metodología de vaciado de locuciones se ha procedido un exhaustivo análisis, la identificación y extracción de manera manual de las locuciones, basándose en su significado vinculado al significado global del macrocampo, es decir, aquellas locuciones que expresan algún tipo de sentimiento, emoción, gustos, intereses, planes, intenciones o sensaciones físicas. Por medio del rastreo sistemático de locuciones afines al criterio anterior hemos ido examinando y validando la pertenencia de cada locución a un campo específico.

Siguiendo a Penadés (2003: 118), que, además, subraya que en la práctica lexicográfica es habitual señalar la información relativa a la combinatoria sintagmática, los elementos de la combinatoria se representan mediante las siguientes formas: *alguien* (para persona) y *algo* (para cosa). Se indican, asimismo, las preposiciones que introducen los complementos.

Para ilustrar cómo están formados los microcampos, a continuación podemos ver la combinatoria sintagmática y el repertorio abreviado de locuciones del campo semántico "**Expresar sensaciones físicas**":

Tabla 2. Combinatoria sintagmática dentro del campo y el repertorio de locuciones.

Campo semántico "Expresar sensaciones físicas"	
alguien	caerse a pedazos estar en un grito no poder con su alma sudar la gota gorda ver las estrellas
a alguien	dar algo irse la cabeza
algo/alguien, a alguien	hacer papilla hacer trizas

Por otro lado, han sido incluidas en un mismo campo las locuciones que puedan diferenciarse por si expresan en distintos contextos una acción durativa, realizaciones, aquella que está a punto de producirse, estados o logros (Real Academia Española, 2010: 431-436) como, por ejemplo, en el campo semántico "**Expresar indiferencia o ausencia de preferencia**". Fíjense: *dar igual, dar lo mismo, ser igual* y *ser lo mismo* se usan normalmente en los enunciados que aluden a una situación no delimitada en el tiempo; en cambio, *hacerse el loco, hacerse el longui(s), hacerse el tonto, entrar por un oído y salir por el otro, hacer caso omiso* o *pasarse por el arco del triunfo* podrían expresar algo que se efectúa, o realizaciones.

Se ha procedido, además, a incorporar en un mismo campo semántico las locuciones que se distinguen en el registro de uso. De este modo, en el campo "**Expresar hartazgo**" se puede observar las locuciones *no poder más* y *tocar fondo* junto a *estar hasta el gorro* y *estar hasta la coronilla*, que, sin duda, pertenecen a distintos registros en la lengua.

3 Resultados y discusión

3.1 Relaciones paradigmáticas entre las locuciones

Antes de proceder, es necesario señalar que en las teorías semánticas de la lingüística moderna y los estudios que giran en torno a la fraseología, los autores definen las relaciones paradigmáticas entre unidades fraseológicas desde distintas perspectivas. Para algunos científicos, desde una perspectiva que se refiera a las formas de contenido, las relaciones léxico-semánticas están representadas por sinonimia, parasinonimia, hiperonimia, cohiponimia, hiponimia y diversos tipos de antonimia (Casas Gómez, 2015: 6). Otros autores (Ruíz Gurillo, 2001: 59) incluyen también las relaciones de polisemia y homonimia como relaciones semánticas entre unidades léxicas complejas. Existe, además, el enfoque según el cual la sinonimia no podría ser representada como relación

semántica (Penadés, 2012). En nuestro trabajo no nos detendremos en el análisis exhaustivo de distintos enfoques respecto a las relaciones léxico-semánticas de las locuciones, sino mostraremos tales relaciones en contexto teniendo en cuenta su aplicabilidad a la enseñanza de la lengua.

3.2 Hiponimia-hiperonimia

Se podría decir que todo campo semántico, o microcampo, representa una estructuración de locuciones hipónimas respecto a su hiperónimo que sería la misma denominación del campo. Así, por ejemplo, podrían relacionarse las locuciones que por su significado amplio se refieren al campo semántico **"Expresar aversión"**. No obstante, en el presente trabajo nos fijamos únicamente en las relaciones paradigmáticas entre las locuciones en sí.

Así, la locución *hacer de menos* que significa "menospreciar {a una persona}" funcionaría como hiperónimo para las locuciones hipónimas *no poder ver ni en pintura, no poder ver ni pintado* "sentir odio o aversión hacia una persona o cosa", *hundir en la miseria* "hacer que una persona se sienta humillada", *tener en poco* "considerar {a una persona o una cosa} poco digna de aprecio" y *tener entre ceja y ceja* "sentir antipatía hacia una persona". Cada una de estas locuciones presentan rasgos de los que carece el hiperónimo: *no poder ver ni en pintura* presenta el rasgo de un sentimiento negativo fuerte que es "odio o aversión"; *hundir en la miseria* implica acciones que llevan a la sensación de humillación por parte de otra persona; *tener en poco* tiene el rasgo de consideración de poco aprecio hacia alguien; *tener entre ceja y ceja* expresa la antipatía.

Fíjense en el significado y el uso de estas locuciones en ejemplos concretos:

- Locución hiperónima: *hacer de menos*
  Significado: "menospreciar"
  Combinatoria sintagmática: [*alguien, a alguien*]
  Contexto: «*Pero también pienso que mi desagrado podría deberse a que mi mujer nunca la trató bien, claro que no, siempre la hizo de menos, le fingía una adhesión que no era sincera...*» (Luis Hernáez, *Destidós*, 2002, Paraguay, *CORPES*).

  - Locución hipónima: *no poder ver ni en pintura*
    Significado: "sentir odio o aversión hacia una persona o cosa"
    Combinatoria sintagmática: [*alguien, algo/a alguien*]
    Contexto: «*…Bea bebía anís en copa alta, gastaba medias de seda de La Perla Gris y se maquillaba como las vampiresas cinematográficas que perturbaban el sueño de mi amigo Fermín. Yo no podía verla ni en pintura, y ella correspondía a mi franca hostilidad con lánguidas miradas de desdén e indiferencia.*» (Carlos Ruiz Zafón, *La sombra del viento*, 2001, España, *CORPES*).

  - Locución hipónima: *hundir en la miseria*
    Significado: "hacer que una persona se sienta humillada"
    Combinatoria sintagmática: [*alguien, a alguien*]

Contexto: *«…proyectan una imagen al exterior que no se corresponde con la realidad. Y esta imagen superlativa conseguida a base de "efectos especiales" es la que condiciona, baja la moral y, confesémoslo, en ocasiones nos hunde en la miseria más innoble.»* (Mónica Ceño Elie-Joseph, *Desnudas. Aprende a quererte tal como eres*, 2007, España, *CORPES*).

3.3 Sinonimia

En el presente trabajo hemos optado por el planteamiento más amplio de la sinonimia entre las locuciones y es la que designa diversos tipos de significados semejantes o de la relación semántica, orientada a la didáctica –puesto que en la enseñanza activamente se emplea la sinonimia como recurso para la introducción, descodificación y memorización de una nueva unidad léxica–, esto es, la sinonimia parcial y absoluta (García-Page, 2008), o, en otras palabras, de distribución libre o complementaria (Penadés, 2012), variantes sociolingüísticas, así como relaciones de sinonimia entre distintas variantes de las locuciones (Corpas, 1996). Asimismo, a través de contextos de uso, hemos podido corroborar la afirmación que la agrupación de locuciones por campos semánticos favorece la búsqueda y el establecimiento de las relaciones sinonímicas entre las locuciones, por lo que esto sería de gran utilidad para los aprendientes, docentes e interesados en el español como lengua extranjera.

De nuestro análisis de las locuciones, hemos seleccionado, como ejemplo, aquellas que se agrupan bajo la denominación del microcampo **"Expresar afecto"** y comparten de algún modo el significado que se refiere a "querer a alguien":

- *beber los vientos* y *morirse por los/sus pedazos* "estar muy enamorado de una persona";
- *no ver más que por los/sus ojos* "querer mucho a una persona, estando pendiente de ella";
- *querer bien* "sentir afecto hacia una persona, estando pendiente de ella / amarla".

Todas estas locuciones presentan la combinatoria con el sujeto *alguien [realiza lo expresado por la locución]*, pero con los complementos la situación es distinta: *alguien bebe los vientos por alguien, alguien se muere por sus/los pedazos de alguien, alguien no ve más que por sus ojos / alguien no ve más que por los ojos de alguien, alguien quiere bien a alguien*. Las cuatro locuciones y sus variantes se considerarían sinónimas en distribución complementaria, ya que no presentan la equivalencia semántica absoluta, sino una semejanza en significado, y, además, sería necesaria la adaptación de los enunciados en caso de la sustitución.

3.4 Antonimia

Según el análisis llevado a cabo en nuestra investigación, es posible establecer relaciones de antonimia en numerosos casos. Así pues, la locución analizada anteriormente en la parte dedicada a la sinonimia, que es *querer bien* "sentir afecto hacia una persona" perteneciente al campo semántico **"Expresar afecto"**, entra en relación antonímica con

las locuciones del campo semántico **"Expresar aversión"**, que son *no poder ver ni en pintura* y *no poder ver ni pintado* que significan "sentir odio o aversión hacia una persona o una cosa". En este sentido, la oposición se basa en el significado de toda su expresión fija. Otros ejemplos de relaciones de antonimia se observan en los siguientes ejemplos en los que, además, puede estar presente la oposición entre series sinonímicas y antonímicas:

- Campo semántico **"Expresar gustos e intereses"**
  - *llamar la atención* ("despertar interés, curiosidad o sorpresa") / *no decir nada – no decir gran cosa* ("no despertar interés, curiosidad o sorpresa");

- Campos semánticos **"Expresar planes e intenciones"** y **"Expresar indiferencia o ausencia de preferencia"**
  - *meterse en la cabeza – meterse en la mollera – meterse entre ceja y ceja – ponerse entre ceja – tener entre ceja y ceja* ("obstinarse en una cosa") / *quitarse de la cabeza* ("dejar de obstinarse en una cosa");

- Campos semánticos **"Expresar indiferencia o ausencia de preferencia"** y **"Expresar fuertes emociones o sentimientos"**
  - *no dar ni frío ni calor* ("no causar ninguna impresión {a una persona}") / *hacer mella – llegar al alma* ("causar una fuerte impresión {a una persona o una cosa}").

4 Conclusiones

Una vez hemos hecho la aportación a las posibles relaciones paradigmáticas entre las locuciones presentadas a través de su agrupación en campos semánticos y vistas dentro del contexto y el uso, podemos extraer una serie de conclusiones. Así pues, todo lo expuesto demuestra que es posible la estructuración semántica de las locuciones según relaciones de hiponimia-hiperonimia, sinonimia y antonimia. Además, cada relación paradigmática se manifiesta por distintos tipos, lo que ofrece múltiples posibilidades para los docentes a la hora de presentar y explicar el uso de las locuciones a los alumnos de ELE. Por ello, partiendo de la importancia para el aprendizaje y asimilación de los significados de cada locución y el desarrollo de la competencia léxico-semántica, resulta esencial recurrir a ejemplos reales bien contextualizados.

Dicho esto, la agrupación de locuciones por campos semánticos facilitaría la elaboración de distintas actividades, no solo las que traten con las relaciones paradigmáticas entre las locuciones, sino también aquellas que permitan trabajar su significado dentro de un contexto, así como actividades que comprendan el análisis de la estructura de locuciones, considerando también las actividades de mediación y traducción, el análisis etimológico y registros de uso de las locuciones.

El desarrollo futuro de nuestra investigación residirá en profundizar respecto a las relaciones semánticas entre las locuciones para poder transferir los resultados teóricos

a la didáctica de la fraseología. Al mismo tiempo, sería substancial realizar la estructuración por campos semánticos de otras clases de locuciones tales como las nominales, adjetivas, pronominales y adverbiales.

Referencias

1. Casas Gómez, M.: Propuesta para una clasificación de las relaciones en semántica. Linred: Lingüística en la Red, 13 (2015).
2. Corpas Pastor, G.: Manual de fraseología española. Gredos, Madrid (1996).
3. Coseriu, E.: Principios de semántica estructural. 2ª ed. Alonso, D. (dir.). Gredos, Madrid (1981).
4. García-Page Sánchez, M.: Introducción a la fraseología española: Estudio de las locuciones. Anthropos Editorial, Rubí (Barcelona) (2008).
5. Instituto Cervantes: Plan Curricular del Instituto Cervantes. Niveles de referencia para el español. Instituto Cervantes, Biblioteca nueva, Madrid (2006). http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/
6. Penadés Martínez, I.: Diccionario de locuciones verbales para la enseñanza del español. Editorial Arco Libros, Madrid (2002).
7. Penadés Martínez, I.: La elaboración del Diccionario de locuciones verbales para la enseñanza del español. Revista de Lexicografía 9, 97-129 (2003).
8. Penadés Martínez, I.: Gramática y semántica de las locuciones. Universidad de Alcalá. Servicio de publicaciones, Alcalá de Henares (2012).
9. Real Academia Española: Nueva Gramática de la lengua española. Manual. Asociación de academias de la lengua española. Espasa (2010).
10. Ruiz Gurillo, L.: Las locuciones en el español actual. Arco Libros, Madrid (2001).

Corpus consultado

11. REAL ACADEMIA ESPAÑOLA: Banco de datos (CORPES XXI) [en línea]. Corpus del Español del Siglo XXI (CORPES), http://www.rae.es

State, Predicatives and Idiomaticity*

Maria A. Todorova[1 [0000−0001−5866−180]]

Institute For Bulgarian Language, Bularian Academy of Sciences, Sofia, Bulgaria
maria@dcl.bas.bg

Abstract. This research represents a typological and constructional approach to the description of idiomatic predicative constructions (IPCs) in Bulgarian. The study employs a subset of 1331 idiomatic predicative constructions extracted from a large dictionary of Bulgarian multi word expressions and a corpus of Bulgarian predicative constructions extracted from the Bulgarian National Corpus. The analysis of their structure, functions and semantics is a stage of the study of the syntactic and semantic structure of the predicates in Bulgarian with a view to ontological presentation of the meanings of state.

Keywords: State · Predicatives · Constructions · Verbal Idioms

1 Introduction

The study presented in the article examines a specific group of predicates which are a subset of verbal multi word expressions (VMWEs) with specific component structure combining a predicative NP or PP and an auxiliary verb. The data observed is a subset of entries extracted from a large dictionary of compound lexical items [19] using as a criterion the typology of predicative structures in Bulgarian, described in [18]. The predicative constructions are well-known and thoroughly described phenomenon in the linguistic research of Russian [24], [6], Bulgarian ([4],[11], [9], etc.), English [20] and other languages [15], [1] mainly with a view to their syntactic functions. But with focus on idioms sharing predicative structure within the semantic domain of state and the lexical and grammatical ways of its expression its less studied. The study gives another point of view of the question of the place of predicative constructions in the continuum between the semantic integrity and syntactic structures. In the remainder of this paper, §2 presents objectives and methodology of the research, §3 - the predicative constructions in Bulgarian and the verbal MWEs as their subset, §4 describes the semantic domain of state and §5 presents the observed data, §6 is constructional and §7 is semantic analysis of the Bulgarian idiomatic predicative constructions (IPC).

* This text is a result of the project "The Ontology of the stative situations in the models of language: a contrastive analysis of Bulgarian and Russian Languages", supported by the National Science Fund (Contract No. KP 06 RUSSIA / 23) under the Bulgaria-Russia 2020 Bilateral Cooperation Program. Thank you for your support.

2 Objectives and Methodology of the Research

The main objectives of this research is the description of idiomatic constructions with predicatives as a subtype of predicative constructions. Their typology is oriented to their place in the linguistic modeling of the ontology of state situations. We are investigating the hypothesis whether predicates with composite component structure are a semantic whole to which universal semantic features apply. Collecting data with idioms sharing common specifics is not an easy task as usually they have low frequency in corpora and are ambiguous with non-idiomatic expressions. That is why we extract idiomatic constructions from a ready made dictionary of Bulgarian MWEs (cf. § 5). In order to detect MWEs with predicative constructions we apply a list of construction types of predicative constructions in Bulgarian to the morphosyntactic description framework of dictionary entries. In this way we both extract candidates for idiomatic predicatives and apply to them the possible syntagmatic and paradigmatic forms. The possible variations in terms of component's order, optional components, and possible insertions both as modifiers of components or external phrases between the components of each of the extracted idiomatic predicative constructions are needed for the recognition of idiomatic predicative constructions in a corpus with a view to study of their argument structure.

3 Predicative Constructions in Bulgarian

Predicative constructions are a type of predicate structures, known also as copular predicate constructions, are language units whose features place them on the border between morphology and syntax. They are united by specific functions and semantics and largely enter the field of lexicology. Their characteristics in Bulgarian derive from their compositional component structure - a specific sequence of: auxiliary verb, adverb/adjective/noun/noun phrase/ prepositional phrase, obligatory or optional dative or accusative pronoun clitic. Morphosyntactically they are characterised with the grammatical categories of person, number and tense and lack synthetic conjugate forms [10]. The general semantic field to which belong the predicative structures is characterized by "the presence of the object in some unchanged state, which is not the result of direct influence of someone for a certain period of time" ([24]: 549). Additionally the predicative constructions are denoting some kind of evaluation of a state ([14], [4], [13]).The detailed typology of Bulgarian predicative constructions we are using in this study is presented in [18]

3.1 Verbal MWEs with Predicative Constructions

The interest of modern research on the lexicalization of expressions in the last 10 years has focused on the description of MWES, their components and structure ([23], [5], [19] with a view to their automatic detection in corpora. The challenges of their description has been also discussed when it comes to morphologically rich

languages as Bulgarian where VMWEs are characterized by a rich set of synthetic and analytical verb forms; with a complex and flexible word order. They combine structural features such as mandatory and optional components, the ability to insert external phrases (clitics, adverbs, nouns, etc.) and defective components ([7], [16], [17]). In this study we focus on the combination of structure and the meaning of idiomatic predicative constructions investigating their place in the structure of the lexical items expressing state semantics.

4 The Semantic Domain of State

According to the theory of Van Valin and Lapola ([22]: 82–138), state predicates represent situations defined as static, non-dynamic, and time-limited. State predicates express location, condition, position, or inner experience. Van Valin ([21]: 39) groups predicates of state into two major ontological classes - locative (to be on/in '(x, y) and non-locative with the subclasses: a) state or position (broken' (x)); b) perception (see '(x, y)) and cognition (I believe '(x, y)); c) possesion (I have possession' (x, y)); d) equality (I am '(x, y)). In the description of predicativeness as a linguistic phenomenon the authors [22] consider predicative structures as a semantic whole along with verbs.

5 The Selection of Data

5.1 Excerption of Bulgarian Idioms with Predictive Constructions

We used two POS tagged linguistic resources for Bulgarian to extract potential idiomatic predicative constructions in Bulgarian - the dictionary of Bulgarian MWEs [19] and a collection of predicative constructions' examples derived from the Bulgarian National Corps [8] comprising over 10,000 sentences [18]. The data of predicative structure groups was derived from the resources using the list of predicative constructions and their formal description in Bulgarian[18].

The dictionary of Bulgarian MWEs [19] contains over 86,373 noun and verb compound lexical units, derived from various sources: printed and electronic dictionaries of Bulgarian idioms, the Explanatory Dictionary of the Bulgarian Language [2] and the Bulgarian WordNet [8] as well as automatically extracted from the Bulgarian National Corps [8]. Each component of MWE's entries in the dictionary of the Bulgarian MWE is applied a POS tag. To extract idiomatic predicative constructions from it at first the verbal entries from the dictionary were extracted and as a result a collection of 27,902 verb idioms was obtained. Then we applied to this collection the formal structural descriptions of predicative constructions [18] and extracted all VMWE's with auxiliary verb, noun/adjective/ adverb/ and, or preposition. Then the potential predicative VMWE records from the dictionary along with the examples from BNC were manually checked. As a result a collection of 1331 idiomatic predicative constructions was obtained and the new structure groups were extracted and analysed.

6 Constructional approach to classification of IPCs

After the excerption we collected over 2500 candidates for IPCs and after manual validation 1331 IPCs were selected. The observation over the structure types led to definition of four major types of idiomatic constructions with predicatives according to their component structure: a) idioms with plain predicative constructions; b) idioms with predicative constructions with pronoun clitics; c) idioms with prepositional predicative constructions; d) idioms with conjunctional predicative constructions. The subtypes share similar structure variations. The word order of the components is semi-fixed. The position of the constituents changes depending on the sentence context. The auxiliary verb is a clitic and cannot stand at the beginning of the sentence in present tense. The distribution of the subtypes within the excerpted data and the grouping of the formal constructional types is represented in Table 1.1.

Table 1. Distribution of the Types of Verbal Idiomatic Constructions with Predicatives in the Data

Costruction types	Occurrences of the types in the collected data	Examples
Aux V + Participle/N/Adv /Adj	337	izlyazal sam ot stroya (sth can't be used any more)
Aux V + Participle/N/Adv /Adj + Acc/Dat Pron	441	pisano mi e (its my faith)
Prep + Aux V + NP	496	v plen sum (feel captured from sth)
Conj + Aux V + NP/ Participle	57	kato che sam padnal ot nebeto (s.o. is very inadequate)

6.1 Idioms with Plain Predicative Constructions

This type of constructions consists of an auxiliary verb and participle (ex. 1) or NP (ex. 3) or AdvP (ex. 2). The entries in this group express common state semantics - characteristics or attributes of someone or something.

Examples: (1) varzan sum v racete (s.o. is not very skillful); (2) van ot sebe si sam (s.o. is very furious);((3) gola voda sam (s.o. or sth. is not valuable).

6.2 Idiomatic Predicative Constructions with Pronouns

This type of constructions include 4 subtypes:

a) IPCs with an auxaliry verb, combined with nouns with constrained grammatical (fixed in singular (ex. 4)) properties and/or a dative or an accusative pronoun clitic (ex. 5, 6)

Examples: (4) grehota e (its a sin), (5) yad me e (to feel rage),(6) zhal mi e(to feel sorry for s.o.)

b) IPCs with an auxiliary verb, combined with peredicative -o words (ex. 7) and/or a dative or an accusative pronoun (ex. 8)

Examples: (7) kasno e (it's late); ((8) dobre mi e (I'm fine).

c) IPCs with an auxiliary verb, predicative noun and a dative or an accusative pronoun, most often expressing emotion (ex. 9) or physical state (ex. 10).

Examples: (9) zhal mi e (I feel sorry); (10) ne me e enya (I don't care).

d) IPCs with an auxiliary verb in 3rd person sg. and a predicative adverb, most often meaning manner (ex. 12) or state (ex. 11). The same meaning share constructions with a predicative word and an omitted auxiliary verb (ex. 13)

Examples: (11) tamno e (it's dark), (12) fasulsko e (it's easy); (13) Veche se samnalo. (It's already dawn.)

6.3 Idioms with Prepositional Predicative Constructions

This type of constructions consists of an auxiliary, preposition and an NP and/or pronoun clitic. This is the less studied group of IPCs as they are predicative constructions, characterised with difficult to reduce ambiguity, caused by the position of the predicate and the predicative function of the preposition both ([11], [3]). In these constructions the forms of the auxiliary verb function same as full paradigm verbs and represent the semantic of presence or existence [12] in combination with the semantics of the individual prepositions. The specifics of the construction type arise from the preposition. Prepositional phrase can be in the first position in the sentence followed by a pronoun clitic. The entire group can be subdivided into different semantic sub-types due to the prepositional semantic:

IPCs denoting location and direction. The most frequent constructions are IPCs with the prepositions v (in) (ex. 14), mezdy (among), and po (along) (ex.15).

Examples: (14) V kravta vi/ti e. (It is in your blood.); (15) Po pat ni e. (It is on our way.).

IPCs denoting evaluation (ex. 16, 17), characteristics or attributes of someone or something (ex. 18, 17).The most frequent constructions are IPCs with the prepositions ot (from) (ex. 15, 19), s (with) and za (for) (ex. 18).

Examples: (16) Izobshto ne im beshe do horata. (They didn't care about people at all.); (17) Tova veroyatno im e ot golyama polza. (This is probably of great benefit to them); (18) Beshe mu za prav pat. (It was his first time.); (19) Samnyavam se, che shte vi bade ot koy znae kakva polza.(I doubt it will be of much use to you.)

6.4 Idioms with Conjuctional Predicative Constructions

This type of constructions consist of a conjunction, an auxiliary verb and participle or noun phrase or adverbial phrase. The entries in this group express common state semantics - characteristics or attributes of someone or something. Within this set two subgroups with idiomatized position of participle can be observed:

a) Positive Idioms with Conjuctional Predicative Constructions - Kato che sam (As if I am) and participle/NP

b) Negative Idioms with Conjuctional Predicative Constructions - Da ne sam (As if I don't) and participle

6.5 Idioms with Predicative Constructions and an Idiomatised Subject

This type of constructions consist of an auxiliary verb, a participle and a noun phrase, idiomatised in the subject position.

Example: (20) Zhivotat mi e mil (not to dare to brake rules)

7 Idioms with Predicative Constructions and State Semantics

As discussed in §7 we group the semantic types of IPCs based on the classification of Van Valin of predicates of state ([21]) into locative and non-locative ontological classes.

7.1 Locative IPCs

Locative IPCs denote assessment of state or position, represented mainly from IPCs with prepositions.

Examples: (21) na pat sum (i'm near the goal); (22)do krivata krusa sam (to be far from the aimed location).

7.2 Non-locative IPCs

Non-locative IPCs denote several subgoups:

a) assessment of physical condition - grogi sam (feel very tired); tip top sam (feel well);

b) assessment of mental condition - kato che sam padnal ot nebeto (to be very inadequate);

c) assesment of emotional condition - izvan relsi sam (feel very confused);

d) assessment of characteristics or attributes of s.o. or sth. - gola voda sam (s.o. or sth. is not valuable).

The semantic features are distributed over the construction types of IPCs. While assessment of characteristics or attributes is typical semantics for plain IPCs (ex. 1–3), IPC's with pronouns represent: assessment of characteristics or attributes (ex. 7, 8, 11, 12), assessment of mental condition (ex. 5, 6); assesment of emotional condition (ex. 9, 10). The IPCs with prepositions represent both locative (ex. 15) and non-locative groups - assessment of characteristics or attributes (ex. 17, 18), assesment of emotional condition (16) etc. This semantic typology of idioms with predicative constructions is compatible with the traditional grouping of predicates in semantic classes and additionally they represent the semantic feature assessment. This could be an argument in support of the hypothesis that regardless of composite component structure predicates are a semantic whole to which universal semantic features apply.

8 Conclusion and Future Plans

The article offers a classification of idiomatic predicative constructions in Bulgarian with a view to study of the ontological semantics of state. A method for excerption of idiomatic predicative data based on the formal of component's structure of VMWEs and Predicative constructions both was used. The constructional analyses of the collected data led to definition of four main structural groups and their respective subtypes. Each group represents variations of state semantics. The combined constructional and semantic analyses gives thoughts on questions for the vague boundaries between predicative constructions, composed verb forms and idiomatic predicative constructions and shows arguments for the statement that regardless of composite component structure predicative constructions are a semantic whole belonging to the universal semantic features of lexemes. The collected set of examples of IPCs will be used to study their argument structure and the lexicalization patterns associated with their syntactic representations.

References

1. Anderson, G., D.S.: Auxiliary Verb Constructions (2006)
2. Andreychin, L., Georgiev, L., Ilchev, S., Kostov, N., Lekov, I., Stoykov, S., Todorov, T.: Balgarski talkoven rechnik. Dopalneno i preraboteno izdanie ot D. Popov (Bulgarian Explanatory dictionary). Nauka i izkustvo. Sofia (2005)
3. Atanasov, A.: Classification of impersonal predicates in modern bulgarian. Bulgarian Language **62**(3), 38–53 (2015)
4. Georgiev, I.: Impersonal sentences in Russian and Bulgarian. Narodna prosveta. Sofia (1990)
5. Gregoire, N.: Duelme: a dutch electronic lexicon of multiword expressions. Language Resources and Evaluation pp. 23 — 39 (2010)
6. Ivanova, E.: Dativno-predikativnyye struktury v bolgarskom yazyke: statistika eksperimenta. Russkiy yazyk za rubezhom **5**, 11–17 (2018)
7. Koeva, S.: Inflection morphology of bulgarian multiword expressions. Proceedings of Azbuki@net, International Conference and Workshop. Sofia pp. 201–216 (2006)
8. Koeva, S.: Wordnet i bulnet. Ezikovi resursi i tehnologii za balgarski ezik pp. 154 — 173 (2014)
9. Kutsarov, I.: Teoretichna gramatika na balgarskia ezik: Morfologia. Publisher of Plovdiv University Paisiy Hilendarski (1999)
10. Maslov, Y.S.: Grammar of the Bulgarian language. Moscow (1981)
11. Nitsolova, R.: Bulgarian Grammar. Morphology. University Publishing House St. Kliment Ohridski, Sofia (2008)
12. Penchev, J.: Syntax of the modern Bulgarian literary language. Plovdiv (1998)
13. Petrova, G.: Nominal predicates joining the dative experiential in the bulgarian language: semantics and syntax. Russian Abroad **5**, 24–29 (2018)
14. Rozhnovskaya, M.: Impersonal sentences in the modern bulgarian literary language. Questions of grammar of the Bulgarian literary language pp. 379–432 (1959)
15. S., M.: Complex Predicates Verbal Complexes, Resultative Constructions, and Particle Verbs in German (2002)
16. Todorova, M.: Typology and Properties of Multiword Expressions in Bulgarian. Verb idioms. PHD Thesis. Institute for Bularian Language, Sofia (2015)
17. Todorova, M.: Formalized morphosyntactic dictionary of verb idioms in bulgarian. Papers of the Institute for Bulgarian Language Prof. Lubomir Andreychin **XXIX**, 207–243 (2016)
18. Todorova, M., Dimitrova, T., Stefanova, V.: Study of the boundaries of predicates for the state in bulgarian. (izsledvane na granitsite na predikativite za sastavyanie v balgarski). Sapostavitelno jezikoznanie **XLVI**(4), 21–46 (2021)
19. Todorova, M., Stoyanova, I.: Razrabotvane na rechnitsi na sastavnite leksikalni edinitsi v balgarskiya ezik za tselite na kompyutarnata lingvistika.(compiling dictionaries of bulgarian multiword expressions for the purposes of computational linguistics). Ezikovi resursi i tehnologii za balgarski ezik (2014)
20. Van Eynde, F.: Predicative Constructions: From the Fregean to a Montagovian Treatment. CSLI Publications (2015)
21. Van Valin, R., D.: A synopsis of role and reference grammar. Advances in Role and Reference Grammar (1993)
22. Van Valin, R., LaPolla, R.: Syntax: structure, meaning and function. Cambridge: Cambridge University Press (1997)
23. Villavicencio, A., Copestake, A., Waldron, B., Lambeau, F.: The lexical encoding of mwes. Proceedings of the ACL 2004 workshop on multiword expressions: Integrating processing. Barcelona, Spain pp. 80 — 87 (2004)

24. Zimmerling, A.B.: Imennye predikativy i dativnye predlozheniya v evropejskih yazykah. Computational Linguistics and Intelligent Technologies: Based on the materials of the annual International Conference "Dialogue" **9**(16), 549 – 558 (2010)

Frasemas en el habla de los jóvenes franceses

Antonio García Fernández

Universidad Nacional de Educación a Distancia, Madrid
agarcia@flog.uned.es

Resumen. El francés hablado por los jóvenes siempre se ha considerado un lenguaje vulgar y poco común, e incluso no se considera verdaderamente francés, sino un francés deteriorado, corrompido y fuera de la norma. Sin embargo, al estudiar a fondo este tipo de lenguaje, observamos que posee unas características propias interesantes para un estudio fraseológico. El presente estudio se centra en analizar las particularidades del lenguaje de los adolescentes a través de un corpus multimodal oral y escrito, y que tiene como objetivo presentar los frasemas más recurrentes por estos hablantes. Además, desarrollaremos otra serie de objetivos secundarios que complementan al objetivo principal como presentar estructuras léxicas prefabricadas de formas de habla que solo se muestran en situaciones determinadas; mostrar descripciones lingüísticas precisas sobre los rasgos característicos de la lengua hablada y que son innovadores en el proceso de adquisición de la lengua; y abordar el concepto de expansión y de durabilidad de expresiones fraseológicas en diferentes niveles lingüísticos.

Palabras clave: lenguaje, frases prefabricadas, norma, jóvenes, adolescentes, identidad, interacciones sociales

1 Introducción

Los adolescentes en Francia utilizan diariamente fórmulas como «*c'est chanmé*» (en español, «increíble») o «*j'ai la seum*» (en español, «soy un pringado»). Estas expresiones evidentemente no las podemos encontrar en un diccionario normativo, sino que se aprenden en las interacciones sociales entre jóvenes, es decir, en diversas situaciones y contextos propios de los adolescentes: los cambios de clase, el recreo o la salida del instituto, e incluso en las redes sociales.

Esta forma de habla en continua evolución se convierte en un rompecabezas para los adultos e incluso para aquellos que se adentran en el aprendizaje del francés como lengua extranjera. Si ya de por sí es difícil aprender una lengua extranjera, más aún es adaptarse a este tipo de contextos donde se utiliza un lenguaje que muchos consideran «no normativo».

1.1 La estabilidad de los frasemas

Todos los miembros de una misma comunidad lingüística no dominan sistemáticamente un gran número de frasemas. Al respecto, H. Burger (1998, citado por Schmale, 2013: 39) expone que el uso de una expresión determinada varía dependiendo de la comunidad de hablantes en la que se utilice, ya sea desde aquellos hablantes de una misma lengua (anglófonos, francófonos, hispanófonos) hasta grupos más reducidos como los hablantes de un país, de una región, de un barrio o incluso de un grupo de familias o individuos:

> *L'emploi courant, la dissémination, critère définitoire du phrasème d'après H. Burger (1998), ne reste par conséquent qu'un critère relatif, se référant à une utilisation ressentie comme fréquente à des échelles variées, allant du niveau des personnes parlant une même langue (p. ex. les anglophones, francophones ou germanophones), à celui de personnes de même langue vivant dans le même pays, la même région, le même village, le même quartier, etc., mais aussi à celui des groupes, des familles, même des individus.*

Asimismo, existen incluso «*phrasèmes d'auteurs*» (frasemas de autor) (Schmale, 2013: 39), es decir, frasemas que se limitan en el tiempo y en una situación determinada y que desaparecen en cualquier momento sin entrar en ningún tipo de diccionario. Este fenómeno se da sobre todo en frases prefabricadas, también conocidas como «clichés», empleadas dentro de diversos grupos sociales como, en este caso, en el lenguaje de los jóvenes franceses.

De hecho, los adolescentes utilizan fórmulas específicas por inercia sin ser conscientes de ello cuando interactúan. Sin embargo, desde un punto de vista social, según Polguère (2016: 9), si en un momento determinado el interlocutor no sabe qué decir o no sabe continuar con el hilo de la conversación utilizando el mismo lenguaje de esta comunidad de hablantes. Esto es debido a que no maneja del todo el mismo ritual lingüístico que su interlocutor.

Por lo tanto, podríamos decir que estamos bloqueados socialmente. Esto también ocurre generalmente en determinados ámbitos, como en una celebración religiosa o durante un encuentro profesional. Encontraremos términos propios del campo, pero también frases prefabricadas concretas (Polguère, 2016: 9). Los jóvenes, en este caso, utilizan este tipo de lenguaje con el fin de no sentirse excluidos y de crear un ambiente integrador dentro de sus grupos. Polguère (2016: 9) añade que el dominio de las fórmulas propias de los hablantes es por tanto la clave de la integración natural en el grupo:

> *Sans cliché, on se prive de la connivence que procure le fait d'employer les messages préfabriqués qui mettent de la ritualisation dans notre élocution. Et c'est justement la nature de rituel du cliché linguistique qui en détermine véritablement l'importance.*

Evidentemente, al ser un lenguaje de origen reciente no sabemos su durabilidad, puesto que la lengua siempre está sujeta a cambios y lo que hasta ahora puede considerarse como fijo, mañana puede estar totalmente en desuso, aplicarse en otro contexto diferente o adquirir un sentido distinto. De esta forma, Schmale (2013: 40) defiende que a pesar de la expansión y la posible corta durabilidad de estas expresiones no debemos alejarnos de estos fenómenos recurrentes, aunque estén a un nivel muy bajo de la comunidad lingüística:

> Ce qui ne signifie pas que toute production langagière préformée doive obligatoirement entrer dans le dictionnaire mais, d'un autre côté, on ne doit pas non plus écarter des

phénomènes de tous évidences récurrentes même s'ils se situent à une échelle d'utilisation bien en-dessous de la communauté langagière.

2 Corpus multimodal

Dada la dimensión del corpus y la limitación de este trabajo, no es posible presentar el corpus de forma extendida. Para ello, en este trabajo presentamos los frasemas más relevantes y característicos del habla de los jóvenes franceses en diferentes situaciones de comunicación dentro de registros de habla específicos.

Este corpus multimodal está formado por un corpus tanto oral como escrito. En cuanto a su forma oral, el corpus está constituido por el corpus MPF (*Multicultural Paris French*) y la emisora Radio jeune (Guillaume Radio 2.0); mientras que el corpus escrito está formado por los textos que aparecen en las redes sociales (Twitter, Facebook, Whatsapp, SMS, fórums).

El corpus MPF se inscribe en el marco de un proyecto sobre el lenguaje de la población joven. Este corpus se basa en el método microsociolinguístico, proveniente de la etnología y de la etnografía de la comunicación. En otras palabras, el investigador participa durante un periodo en la vida cotidiana y en las interacciones de un grupo de personas. Por otro lado, la emisión Guillame Radio 2.0 emplea el uso masivo del lenguaje no estándar por los oyentes e incluso de los mismos presentadores. El hecho de que el presentador indique la edad de cada participante al que llama durante la emisión nos permite centrarnos por ejemplo en una edad concreta. A su vez, dado que los jóvenes conocen la finalidad de la emisión, ellos hablan de forma espontánea, relajada y muy expresiva.

Nos interesaremos por tanto en situaciones de comunicación precisas en las que se emplean diferentes frases prefabricadas propias del habla de los jóvenes franceses.

3 Las frases prefabricadas en el habla de los jóvenes franceses

En el presente trabajo, nos centraremos en un tipo particular de frasema, en los fraseologismos pragmáticos denominados por Tutin (2019: 66) «frases prefabricadas en interacción». Estas frases constituyen el tipo de frasema más enunciado por los jóvenes hablantes. Charles Bally, pionero en este ámbito, nos habla por primera vez del término «*phraséologie exclamative*», como aquellas expresiones que se caracterizan por una entonación exclamativa y que suelen contener un componente afectivo y emocional. Posteriormente, dado el número creciente de estudios en este ámbito, encontramos numerosas denominaciones para referirse al mismo concepto: «*énoncés lié*» de Fónagy (1995), «*structure figées de la conversation*» de Bidaud (2002), «*actes de langage stéreótypes*» (Kauffer, 2019) o «*pragmatème*» de Mel'cuk (2013), Blanco & Mejri (2018).

Las frases prefabricadas en interacción se caracterizan por una serie de propiedades en diferentes niveles, a saber, léxico, semántico, pragmático e incluso psicológico y cognitivo.

Estas expresiones de carácter preformado y prototípico se tratan de enunciados completos ligados a una situación de enunciación específica. Este imperativo pragmático

hace que el enunciado actúe como reacción dentro de un acto de habla y, a su vez, adquiera un valor ilocutorio.

Para este estudio, hemos seleccionado un corpus de 11 frases prefabricadas, que son empleados habitualmente por los jóvenes y que están recogidos en los corpus orales y escritos mencionados en el apartado 2.

a) *J'ai la seum*
b) *Tu as le bon ice*
c) *C'est calé*
d) *C'est chanmé*
e) C'est chaud
f) *C'est de la balle*
g) *C'est un truc de ouf*
h) *Ça passe crème*
i) *Hey gros !*
j) *Je suis en bad*
k) *Je vais le faire*

A continuación, analizaremos las frases prefabricadas seleccionadas a partir de las nuevas tendencias sobre la investigación en la preformación lingüística.

3.1 Propiedades morfosintácticas

Desde un punto de vista morfosintáctico, las frases prefabricadas adquieren un estatus de enunciado, es decir, constituyen una expresión autónoma. Todas las expresiones que recogemos para el presente estudio están compuestas al menos de dos lexías, es decir, son expresiones poliléxicas. Podemos observar fórmulas que corresponden a los patrones sintácticos: c'est + Adj (*C'est chanmé*); c'est + SP (*C'est de la balle*); Je + V (*J'ai la seum*); Tu + V (*Tu as le bon ice*); Ça + V (*Ça passe crème*).

Sin embargo, no solo debemos atender a su estatus de enunciado, sino que debemos a atender también a lo que Schmale (2013: 41) considera «polifactorialidad», ya que estas fórmulas recogen factores tanto verbales (segmentales y suprasegmentales) como no verbales (contextuales, sociales, secuenciales, textuales, etc.). Por ejemplo, si seguimos las combinaciones propuestas por Schmale, estas fórmulas son polifactoriales ya que están formadas por dos lexemas *(j'ai la seum, c'est calé, c'est chaud*), un lexema y un elemento situacional concreto (Enunciar *ça passe crème*, mientras está comiendo por ejemplo su plato favorito) o un lexema y una actividad no verbal (en un saludo, enunciar *Hey gros !* mientras mueve la mano de lado a lado para saludar).

Asimismo, el carácter morfológico de enunciado genera una visión cognitiva, ya que, como expone Dostie (2019: 31), estas fórmulas se memorizan en bloque y se emplean en un momento preciso: «*la séquence préfabriquée est tenue pour un assemblage de eux ou plusieurs unités linguistiques mémorisées en bloc, qui sont utilisées en contiguïté ou à proximité dans un texte*».

3.2 No composicionalidad y fijación

Desde un punto de vista semántico, estas expresiones se consideran no composicionales, ya que el sentido global no se constituye a partir de sus constituyentes léxicos. Esta no composicionalidad es también conocida por otros autores como idiomaticidad o fijación semántica.

Como podemos observar en algunas expresiones como *je vais le faire*, la no composicionalidad no implica necesariamente la opacidad semántica de los frasemas. Comprobamos como la expresión *c'est cool*, cuyo semantismo se deduce por la lexía *cool*; mientras que el mismo sentido de la expresión en este tipo de habla de los jóvenes, *c'est calé* puede ser opaco por el locutor no familiarizado con este registro de lengua. Forma coloquial: *Les fringues de chez Stéphane, on a beau dire, mais c'est trop calé*; forma estándar: *Force est de constater que les vêtements qui viennent de chez Stéphane et avec lesquels on peut frimer car il est très difficile de se les procurer en France, c'est quand même trop cool.*

De esta forma, observamos que el registro forma parte también del sentido pragmático, que a su vez constituye un factor fundamental para el componente semántico, ya que si empleamos el enunciado en otro contexto diferente perdería su sentido.

3.3 Propiedades pragmáticas: el valor ilocutorio

Estas fórmulas, además de tener una apariencia meramente idiomática, también poseen una apariencia pragmática y social, puesto que se utilizan en unas situaciones de comunicación específicas.

Las frases prefabricadas adquieren un carácter social. Polguère (2016: 8) defiende que son sobre todo las frases prefabricadas (también conocidas como *clichés*) las que desempeñan una función fundamental en la interacción social:

> *Il est certain que les locutions et les collocations ont un rôle social, puisque le respect ou le non-respect des règles n'est pas neutre dans l'interaction sociale ; mais ce sont avant tout les clichés – parce qu'ils sont des contenus préfabriqués avant d'être des segments de langue préfabriqués – qui jouent un rôle sur le plan de l'interaction sociale.*

La utilización de estas fórmulas responde a un conjunto de reglas dentro de una interacción social. Este imperativo social ha sido identificado por Wray (2012: 231-232), que expone:

> *Humans, being psychologically and socially complex, are unable fully to meet their emotional, mental, and physical needs without involving others. One effective tool for drawing others into behaviors beneficial to us is to employ wordstrings that are in current use in our community. They enable us socially to align ourselves with others (I am like you because I talk like you, so you will want to help me), and as a way of minimizing the risk of misunderstanding, since wordstrings or partly lexicalized frames that have their own semantic entry require less decoding.*

Por lo tanto, en este tipo de lenguaje, los jóvenes deben poseer un excelente dominio de este tipo de lenguaje para utilizar determinadas frases prefabricadas propias y evitar así la marginalización social.

Schmale (2013: 36) añade que la adecuada enunciación de este tipo de frasemas necesita el cumplimiento de condiciones secuenciales, contextuales, sociales, estilísticas y otras condiciones muy específicas:

> *Les expressions préformées […] sont rattachée pragmatiquement à des contextes d'utilisation, leur réalisation adéquate nécessitant le respect de conditions séquentielle, contextuelle, sociales, stylistiques, etc. très spécifiques.*

Desde un punto de vista interaccionista, dentro de la Teoría de los actos de habla de Austin (1962) y Searle (1969), los actos de habla constituyen también otra propiedad definitoria de las frases prefabricadas, ya que pueden representar una reformulación, una aprobación, una excusa, una expresión de una emoción, etc. El locutor no solo emite el mensaje, sino que en el momento de emitirlo adquiere también un valor ilocutorio.

Por ejemplo, la fórmula *ça passe crème* se emplea exclusivamente dentro de un acto de habla estrechamente ligado a unas condiciones y a unas limitaciones temáticas, secuenciales y situacionales. En este caso, esta fórmula se da en una situación en la que el locutor está pasando un momento agradable; sin embargo, una expresión como *tu as le bon ice* o *c'est un truc de ouf* no solo reúne las condiciones anteriores, sino que además está estilísticamente marcada dentro de un registro familiar, es decir, está reservado únicamente a situaciones y a hablantes que aceptan este nivel de lengua.

La fórmula *je suis en bad* muestra un valor ilocutorio de indignación del hablante ante una triste situación. El uso de esta fórmula se acompaña de un sentimiento de tristeza o aflicción. Del mismo modo que la fórmula *j'ai la seum*, que muestra una indignación.

Otras de las fórmulas muy recurrentes por los jóvenes son *c'est calé*, *c'est un truc de ouf* ou c'est chanmé o *tu as le bon ice*. Estas fórmulas poseen un valor ilocutorio de sorpresa y aprobación, es decir, la intención del locutor es la de realizar un acto de aprobación como reacción a una situación que le sorprende o a una idea extraordinaria.

4 Conclusión

Los jóvenes utilizan determinadas prácticas lingüísticas que forman parte de la producción simbólica de su identidad. Sus hablantes intentan crear más excepciones no solo para mantener esta jerga como un marcador de identidad y distinguirse de otros, sino también como un método para crear un lenguaje encriptado.

Gracias a este estudio, observamos que estas frases prefabricadas en interacción constituyen por tanto unas formas preconstruidas y forman parte del stock del lenguaje de los jóvenes franceses.

El uso de estas fórmulas espontáneas y poco estudiadas nos han servido para describir y profundizar en una forma de habla muy recurrente por los jóvenes. Se trata de fórmulas constituidas por unos enunciados completos y ligados a una situación de enunciación específica. Comprobamos que el criterio determinante en la mayoría de los casos y que va a determinar la definición de este tipo de fórmulas es su valor pragmático y, en particular, su valor ilocutorio. Sin embargo, continuaremos en el estudio de los aspectos pragmáticos y nos dirigiremos hacia otras perspectivas como la prosodia.

Referencias bibliográficas

1. Dostie, G. Paramètres pour définir et classer les phrases préfabriquées : La vengeance est un plat sui se mange froid. Bon appétit !. *Cahier de lexicologie*, 114 (1), 27-62 (2019).
2. Polguère, A, Il y a un traître par minou : le statut lexical des clichés linguistiques. *CORELA - COgniton, REprésentation, LAngage. In* http://corela.revues.org/4486, último acceso 04/04/2022 (2016b).
3. Polguère, A., *Lexicologie et sémantique lexicale. Notions fondamentales.* Montréal: Les Presses de l'Université de Montréal (2016a).
4. Schmale, G., Qu'est-ce qui est préfabriqué dans la langue ? – Réflexions au sujet d'une définition élargie de la préformation langagière. *Langages*, 189, 1, 27-45 (2013).
5. Tutin, A. Phrases préfabriquées des interactions : quelques observations sur le corpus CLAPI. *Cahier de lexicologie*, 114 (1), 27-62 (2019).
6. Wray, A., What Do We (Think We) Know About Formulaic Language? An Evaluation of the Current State of Play. *Annual Review of Applied Linguistics*, 32, 231-254 (2012).

The Phraseological Units of Arabic and Their Equivalents in Russian

Rafis Zakirov[1] [0000-0001-9725-8275] Nailya Mingazova[2] [0000-0002-3097-0193] Vitaly Subich[3] [0000-0003-2003-7866] Alfiya Khabibullina[4] [0000-0002-6866-9419]

[1] Kazan Federal University, Russia
[2] Kazan Federal University, Russia
[3] Kazan Federal University, Russia
[4] University of Wolverhampton, UK

Abstract. This study deals with the problem of equivalents of phraseological units, namely, idioms in the Arabic-Russian language pair. The article describes the main ways of translation of Arabic phraseological units into Russian. Possible scenarios for equivalency existing between the languages are described: full equivalents, partial equivalents, analogues, non-equivalent (lacunar) units. The following methods are used for translation: loan translation, descriptive translation, lexical translation and combined translation. It is revealed that the phraseological units, the meaning of which goes back to a historical event or reflects a custom are the most difficult for translation. At the same time, the translation of Arabic phraseological units into Russian requires a mandatory knowledge of vocabulary, an ability to analyze the context, a thorough study of the history, culture, life and customs of the people who speak the source language, as well as the target language speakers.

Keywords: Phraseology, Phraseological Units, Equivalent, Lexical Component, Translation, Arabic, Russian.

1 Introduction

Phraseology reflects the national identity of the language, folk wisdom, historical and cultural experience of people. The majority of phraseological units carries a certain expressive and stylistic load. In each case, the context specifies or changes the meaning of a phraseological unit. If it has a contextual meaning, it can differ significantly from the meanings of the phraseological unit given in the dictionaries.

Linguists still cannot come to a common opinion on a number of problems of phraseology, and this is quite natural. Phraseology of the Arabic language was studied in Ushakov V.D. (1996), Hussein S.H. (2014), Zakirov R.R. et Mingazova N.G. (2014, 2015), Mingazova N.G., Subich V.G., Al-Foadi R.A. et Zakirov R.R. (2019) and other works. This work has explored over 2000 phraseological units from the phraseological dictionary 'Arabic-Russian dictionary of phraseological units' by Abi Jaber (2018). As a result a systematized presentation of translation methods was developed.

2 Methods of translation

Phraseological units are the most difficult to translate. This is why the translation of phraseological units is the result of a thorough analysis of the various components of the content structure of a phraseological unit.

When translating a phraseological unit, it is necessary to convey its meaning and reflect its imagery, finding a similar expression in the Russian language and not losing sight of the stylistic function of the phraseological unit. If there is no identical image in the Russian language, the translator is forced to resort to an "approximate match".

Phraseological translation involves the use of stable units of varying degrees of proximity between the unit of a source language and the corresponding unit of the target language, from full and absolute equivalent to approximate phraseological correspondence.

One of the first classifications of phraseological units of the Arabic language was proposed by V. D. Ushakov, who notes that non-free word groups in modern literary Arabic are "units that have the properties of idiomaticity, stability; phrases in which there is a violation of morphological and syntactic norms; figurative units that include repetitions, paired combinations in which euphonic means are used (rhymed consonances, alliteration); paired combinations of synonyms and antonyms; phrases containing words of the same root; periphrases, proverbs and sayings, interjective phrases and cliches, units that serve the purpose of naming certain objects and phenomena; combinations in which each of the components, or one component, tends to turn into a part of the word (to morphologize)" [Ushakov 1964: 3].

Such a variety of units included in the category of "non-free word groups" is explained by their common feature, namely, the semantic unity of the components of the phrase. Different degree of the semantic coherence of the phrases served as a basis of their division into groups.

When translating phraseological units of Arabic into Russian, it is advisable to rely on structural-typological and functional-semantic approaches to establish the equivalence of phraseological units. This allows us to distinguish the following interlanguage relations:

1) full equivalents;
2) partial equivalents;
3) analogues;
4) non-equivalent (lacunar) units.

For the translation of non-equivalent (lacunar) units, the following methods are used: loan translation, descriptive translation, lexical translation, and combined translation.

2.1 Full phraseological equivalents

Full phraseological equivalents are phraseological units of the Arabic and Russian languages that coincide semantically, lexically, and stylistically. But with this coincidence of structural and grammatical features, we take into account the specificity of typological features inherent in one language and not characteristic of another. Full equivalents in their functional and stylistic characteristics are either interstitial or bookish. Literary-

colloquial phraseological units are quite rare. The phenomenon of complete equivalence is not typical for stylistically reduced phraseological units of the Arabic language. For example, full phraseological equivalents in the Arabic and Russian languages are idioms بَيْنَ نَارَيْنِ and между двух огней (between two fires). Here we observe complete identity of the semantics of their lexical components and stylistic peculiarities.
Full phraseological equivalents in Arabic and Russian can be represented by the following examples:
بَنُو آدَمَ – дети Адама (Adam's children)
فِي الْبَرّ وَالْبَحْرِ – на суше и на море (on land and at sea)
أَقْحَمَ أَنْفَهُ فِي – совать свой нос (куда-либо) (poke his nose (anywhere)
لاَ يَرَى أَبْعَدَ مِنْ أَنْفِهِ – дальше носа своего не видит (he can't see further than his nose)
فِي لَمْحِ البَصَرِ– в мгновение ока (in the blink of an eye)
تَجَمَّدَ الدَّمُ فِي عُرُوقِهِ – кровь в жилах стынет (the blood runs cold in his veins)
يَسْفِكُ الدِّمَاءَ – проливать кровь (to shed blood)
فِي اللَّيْلِ وَالنَّهَارِ – днем и ночью, денно и нощно (day and night)
لَقِيَ حَتْفَهُ – встретить смерть (to meet death)
حَطَّمَ الأَغْلاَلَ – разорвать оковы (to break the shackles)
بَعَثَ الحَيَاةَ فِي– вдохнуть жизнь во что-либо (to breathe life into something)
خَرِيفُ العُمْرِ – осень жизни (autumn of life)
خَيَّمَ الظَّلاَمُ عَلَى – опустилась тьма (darkness descended)
ذَرَّ الرَّمَادَ فِي العُيُونِ – пускать пыль в глаза (to throw dust in the eyes)
تَدُورُ أَعْيُنُهُمْ – глаза закатываются (от страха) (eyes roll up (from fear)
يَرَوْنَهُمْ رَأْيَ الْعَيْنِ– увидеть кого-либо воочию (to see someone in flesh)
قَطَعَ دَابِرَهُمْ – истребить всех до последнего (to exterminate every last one of them)

2.2 Partial phraseological equivalents

Partial phraseological equivalents include the units of the Arabic and Russian languages that fully convey the semantic and stylistic coloring of phraseological units, but differ in structural and grammatical organization and composition.

Partial phraseological equivalents among the phraseological units of the Arabic and Russian languages can be divided into two groups:

1) phraseological units that coincide in meaning, stylistic peculiarities and are close in imagery (the grammatical structure may or may not coincide), but differ somewhat in lexical composition.

For example, in the Arabic and Russian phraseology عَضَّ عَلَى أَنَامِلِهِ and кусать локти (от отчаяния) (biting elbows (from despair), which coincide in meaning, stylistic orientation and are close in imagery, only the components أَنَامِلِهِ (his finger tips) and локти (elbows) differ.

Partial phraseological equivalents in Arabic and Russian are given below:
يَوْمُ الدِّينِ (lit. The day of Religion) – Судный день (Judgment Day)
يَوْمُ الْقِيَامَةِ (lit. The day of the rising from the graves) – Судный день (Judgment Day)

إِنْقَلَبَ عَلَى عَقِبَيْهِ (أَعْقَابِهِمْ)
(lit. to roll over on his heel (their heels)) – обращаться вспять – (to reverse)
قَلَّبَ كَفَّيْهِ (lit. to turn his palms) – заламывать руки (от отчаяния, сожаления, раскаяния) (wringing your hands (from despair, regret, remorse)
حَجَرُ الزَّاوِيَةِ (lit. the corner stone) – краеугольный камень (the corner stone)
وَضَعَ حَجَرَ الأَسَاسِ
(lit. to lay the foundation stone) – заложить камень в фундамент (to lay the foundation stone
مَوْقِفٌ لاَ يَحْسُدُ عَلَيْهِ
(lit. a position that is not envied) – незавидное положение (an unenviable situation)
حَصَدَ مَا زَرَعَهُ (lit. he reaped what he sow) – пожинать плоды (to reap the fruits)
2) phraseological units that coincide in meaning, imagery, and stylistic peculiarities, but differ in the category of nominal number.
Thus, in the following phraseological units of Arabic and Russian اقْتَلَعَ الشَّيْءَ مِنْ جُذُورِهِ and вырвать с корнем (snatch anything with his root) we see that in Arabic the noun is plural: جُذُورِهِ (his roots), and in the Russian language the noun корень (root) is singular.
Examples of this type of partial phraseological units in Arabic and Russian are the following phrases:
عَنْ يَدٍ (lit. by hand) – из рук в руки (from hand to hand)
تَغَيُّرٌ جِذْرِيٌّ (lit. root change) – коренные изменения (fundamental changes)
لَمْ يَغْمُضْ لَهُ جَفْنٌ (lit. he didn't even close his eyelids) – века не сомкнул (I didn't even close eyelid)

2.3 Phraseological analogues

Phraseological analogues are considered to be phraseological units of the Arabic and Russian languages that coincide in semantics and stylistic variation, but differ in imagery (the grammatical structure may or may not coincide). Analogues convey unique images and concepts that make up the national identity of the compared languages.

For example, in the phraseological units of Arabic and Russian انْشَقَّتِ العَصَا بَيْنَهُمْ (lit. a stick cracked between them) and черная кошка пробежала (между ними) (a black cat ran (between them) different imagery is used (a cracked stick and a black cat), although these phraseological units coincide in semantics and stylistic meaning.

Phraseological analogues in Arabic and Russian can be represented by the following examples:

ذَاتُ الصُّدُورِ (lit. what's in the chest)– тайники души (hiding places of the soul)

قُرَّةُ عَيْنٍ (lit. coolness of the eye)– услада очей (delight of the eyes)

الْقُلُوبُ لَدَى الْحَنَاجِرِ (lit. hearts at the throat)– душа в пятки ушла (от страха) (the soul has gone to the heels (from fear)

سُقِطَ فِي أَيْدِيهِمْ (lit. to fall into their hands) – они были приперты к стене (they were pinned to the wall)

2.4 Non-equivalent phraseological units

Non-equivalent (lacunar) phraseological units are phrases that do not have correspondences in the phraseological system of another language. These units reflect the peculiarities of psychology, ways of thinking, and specific conditions for the development of the material and spiritual life of native speakers.

Among the methods of translation of non-equivalent phraseological units of the Arabic language, we distinguish loan translation, descriptive translation, lexical method of translation, and combined translation.

Loan translation. Phraseological calques are phraseological units that appear as a result of the exact or modified structure and meanings of target language prototypes (usually phraseological units) by means of the borrowing language. The initial perception of the content of foreign prototypes in the form of concepts which usually do not immediately get the quality of the language structure according to the dictated schemes, is a common feature of phraseological and word-forming calques.

Loan translation is used when translating non-equivalent phraseology when a phrase cannot be translated using other types of translation, or when another language has a phraseological unit with the same meaning, but its use would lead to a distortion of the phrase or to the complete loss of its color.

So, the phrases in the Arabic and Russian languages الضَّوْءُ الأَخْضَرُ and зеленый свет (green light) are phraseological calques.

Phraseological calques of Arabic and Russian languages are quite numerous and common in use:

عَلَى أَبْوَابٍ – у ворот (at the gate)
إِرْتَفَعَتْ أَسْهُمُهُ – его акции поднялись (his stock has gone up)
نَصِيبُ الأَسَدِ – львиная доля (the lion's share)
بُرْجٌ عَاجِيٌّ – башня из слоновой кости (the ivory tower)
تَجْمِيدُ الأَسْعَارِ – заморозить цены (to freeze prices)
سَبَحَ ضِدَّ التَّيَّارِ – плыть против течения (to swim against the current)
إِنْجَرَفَ مَعَ التَّيَّارِ – плыть по течению (to go with the flow)
صِرَاعُ حَيَاةٍ أَوْ مَوْتٍ – борьба не на жизнь, а на смерть (to fight for life and death)
شَرِيكُ الحَيَاةِ – спутник жизни (life partner)
خَطُّ النَّارِ – линия огня (line of fire)
أَدَارَ دَفَّةَ الحُكْمِ – стоять у руля власти (to stand at the helm of power)
دَقَّ مِسْمَارًا فِي نَعْشٍ – вбить гвоздь в гроб (to drive a nail into the coffin)
دُمُوعُ التَّمَاسِيحِ – крокодиловы слезы (crocodile tears)
سَفِينَةُ صَحْرَاءِ – корабль пустыни (a desert ship)
الجِنْسُ اللَّطِيفُ – прекрасный пол (о женщинах) (the fair sex (about women)

Descriptive translation method. The essence of this method of translation is to convey the meaning of a phraseological unit using words, phrases and sentences. The descriptive method of translation most fully reveals the essence of the described phenomenon,

but in the descriptive translation, the figurative expression loses its imagery, only its general phraseological meaning is transmitted.

For example, for translating idioms of The Holy Quran text سَكْرَةُ المَوْتِ (lit. the intoxication of death) from Arabic into Russian the descriptive method is used, since it is the only and most adequate method in this case: предсмертная агония (the agony of death).

The following phraseological units of the Arabic language are translated in the same way:

زَهْرَةُ الْحَيَاةِ الدُّنيَا (lit. the flowering of the shortest life) – лучшая пора жизни земной (the best time of life on earth)

قُلُوبُنَا غُلْفٌ (lit. our hearts are uncircumcised) – наши сердца не обрезаны (our hearts are inaccessible to faith)

كَلَ لَحْمَ (فُلاَنٍ) (lit. to eat someone else's meat) – порочить, поносить кого-либо, клеветать на кого-либо (to defame, vilify someone, slander someone)

ثَبَّتَ فُؤَادَهُ (بِهِ) (lit. to strengthen someone's heart) – укреплять кого-либо в вере (to strengthen someone's faith)

خَتَمَ (طَبَعَ) عَلَى سَمْعِهِمْ (lit. to put a seal on their hearing) – лишить возможности воспринимать истину (to deprive them of the ability to perceive the truth)

قَابَ قَوْسَيْنِ (lit. at a distance of two bows)– очень близко (very close)

عَلَى أَبْصَارِهِمْ غِشَاوَةٌ (lit. there is a veil over their eyes)– на их взорах покровы (about the rejection of faith)

The lexical method of translation. In some cases, a single word can be used to translate phraseological units of the Arabic language.

As an example, let us examine the idiom of the Arabic language of The Holy Quran سِرَاجٌ مُنِيرٌ (illuminating lamp) that is translated into Russian as пророк (prophet).

The lexical method is used to translate a number of phraseological units of the Arabic language into Russian:

غَلِيظُ الْقَلْبِ (*lit.* hard-hearted)	– жестокосердный (hard-hearted)
الْقَاسِيَةُ قُلُوبُهُمْ (*lit.* cruel hearts)	– жестокосердные (hard-hearted)
هَبَاءً مَنْثُوراً (*lit*. scattered dust)	– ничем (by nothing)
وَلا رَطْبٍ وَلا يَابِسٍ (*lit.* neither wet nor dry)	–ничего (of nothing (no one)
إِبْنُ السَّبِيلِ (*lit*. son of the way)	– путник, странник (a stranger)
صِرَاطٌ مُسْتَقِيمٌ (*lit*. the true path)	– религия, вера (religion, faith)
عَلَى حَرْفٍ (*lit*. on the edge)	– неуверенно (uncertainly)
إِنْتَقَلَ إِلَى جِوَارِ رَبِّهِ (*lit*. to go to the Lord)	– умереть (to die)

Combined translation. Combined translation is necessary to convey the meaning of the phraseological unit into another language as fully as possible and to present all the

existing possibilities of its translation. Usually, when using a combined translation, a descriptive translation that explains the phraseological unit is given.

So, to convey the meaning of The Holy Quran idiom أَهْلُ الْكِتَابِ from Arabic into Russian the combined translation is used: люди Писания (иудеи и христиане) (the people of the Scripture (Jews and Christians).

The following phraseological units of the Arabic language are translated into Russian in the same way:
أُمُّ الْقُرَى (lit. mother of cities) – мать поселений (Мекка) (the most important city, the main settlement) (Mecca)
ذُو الأَيْد (lit. possessor of hands) – обладатель рук (король Давид) (possessor of power (or might) (King David)
خَاتَمُ النَّبِيِّين (lit. the seal of the prophets) – печать пророков (последний из пророков) (Мухаммад) (the last of the prophets) (Muhammad)
قَاصِرَاتُ الطَّرْفِ (lit. with downcast eyes) – с поникшим взором (modest, shy) (virgins)
وَأَفْئِدَتُهُمْ هَوَاءٌ (lit. and their hearts are air) – в их сердцах – воздух (их сердца пусты) (their hearts are empty) (they do not perceive anything, do not understand)
جَعَلَ صَدْرَهُ ضَيِّقاً حَرَجًا (lit. to constrict the chest) – сдавить грудь (поставить кого-либо в затруднительное положение) (to put someone in a difficult position) (to deprive him of the opportunity to go the right way).

Phraseological units whose meaning goes back to some historical event or reflects some customs of the Arabs are particularly difficult to translate. Moreover, sometimes the apparent semantic clarity turns out to be deceptive.

Thus, the analysis of the component values of the Arabic idiom مَخْتُومٌ عَلَى قَفَاهِ (*lit.* with a seal on the back of the head) does not give the key to the translation, if you do not know that the illiterate, dark Egyptian peasants treated any serious diseases by applying a hot iron to the back of the head, after which a mark remained for life, indicating the extreme stupidity of the marked with this seal. It is translated into Russian as с клеймом на затылке.

3 Conclusion

To conclude, the translation of the phraseological units of the Arabic language into the Russian language demands using structural-typological and functional-semantic approaches and find their equivalents. Comparing the Arabic and Russian phraseological units the following interlanguage relations could be distinguished: full equivalents, partial equivalents, analogues and non-equivalent (lacunar) units. Non-equivalent (lacunar) units are translated by loan translation, descriptive translation, lexical translation, and combined translation methods.

It is important to note that the translation of Arabic phraseological units into Russian requires a mandatory knowledge of vocabulary, an ability to analyze the context, and a thorough study of the history, culture, life and customs of the people who speak the source language, as well the target language.

8

References

1. Finkel'berg N.D. Kurs teorii perevoda – M.: Vostok-Zapad, 232 p. (2004).
2. Grande B.M. Kurs arabskoj grammatiki v sravnitel'no-istoricheskom osveshhenii. 2-e izd. – M.: Vostochnaya literatura, 592 p. (1998).
3. Hussein, S.H. An introduction in lingual contrast between Arabic and the other languages // Journal of the college of languages. Vol. 29, pp. 98-128. (2014).
4. Mingazova N.G., Zakirov R.R. Sopostavitel'naja tipologija tatarskogo i arabskogo jazykov. Chast' 1: uchebnoe posobie – Kazan': Izd-vo Kazanskogo un-ta. – 236 p. (2014).
5. Mingazova, N.G., Safin, M.F., Zakirov, R.R. Sopostavitel’naya grammatika angliskogo i arabskogo yazykov. Chast’ 3 – Kazan: izdatelstvo kazanskogo universiteta, 160 p. (2019).
6. Mingazova N.G., Subich V.G., Al-Foadi R.A., Zakirov R.R. Phraseological Units of the Idafa Type in the Quran. Opcion 35 (Special Issue 22) Universidad del Zulia, ISSN: 1012-1587, pp. 682-698. (2019).
7. Abi Jaber, J., Kapshun, A.V. Arabsko-russkiy slovar isiomaticheskish vyrazheniy. – M: YASK, 456 p. (2018).
8. Zakirov, R.R., Mingazova, N.G., Yuzmukhametov, R.T. Frazeologia Qorana – Kazan: Izdatelstvo Kazanskogo Universiteta, 367 p. (2008).
9. Zakirov R.R., Mingazova N.G. Teorija i praktika perevoda (arabskij jazyk) – Kazan': RII, 268 p. (2015).
10. Ushakov V.D. Frazeologiya Qorana: Opyt sopostavlenia frazeorecheniy Korana i arabskogo klassicheskogo yazyka. (1996)
11. Mingazova N.G., Zakirov R.R., Shuinshaliyeva A.N. The Analysis of the English and Kazakh Idioms with Lexical Components «Truth» and «Lie». Applied Linguistics Research Journal. – Vol. 4, Is. 7, pp. 32-37. (2020).

Variabilidad fraseológica y forma citativa en los diccionarios bilingües (español ↔ catalán) en línea

Joseph García Rodríguez[1][0000-0001-7264-0347] y Marta Prat Sabater[2][0000-0002-4462-5403]

[1] Universidad Nacional de Educación a Distancia, 28040 Madrid, Spain
joseph.garcia@flog.uned.es
[2] Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain
marta.prat@uab.cat

Resumen. La variabilidad suele ser propia de las unidades fraseológicas (UF) a pesar de que tradicionalmente estas se hayan relacionado con la fijación. La clave reside en que, a pesar de la existencia de variaciones, variantes o desautomatizaciones, el contenido semántico no se caracteriza por su modificación. La forma citativa de esta variabilidad debe aplicarse de un modo adecuado a cualquier tipo de obra lexicográfica y, en concreto, a los diccionarios bilingües en línea, en este caso, relacionados con las lenguas española y catalana en ambas direcciones. Los resultados del análisis oportuno de los pares de obras en línea de cuatro editoriales distintas que se han realizado permiten demostrar la ausencia de sistematización y, por consiguiente, la necesidad de tener que unificar al respecto los distintos criterios (constitutivo, de frecuencia de uso, organizativo, tipográfico y ortográfico) acerca de cómo deben presentarse estas expresiones en los diccionarios.

Palabras clave: variabilidad fraseológica, forma citativa, lexicografía.

1 La variabilidad fraseológica

La fijación es una de las características que tradicionalmente se ha asociado a las UFS, aun así, se pueden producir ciertas modificaciones que propicien que no siempre los elementos que las forman aparezcan del mismo modo. En este sentido, gracias a la información que proporcionan los corpus textuales, se ha comprobado que la variabilidad es un rasgo importante de estas expresiones.

En la presente investigación se tomarán como punto de partida los estudios de Mena Martínez (2003) y Montoro del Arco (2005). En primer lugar, Mena Martínez (2003) prefiere diferenciar entre *desviaciones* (existencia de UFS que se producen por un lapsus percibido por el receptor), *variantes* (fraseologismos que forman parte de la lengua y que comparten significado y algunos elementos léxicos con la UF original) y *UFS modificadas o desautomatizadas* (las que resultan de un proceso creativo y de carácter voluntario por el emisor). En segundo lugar, Montoro del Arco (2005) presenta una clasificación atendiendo a si las modificaciones que se producen en las UFS forman parte de la misma lengua funcional (internas) o si, por lo contrario, conllevan un cambio en esta misma (externas). A partir de esta diferenciación, el autor distingue dos tipos de alteraciones: la *variación fraseológica* y, al igual que Mena Martínez (2003), la *desautomatización fraseológica*. El hiperónimo de *variación fraseológica* incluye las

variantes y *variaciones estructurales*, ambas internas o externas. Las *variantes* internas pueden ser léxicas, morfoléxicas y/o gramaticales. Las *variaciones estructurales*, léxico-cuantitativas y morfosintácticas. Las *variantes* externas se pueden diferenciar en *léxicas* y *gramaticales*, según los distintos tipos de variedades de la lengua. Las *variaciones estructurales*, al igual que las internas, se distinguen entre léxico-cuantitativas y morfosintácticas.

A partir de lo expuesto por los distintos especialistas, se opta por la siguiente clasificación sobre la variabilidad fraseológica:

(A) Las variaciones
  I. Adición o supresión [*poner tierra (de) por medio*]
  II. Variaciones gráficas [*a boca de jarro*/*a bocajarro*]
  III. Alteración del orden de los constituyentes [*con una mano delante y otras detrás*/*con una mano detrás y otra delante*]
  IV. UFS con casillas vacías [*a mi*/*tu*/*su aire*]
  V. Nominalizaciones [*tomar el pelo*/*tomadura de pelo*]

(B) Las variantes
  I. Léxicas [*quedarse (algo) en agua de borrajas*/*cerrajas*]
  II. Morfosintácticas [*hacer (algo) agua(s) por todas partes*]

(C) La desautomatización [*al mal tiempo, buena carrera* (*El Correo*, 16/05/2016) en lugar de *al mal tiempo, buena cara*]

Al igual que Montoro del Arco (2005), se parte de la premisa de que todo lo que forma parte de la variabilidad fraseológica afecta directamente al conjunto de la estructura o a algún elemento interno de la unidad; sin embargo, una condición indispensable que debe darse en todo momento es el mantenimiento del significado.

2 La forma citativa de los fraseologismos en la lexicografía

Resulta complicado incorporar la variabilidad fraseológica a los diccionarios (Wotjak B., 1998; Wotjak G., 1998, 2006, 2008 y 2010; García-Page, 2008; Mellado Blanco, 2008; Montoro del Arco, 2004 y 2008; Ortega Ojeda & González Aguiar, 2008; Wotjak B. & Wotjak G., 2014). Los problemas que plantea la forma citativa en una obra lexicográfica en papel quedan solventados en el soporte electrónico, ya que permite incluir un mayor número de casos. Los autores anteriormente citados están de acuerdo en incluir, bajo diferentes marcas tipográficas, las posibles variantes y variaciones que pueda registrar una UF, aunque destacan la complejidad que esto conlleva (Wotjak G., 2010).

A la hora de integrar la forma citativa en los diccionarios es preciso considerar bajo qué lema o lemas deben ubicarse las UFS recogidas; qué variantes y variaciones de una UF tienen que aparecer en el diccionario; y cómo deben presentarse las distintas formas de variabilidad de dichas unidades. En las obras lexicográficas electrónicas resulta imprescindible diferenciar los criterios esenciales para sistematizar la forma citativa de las UFS:

(1) Criterio constitutivo (presentación de los elementos propios de la UF y de los que forman parte del contorno)

(2) Criterio de frecuencia de uso (presentación de las distintas variantes y variaciones de una misma UF)
(3) Criterio organizativo (presentación de los elementos fijos de la UF y de los variables)
(4) Criterio tipográfico y ortográfico (determinación de los recursos que permitan distinguir el núcleo y el contorno de las UFS, como negrita, cursiva, barras oblicuas, paréntesis o corchetes)

En el siguiente apartado, se analizará la forma citativa de los diccionarios bilingües (español ↔ catalán) en línea para averiguar su adecuación (para más información, véase García Rodríguez, 2021).

3 La variabilidad fraseológica en los diccionarios bilingües español-catalán en línea: análisis de la forma citativa

Se han examinado ocho diccionarios bilingües español-catalán en línea de cuatro editoriales distintas (grup enciclopèdia, Vox-Larousse, Glosbe y dict.com), con el fin de analizar la forma citativa que se emplea a la hora de incorporar las variantes y variaciones de las UFS en este tipo de obras lexicográficas.

Según se ha podido observar, en los diccionarios *DCC1* y *DCC2* se emplean los paréntesis para incluir todo tipo de información. En cuanto a las variaciones, se utilizan para indicar que una UF puede ampliarse, lo que hemos denominado adición o supresión, como es el caso de *bon vent! (o bon vent i barca nova!)*; y para señalar las expresiones con casillas vacías, como sucede con *quin vent us (o et, etc.) porta per ací?*.

Por lo que se refiere a las variantes, también se usan los paréntesis para señalar los cambios léxicos, según puede observarse en *a cielo abierto (o descubierto)*; *amor (o estimat) meu!*; *perdre la xaveta (o el cap, o el seny)*; *caído (o bajado, o llovido) del cielo*; *levantarse (o moverse) viento*; *coger (o agarrar, o tocar) el cielo con las manos*; y *engegar a dida (o a passeig, o a la quinta forca)*; entre muchos otros. En alguna ocasión se emplean los corchetes para concretar el complemento circunstancial que forma la UF, como *beber los vientos por [una cosa, una mujer]*. Este criterio puede ser confuso, ya que el mismo diccionario incluye la información del significado entre corchetes: *al pie de la letra [exactamente]*.

Las modificaciones morfosintácticas también se incorporan entre paréntesis, tónica habitual de este diccionario. Se hallan casos relacionados con el número, como *ver el cielo abierto (o los cielos abiertos)* y *córrer la mar (o les mars)*, entre otros; con el cambio de alguna preposición, según se puede observar en *anar sobre (o contra) vent*; y con la estructura sintáctica *mar bonança (o mar calma, o bona mar)*. Cabe destacar que, en algunos casos, en vez de usarse el paréntesis para hacer referencia a cambios estructurales, se usan las comas, tal y como ocurre con *anar-se'n a fer punyetes*, *anar-se'n en orris*. En relación con esto, el mismo criterio, el de separar UFS con comas, se utiliza para añadir fraseologismos que se pueden considerar sinónimos: *tener viento de proa, hurtar el viento*; *a dojo, a doll, a gavadals*; *ir viento en popa, ir que chuta.*

Otro par de diccionarios bilingües en línea consultados son el *DMCC1* y el *DMCC2*, cuya forma de presentar la variabilidad fraseológica es prácticamente la misma que la de los anteriores. La única diferencia se halla en que en este caso no se emplean

corchetes. Los significados, sin embargo, al igual que los cambios léxicos, también se colocan entre paréntesis. La conjunción disyuntiva *o*, para estos últimos, es lo único que diferencia ambos contenidos: *coger (o agarrar) el cielo con las manos (estar muy enfadado)*, *treure foc pels queixals*.

Los diccionarios *DCE* y *DEC*, en contraposición con los anteriores, no usan marcas ortotipográficas para añadir la variabilidad de las expresiones que recogen. Aun así, incluyen las variantes y variaciones de algunos fraseologismos como entradas distintas. Esto sucede, por ejemplo, con las variantes morfosintácticas:

- *a pie juntillas – a peus junts*
- *a pies juntillas – a ulls clucs*

- *portar aigua al mar – llevar agua al mar · llevar leña al bosque · llevar leña al monte · vender miel al colmenero*

En este último caso, como se puede apreciar, los equivalentes en español se presentan separados, aunque algunos de ellos son variantes léxicas (*llevar leña al bosque – llevar leña al monte*).

Por último, los diccionarios *DAEC* y *DCEA* sí que añaden información sobre la variabilidad, aunque de manera escasa en comparación con las cuatro primeras obras analizadas. Por ejemplo, *tomar (a) alg(n) en brazos – prendre algú/algc en braços*. Como se puede apreciar, en la expresión española se añaden entre paréntesis las posibles variaciones referentes al complemento directo o indirecto que acompaña al verbo (*tomar a alguien* o *tomar algo en brazos*). En la parte catalana, el equivalente también incluye dicho apunte, pero en este caso se utiliza una barra oblicua. En tal sentido, el diccionario no es sistemático, ya que para presentar el mismo tipo de variación los recursos utilizados son distintos.

Finalmente, la barra oblicua se usa, también, para evidenciar los cambios morfosintácticos, como sucede en *a la/en cabeza de alg – al cap d'algc*; y, también, para señalar las variantes léxicas: *apretar/estrechar la mano a algn.*

4 Conclusiones

A pesar de que la variabilidad fraseológica esté bien consensuada desde la perspectiva científica, la forma citativa muestra divergencias en el contexto de la lexicografía. La ortotipografía no permite evidenciar las distinciones, sobre todo, entre variaciones (relacionadas con adición o supresión de elementos y con casillas vacías) y variantes (tanto léxicas como morfosintácticas). El predominio de los paréntesis y su combinación con comas, corchetes o barras oblicuas para indicar lo mismo, de igual modo que la escasez o ausencia de información relacionada con la variabilidad que se han localizado en los distintos diccionarios conllevan confusiones a los usuarios de este tipo de obras. Es evidente, pues, que resulta imprescindible seguir trabajando para conseguir la sistematización lexicográfica, no solo de lo que se vende impreso, sino de lo que permite su consulta en línea.

Referencias

Al mal tiempo, buena carrera, El Correo, https://www.elcorreo.com/alava/deportes/mas-deportes/201605/15/corredores-citan-maraton-martin-20160512233913.html, último acceso 2022/05/19.

García-Page, M.: Introducción a la fraseología española. Estudio de las locuciones. Anthropos, Barcelona (2008).

García Rodríguez, J.: FRAESCAT: propuesta de un diccionario electrónico de fraseología bilingüe español-catalán. Zeitschrift für romanische Philologie 137(2), 514-541 (2021).

Diccionari castellà-català (DCC1), diccionari.cat, https://www.diccionari.cat/diccionari-castella-catala, último acceso 2022/05/19.

Diccionari català-castellà (DCC2), diccionari.cat, https://diccionari.cat/diccionari-catala-castella, último acceso 2022/05/19.

Diccionari català-espanyol advanced (DCEA), dict.com, https://www.dict.com/catalan-espanol, último acceso 2022/05/19.

Diccionari Manual castellano-catalán (DMCC1), diccionaris.cat, https://www.diccionaris.cat/index.php, último acceso 2022/05/19.

Diccionari Manual català-castellà (DMCC2), diccionaris.cat, https://www.diccionaris.cat/index.php, último acceso 2022/05/19.

Diccionario avanzado español-catalán (DAEC), dict.com, https://www.dict.com/espanyol-catala, último acceso 2022/05/19.

Diccionario catalán-español (DCE), Glosbe, https://es.glosbe.com/ca/es, último acceso 2022/05/19.

Diccionario español-catalán (DEC), Glosbe, https://es.glosbe.com/es/ca, último acceso 2022/05/19.

Mellado Blanco, C.: Introducción: colocaciones y algunas cuestiones teórico-prácticas de fraseografía. En: Mellado Blanco, G. (ed.), Colocaciones y fraseología en los diccionarios, pp. 7-31. Peter Lang (Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation, vol. 44), Frankfurt (2008).

Mena Martínez, F.: En torno al concepto de desautomatización fraseológica: aspectos básicos. Tonos digital: Revista electrónica de estudios filológicos 5, 1-12 (2003).

Montoro del Arco, E. T.: La variación fraseológica y el diccionario. En: Battaner, P., DeCesaris, J. (eds.), De Lexicografia (Actes del I Symposium Internacional de Lexicografia), pp. 591-604. Barcelona: Institut Universitari de Lingüística Aplicada (2004).

Montoro del Arco, E. T.: Hacia una sistematización de la variabilidad fraseológica. En: Pastor Milán, M.ª A. (coord.), Estudios lingüísticos en recuerdo del profesor Juan Martínez Marín, pp. 125-152. Universidad de Granada, Granada (2005).

Montoro del Arco, E. T.: El concepto de 'locución' con casillas vacías. En: Mellado Blanco, G. (ed.), Colocaciones y fraseología en los diccionarios, pp. 131-146. Peter Lang (Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation, vol. 44), Frankfurt (2008).

Ortega Ojeda, G. & González Aguiar, M.ª I.: La técnica fraseográfica: el 'DRAE'-2001 frente al 'DEA'-1999. En: Mellado Blanco, G. (ed.), Colocaciones y fraseología en los diccionarios, 233-246. Peter Lang (Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation, vol. 44), Frankfurt (2008).

Wotjak, B.: Unidades fraseológicas en un diccionario de aprendizaje del alemán como lengua extranjera. En: Wotjak, G. (ed.), Estudios de fraseología y fraseografía del español actual, pp. 343-364. Vervuert/Iberoamericana, Frankfurt am Main (1998).

Wotjak, G.: ¿Cómo tratar las unidades fraseológicas (UF) en el diccionario? En: Wotjak, G. (ed.), Estudios de fraseología y fraseografía del español actual, pp. 307-322. Vervuert/Iberoamericana, Frankfurt am Main (1998).

Wotjak, G.: Las lenguas, ventanas que dan al mundo. Servicio de Publicaciones, Salamanca (2006).
Wotjak, G.: Acerca del potencial combinatorio de las UL: procedimientos escenogenésicos y preferencias sintagmático-colocacionales. En: Mellado Blanco, G. (ed.), Colocaciones y fraseología en los diccionarios, pp. 193-210. Peter Lang (Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation, vol. 44), Frankfurt (2008).
Wotjak, G.: Acerca del potencial comunicativo de las unidades fraseológicas idiomáticas y no-idiomáticas. En: Cabré i Castellví, M. T. (ed.), Lèxic, corpus i diccionaris (Cicle de conferències 96-97), pp. 155-180. Institut Universitari de Lingüística Aplicada, Barcelona (2010).
Wotjak, B., Wotjak, G.: La teoría del campo y otras propuestas clasificadoras para la fraseografía. En: Durante, V., Fraseología y paremiología: enfoques y aplicaciones, pp. 51-78. Centro Virtual Cervantes, Biblioteca fraseológica y paremiológica (Serie Monografías, 5), Madrid (2014).

A Phraseology Approach in Developmental Education Placement*

Miguel Da Corte[1][0000−0001−8782−8377] and Jorge Baptista[1,2][0000−0003−4603−4364]

[1] University of Algarve, Faculty of Human and Social Sciences
mlloveradacorte@gmail.com
[2] INESC-ID Lisboa, Human Language Technology Lab
jbaptis@ualg.pt

Abstract. This study focuses on an automatic classification task aiming at placing community college students into the appropriate level (Level 1 and 2) of Developmental Education (DevEd) courses, according to their English L1 proficiency. DevEd courses are designed to remediate and support students' communication skills in reading and writing before they can fully participate in college-level or college-bearing courses. This paper uses machine-learning methods to investigate the impact of considering multiword expressions (MWE) as entire tokens on the automatic classification task. Since many MWE are often non-compositional in meaning and constitute a large percentage of the textual units of many texts, they are likely to have a relevant role in the data representation of texts and, hence, improve subsequent classification task. Information is scarce regarding the tokenization of MWE and how this affects automatic placement. To this end, a random, balanced corpus of 186 sample texts (93 from each level) was used. Experiments compared the performance of a set of classifiers on the plain text corpus and on a version of the same corpus annotated for MWE. Results showed that using MWE as lexical features improved the classification accuracy by 8.1% above the baseline.

Keywords: Developmental education · machine-learning classification · multiword expressions · text mining.

1 Introduction and Objectives

The study of MWE continues to provide an opportunity for the enhancement of linguistic analysis. Several authors consider the lexical variety and repetition of MWE to be key in the process of language acquisition and mastery of language proficiency and fluency [9, 12, 15]. Understanding the lexical and syntactical patterns of community college students paves the way to understanding

* Research for this paper has been supported by University of Algarve, Language Sciences doctoral program, and by national funds through Fundação para a Ciência e a Tecnologia (Proj.Ref. UIDB/50021/2021).

issues related to language that impede effective communication both verbally and in writing. For community college students, the target population of this study, the use of these "prepackaged," "bundled" expressions facilitates communication and provides a sense of fluency, even if their writing skills require improvement to enter higher education. This is one of the main purposes of Development Education (DevEd) [4, 5, 16]. On the other hand, placement in DevEd courses often resorts to machine-learning based systems, such as ACCUPLACER[3], COMPASS[4], ACT[5], which use textual linguistic features to assist the placement strategies followed by higher education institutions. The COH-METRIX [10] [6] has been used to provide descriptive textual features that help capture Text Complexity and Readability, e.g., word count, paragraph length, word length; Text Easability; Referential Cohesion; Lexical Diversity; Connectives, Syntactic Complexity; Syntactic Pattern Density; Word Information; and Readability metrics (Flesch-Kincaid). These categories can be used as items for a linguistic and discourse representation to enrich corpora with linguistic data as evidenced in a previous study [1]. However, there is scarce information on how these systems deal with MWE (if at all), or at least what types of MWE are considered. The prevalent use of MWE in students' writing samples from entrance exams motivates the study of the impact of phraseology on machine-learning classification tasks pertaining to DevEd course placement, thus, succinctly defining the focus of this paper.

2 Related Work

Vocabulary in discourse and the connection of words within the larger discourse has been examined by [14], who particularly focused on how meaning is carried through *formulaic sequences*. Particular emphasis is placed on the formation of multiword units as a whole and common categories of these units such as compound words, phrasal verbs, fixed phrases, and lexical phrases are exemplified. The author asserts [14, p.101] that "language production stems from the precept that native speakers tend to use language that is formulaic in nature."

Understanding common categories of MWE requires a deep examination of their properties and the impact of these properties on NLP applications. According to [8], collocation and discontiguity phenomena are heavily exploited features, among the many properties MWE present, that have aided parsing tasks and, thus, enhanced syntactic analysis [3]. Studying the impact of discriminated tokens versus *single syntactic constituents*, as coined by [8], on the structural and functional aspects of developing writers' skills can aid in a more accurate placement of students in DevEd [8].

[3] `https://www.accuplacer.org/` (Last access: July 12, 2022; all URL in this paper were check on this date.)
[4] `https://www.compassprep.com/practice-tests/`
[5] `https://www.act.org`
[6] `http://141.225.61.35/CohMetrix2017/`

The incidence of multiword expressions was studied by [11] in 746 argumentative essays with 121,638 tokens from Korean-university students with the goal of exposing reocurrent structural sentence patterns and their frequency. These patterns were compared to those exhibited by American-university students in two large corpora (LOCNESS; MICUSP), where the incidence of phrased-based expressions exhibited by L2 compared to the incidence of noun phrases by L1 emerged as a theme. Additionally, L2 student showed a wider variety of MWE in their writing, suggesting that the use of these lexically-bound expressions facilitates communication among developing writers by adding a level of functionality to the writing process. Even though this study focused on L1 and L2 students, it supports the investigation of "MWE across proficiency levels or novice and professional academic writing [...] in an academic context." [11, p.11]

The ongoing variability of MWE, particularly verbal ones, and the challenges it poses for automated machine learning identification tasks are recognized by [13]. The authors focused on strategies to improve the identification of verbal MWE (VMWE) and used the PARSEME[7] corpora as a starting point. They completed three tasks comprised of a training and development phase, a prediction phase, and an evaluation phase all aided by a simple [candidate VMWE potential] extraction of filtering (*Seen*2020) techniques for precision [13]. Promising results were evidenced in boosting the identification of global MWE by using a combination of morphosyntactic filters. Suggestions for further work emphasize the representation of "VMWE as multisets of (lemma-POS) pairs rather than lemma and POS multisets separately." [13, p.3342]

A framework to further investigate the lexical and syntactical features of MWE is provided by [8, 11, 13, 14], among others. However, it is unclear whether automatic placement classification systems, i.e., ACCUPLACER, COMPASS, and ACT, take MWE into consideration or at least what types of MWE are considered. To the best of our knowledge, these systems provide a holistic score of writing samples based on the purpose and focus of the essay; organization and structure; development and support of ideas; sentence variety and style; accurate use of Standard Written English; and how one communicates and connects ideas while addressing a topic [2]. However, [16] emphasizes the need to improve college readiness by more accurately assessing students' performance in writing tasks. We aim at detecting MWE and MWE types to aide in the assessment of lexical features towards language proficiency.

3 Methods

To assess the impact of MWE in our classification task, a small corpus of 186 essays was collected from community college students, in the State of Oklahoma, U.S.A., placed in DevEd after completing their writing placement exam during years 2020-2021. The sample texts, consisting of a short multiparagraph essay of 300-600 words, prompted by a short excerpt with guiding questions, were

[7] `https://typo.uni-konstanz.de/parseme/`

retrieved from the ACCUPLACER platform of the higher education institution where this study took place. The corpus was then balanced for placement level (93 essays from each level: Level 1 and Level 2), and for sample size (keeping only up to 100 tokens per text). All the essays addressed the same topics.

Two experiments were devised, where the Classification Accuracy (CA) performance of a set of classifiers of the raw corpus and on a version of the same corpus annotated for all types of MWE was compared. We first devised a classification criteria for identifying MWE. Then, we independently annotated the corpus and compared the annotations in order to validate the selected MWE. The ORANGE[8] data-mining toolkit [6] was chosen for analysis and modeling since it provides, in a practical way, several NLP tools (mainly for preprocessing and data representation) within the same machine-learning (ML) toolkit and a large set of ML algorithms and data visualization tools.

3.1 Corpus Annotation and Classification Guidelines

In order for the Orange data-mining toolkit to tokenize a MWE as a single token, its elements had to be joined in the input text (an underscore was used to this end). This task was done manually. One of the reasons for this has to do with the large number of typos and spelling errors found in the essays, which we could only ignore with a manual inspection of the text. For the manual identification of MWE, the set of linguistic criteria adopted derived from the literature [3]. The MWE categories of [7] and the caveats by [8] were particularly insightful. Based on these criteria, the annotation of MWE in the corpus included multiple categories, based on traditional classification guidelines and lists available in dictionaries, e.g., **compound nouns**: *golden rule*, *refresher course*; **compound adjectives**: (stay) *tall and straight*, (be) *o*n [one's] *feet*; *strong minded*; **compound conjunctions and prepositions**: *as long as*, *in spite of*; **compound adverbs**: *by the way*, *back in the old day*; **verbal idioms**: *see the bigger picture*, *knock at your door*; *weigh the pros and cons*; **phrasal verbs**: *fit in*, *mess up*; **support verb constructions**: *have no clue*, *make mistakes*; among others. In the case of **nouns**, these also included named entities, both person names (*Colin Powell*), locations (*Guatemala City*), and organizations (*United Nations*).

For the annotation proper, the MWE were joined following these guidelines: Sequential strings of words were just joined by an underscore (*phone_ call*). Productive inserts, as in phrasal verbs, were marked by a double underscore and the inserted word permuted (*bring you down* → *bring_ _ down you*). Long-distance elements of the same MWE were joined in the first element, and the slot left void marked by a hashtag '#': The more *success <sic> you are in your job* the more *you will make* → *The_ more_ the_ more success you are in your job # you will make*. Prepositions introducing complements of predicative elements (adjectives, nouns, and verbs, including phrasal verbs) are not taken into account, e.g. *to get_ out of something* where *of something* is just a complement of the phrasal verb *get_ out* (joined). At the end, 1,260 MWE were annotated.

[8] `https://orangedatamining.com/`

3.2 Data Processing and Modeling

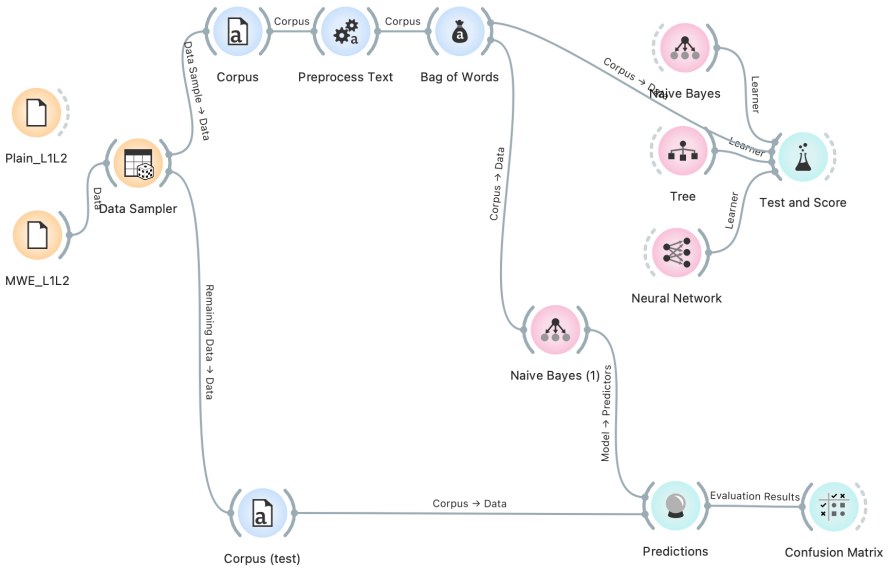


Fig. 1. Orange Text Mining Environment

The workflow adopted for this study is shown in Fig. 1. The DATASAMPLER widget was used to partition the data for a 3-fold cross-validation, leaving 2/3 for training and 1/3 for testing purposes. In the PREPROCESSING stage, basic tokenization options were selected, capturing both words and punctuation as tokens. Experiments were then carried out introducing different PART-OF-SPEECH (PoS) taggers. The data representation chosen was BAG-OF-WORDS (using the count of term frequency). For the training step, and to assess the models, the TEST&SCORE widget was used. The models used were Naive Bayes, Decision Tree (forward pruning) and Neural Network (multi-layer perceptron with back propagation). The data representation and the learning models were chosen because they had already proven to perform well with this type of data in previous experiments. For the PREDICTIONS (testing) step, the best performing model in each configuration setting was assessed.

3.3 Experiments

Table 1 shows the experimental settings and the breakdown of the results. In these experiments, we compared the non-annotated corpus (PlainText) with the corpus where MWE had been joined (Text-MWE). Since the corpus is small, the training partition corresponds to 2/3 of the corpus and the remaining is

Table 1. Experimental settings and results.

Experiment	Description	NB	NN	TR	best model	
Baseline	PlainText;Tok:W&P. noPoS.BoW	**0.685**	0.629	0.669	NB	0.742
Experiment 01	PlainText; idem+PoS:TB-ME	0.669	**0.718**	0.669	NN	0.677
Experiment 02	PlainText; idem+PoS:AP	0.613	0.661	**0.685**	TR	0.726
Experiment 03	Text-MWE;Tok:W&P. noPoS.BoW	0.707	**0.740**	0.553	NN	**0.823**
Experiment 04	idem+PoS:TB-ME	0.573	0.661	**0.667**	TR	0.742
Experiment 05	idem+PoS:AP	0.540	0.637	**0.653**	TR	0.758

left for testing. The data representation mode was Bag-of-Words with a simple frequency count. Results are provided using the Classification Accuracy metrics (CA). The names of learning models selected for the study are shortened to NB (Naive Bayes), Neural Networks (NN) and Tree (TR). The values reported for each individual model concern the Test&Scoring (training) step. In the rightmost columns, the CA values in the testing step are reported to the best performing model in the previous training step. In the description of the experiments, the two PoS-taggers have been shortened: Treebank - Maximum Entropy (TB-ME) and Average Perceptron (AP). Otherwise, 'noPoS' indicates no PoS-tagging was applied.

We first defined a baseline, using a basic configuration for the preprocessing step. The Word&Punctuation tokenization option was chosen, because it keeps punctuation as is, which might be a factor of placement classification. This is due to the fact that the use of punctuation is one of the writing skills not yet entirely mastered by the students and which is a topic that is addressed by the DevEd courses. No difference was found when the models were trained with different combinations of other tokenization configurations, namely just using white-space or just the words. No PoS-tagger was used in this baseline. For Experiments 1-2, the same baseline configuration was used but now adding each PoS-tagger in turn. For Experiment 3, the initial baseline configuration was enhanced by using the same corpus but with its MWE joined as single tokens. No PoS-tagger was used here. In Experiments 4-5, the same enhanced configuration was used, with joined MWE, but now adding each PoS-tagger.

4 Results and Discussion

Overall, the results from the experiments showed that adding MWE information to single-word tokenized texts improves the classification up to a clear 8.1% above the baseline (Experiment 03). The preprocessing tokenization options available with Orange did not seem to affect results too much. On the other hand, PoS-tagging underperforms the baseline, especially the AP (Experiment 02). With the MWE annotated text, however, AP performs better (Experiment 05) than TB-ME (Experiment 04). When one compares the results from the best performing model in the training step with its results in the testing step, the testing step results usually outperform the training, except in Experiment 01, where testing

results were 4.1% less than training. Otherwise, differences vary from 4.1% (Experiment 02) to 10.5% (Experiment 05). While using the plain text corpus in the training step, and though results are not very different from the baseline, no model proved to outperform the other two, as each in turn occurs as the best model. Using the text with MWE, however, showed that Neural Networks was the best model (Experiment 03), but that Tree outperformed the other two models when PoS-tagging was added to the configuration.

Although the corpus is small, the size of MWE found in it corresponds to 6.75% of the words in the corpus, a non-neglegible percentage. Still, results seem to signal the importance of using information on MWE in this task. This is in line with the literature findings on similar scenarios [7]. A proper PoS-tagging of MWE, combined with an appropriate classification of MWE types [9] may contribute to improving results and providing more robust insights on DevEd students' writing patterns.

5 Conclusion and Next Steps

This study focused on an automatic classification task aimed at placing community college students into the appropriate level (Level 1 and 2) of Developmental Education (DevEd) courses, according to their English L1 proficiency. We investigated and presented experimental results on the impact that multiword expressions (MWE), considered as entire tokens, have on an automatic classification task using machine-learning methods. The classification was based on linguistic features derived from small corpus of texts written by native English speakers at the onset of their higher education journey. Issues of orthography, punctuation, lexical usage such as slang, grammatical issues such as agreement, semantic issues, and discursive aspects of the text have been previously explored. Now, the inclusion of MWE provided promising results as they constitute a large percentage of the textual units of many texts and are likely to have a relevant role in the representation of the information in texts. An extended corpus will be required to confirm the hypothesis presented in this preliminary study.

A quantitative evaluation to verify the usefulness of including MWE information was presented and discussed, including the benefits of using different PoS-taggers and the degree to which these tools affect the classification accuracy. Although, at this point in the research study, we are not concerned with the automatic identification of MWE in DevEd students' writing, we are committed to developing, first, a more precise MWE scheme for annotating writing samples to further gain insights into linguistic issues that prevent effective communication for this student population.

[9] The list of MWE is available at: `https://www.researchgate.net/project/Linguistic-Aspects-of-Developmental-Education`

References

1. Abba, K.A.: Community college students' writing: Lexical, syntactic, and cohesion differences in L1, L2, and Generation 1.5 students and examining knowledge of the writing process. Ph.D. thesis, Texas A&M University, Graduate and Professional Studies (2015)
2. Board, C.: ACCUPLACER program manual. The College Board New York (2018)
3. Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Multiword Expression Processing: A Survey. Computational Linguistics **43**(4), 837–892 (12 2017). https://doi.org/10.1162/COLI_a_00302, https://doi.org/10.1162/COLI_a_00302
4. Cormier, M., Bickerstaff, S.: Research on developmental education instruction for adult literacy learners. The Wiley Handbook of Adult Literacy pp. 541–561 (2019)
5. Darkenwald-DeCola, J.A.: 'In College, I'm the One People Go To': Lessons from Successful Developmental Literacy Students About the Transition to College-Level Courses Across Disciplines. Ph.D. thesis, Rutgers The State University of New Jersey, School of Graduate Studies (2021)
6. Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., Zupan, B.: Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research **14**, 2349–2353 (2013), http://jmlr.org/papers/v14/demsar13a.html
7. Kochmar, E., Gooding, S., Shardlow, M.: Detecting multiword expression type helps lexical complexity assessment. arXiv preprint arXiv:2005.05692 (2020)
8. Laporte, E.: Choosing features for classifying multiword expressions. In: Sailer, M., Markantonatou, S. (eds.) Multiword expressions: In-sights from a multilingual perspective, pp. 143–186. Language Science Press, Berlin (2018). https://doi.org/10.5281/zenodo.1182597
9. Martinez, R.: A framework for the inclusion of multi-word expressions in elt. ELT journal **67**(2), 184–198 (2013)
10. McNamara, D.S., Ozuru, Y., Graesser, A.C., Louwerse, M.: Validating CoH-Metrix. In: Proceedings of the 28th annual Conference of the Cognitive Science Society. pp. 573–578 (2006)
11. Nam, D., Park, K.: *I will write about*: Investigating multiword expressions in prospective students' argumentative writing. Plos one **15**(12), e0242843 (2020)
12. Omidian, T., Shahriari, H., Ghonsooly, B.: Evaluating the Pedagogic Value of Multi-word Expressions based on EFL Teachers' and Advanced Learners' Value Judgments. TESOL Journal **8**(2), 489–511 (2017)
13. Pasquer, C., Savary, A., Ramisch, C., Antoine, J.Y.: Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 3333–3345 (2020)
14. Schmitt, N., Schmitt, D.: Vocabulary in language teaching. Cambridge University Press (2020)
15. Thomson, H.: Building speaking fluency with multiword expressions. TESL Canada Journal **34**(3), 26–53 (2017)
16. Zachry Rutschow, E., Edgecombe, N., Bickerstaff, S.: A brief history of developmental education reform (Oct 2021), https://postsecondaryreadiness.org/research/history-developmental-education-reform/

The German equivalenceless construction Prep + Sub + sein in Slovak[1]

Peter Ďurčo[1][0000-0002-7538-3689] and Anita Braxatorisová[2][0000-0002-9502-2725]

[1] University of Ss. Cyril and Methodius in Trnava, Trnava 917 01, Slovak Republic
[2] University of Ss. Cyril and Methodius in Trnava, Trnava 917 01, Slovak Republic
peter.durco@ucm.sk
anita.braxatorisova@ucm.sk

Abstract. The study focuses on contrastive research on fixed prepositional structures in German von + $Subst_{abstr}$ + sein, which have no direct parallels in Slovak. In the translation process, therefore, we are faced with questions of their contextual anchoring, the distribution of adequate functional equivalents, as well as reversibility in the search for the initial German construction. The aim of the study is to find out by means of parallel corpus analysis what equivalent constructions are used in the given cases in the translation of texts and subsequently to reveal patterns in translation preferences on the bases of analysis and interpretation of empirical data.

Keywords: Equivalenceless Constructions, Prepositional Phrases, Translation Equivalence, Parallel Corpus Analysis.

1 Introduction

Nowadays, the study of multi-word expressions (MWE) and their filling with lexical components is a very intensively developing area of research at the interface of various theoretical approaches between (construction) grammar, lexicology and phraseology (cf. Dobrovol'skij and Steyer 2018, Ďurčo 2020, Fraštíková 2022). Of particular interest from a contrastive point of view are those fixed constructions that have no direct parallels in the target language and therefore have to be represented by different constructions. According to Dobrovol'skij and Steyer (2018), the question of their contextual inclusion and distribution of adequate functional correspondences then comes to the fore. The question of reversible consideration, i.e. how the extensional semantics of equivalents can be applied in a reversible way to the German initial construction, is also quite legitimate. In fact, it is quite reasonable to expect that the extensional and intentional semantics of equivalents do not coincide either, resulting in a multivergent network of equivalence relations (cf. Ďurčo 2018a, 2018b). The typical highly frequented are cases like *von Beruf sein* (to be by profession), *von Bestand sein* (to endure), *von Dauer sein*" (be permanent), *von Interesse sein*" (to be of interest), *von Nöten sein*" (to

[1] The article was written within the project of the Slovak grant agency VEGA 1/0352/20 Confrontational research on lexicalization of construction models in German and Slovak.

be of needs), *von Nutzen sein*" (to be of use), *von Relevanz sein*" (to be of relevance), *von Vorteil sein*" (to be of benefit), *von Wichtigkeit sein*" (to be of importance) etc. In this paper, we analyze the German predicative construction with abstract nouns of the type "von + Subst + sein"[2] with the general meaning "to be of significance". In this study we discuss the cases of *von Bedeutung sein* (to be of significance) and *von Belang sein*" (to be of relevance). In this construction model, two basic readings can be recognized, the first with possessive semantics (cf. Chomová 2011): something has X (noun), and qualitative semantics (cf. Nábělková 1993, Sokolová 2003, Jarošová 2008): something is X (adjective).

2 Research object and approach

The study focuses on the empirical analysis of translation equivalents to the specific construction in German without constructional analogy in Slovak. The aim is to reveal the variability of the translations. The sources used for data extraction were the German-Slovak parallel corpus par-skde-all-3.0-de in the Slovak National Corpus[3], as a second source, the deTenTen13 corpus in Sketch Engine. For the MWEs *von Bedeutung / Belang sein* and their variants *nicht von Bedeutung / Belang sein*, *von + Attribute + Bedeutung / Belang sein* studied in this paper, structural solidity with the zero article between preposition and noun is typical. Functioning with the same structural and semantic features are also several other articleless prepositional MWEs with the general meaning to be of importance: *von Interesse, von Relevanz, von Wichtigkeit sein*, which are to be regarded as partial synonyms to the MWEs we selected as representative examples of this semantic group. These MWEs are by structure dative objects, so-called analytical Vonverbindungen (Lohstoeter 1931) with the meaning of the qualitative genitive (cf. Lipavic Oštir 1998, Laing 1920, Ružička et al. 1966, Helbig and Buscha, 2001). According to Schröder's (Schröder 1986, 203) semantic classification of prepositions, these constructions belong to the modal meaning of the preposition[4] *von*, indicating a remarkable property of an object/person (ibid.). These concrete constructions have no structural analogy in Slovak[5]. It is precisely for this reason that we want to explore the Slovak translation possibilities in this study. From a structural point of view, both in Slovak (Ružička et al. 1966, 169) and in German these meanings can be expressed with the help of an instrumental object: *(otázka) s veľkým významom - (Frage) mit großer Bedeutung* (question) with great importance), but these are little used in both languages. In German, the dative construction *mit großer Bedeutung* occurs contrary

[2] On the problem of the autonomy status of PWVs with zero article, see Hornáček Banášová (2019).

[3] The German - Slovak parallel-corpus in Slovak National Corpus comprises 238 137 893 tokens, 162 506 078 words, 15 637 775 sentences, 199 175 documents (belles-lettres and texts of the european parliament).

[4] For a deeper description of modal PWVs, see Hornáček Banášová (2018).

[5] Concerning the lack of equivalence and the general problem of searching for equivalence of prepositional word compounds, see further: Ďurčo (2018a, 2018b, 2020), Fraštíková (2018, 2019, 2022), Schillová (2021).

the analytical genitive *von großer Bedeutung* (75,215 hits) much rarely. In the German Web 2013 (deTenTen13) there are only 597 hits.

Data collection

The occurrence of prepositional MWEs with the structure: von + $Subst_{abstr}$ + sein with the meaning "to be significant" in the SNK is:

MWE	Hits	Searchingmethod
von Bedeutung	31 085 (130,53 per million)	Searching of Phrase: von Bedeutung or CQL: [word="von"][word="Bedeutung"] within <s/>
von + attr. + Bedeutung	10 739 (45,10 per million) + 638 (2, 68 per million)	[word="von"][][word="Bedeutung"] within <s/> [word="von"][]{2}[word="Bedeutung"] within <s/>
nicht von + Bedeutung	113 hits (0,47 per million)	[word="nicht"][word="von"][]{1}[word="Bedeutung"] within <s/> 31 (0,13 per milion) [word="nicht"][word="von"][]{2}[word="Belang"] within <s/> 1 hit (0 per milion)
von Belang	931 hits (3,91 per million)	Searching of Phrase: von Belang or CQL: [word="von"][word="Belang"] within <s/>
von +attr. + Belang	35 hits (0,15 per million)	CQL:[word="von"][]{1}[word="Belang"] within <s/> and [word="von"][]{2}[word="Belang"] within <s/> the variant [word="von"][]{3}[word="Belang"] within <s/> without results
nicht von + Belang	95 hits (0,40 per milión)	[word="nicht"][word="von"][word="Belang"] within <s/> [word="nicht"][word="von"][]{1}[word="Belang"] within <s/> without results

3 Translation equivalence using the example of meaning *to be of importance*

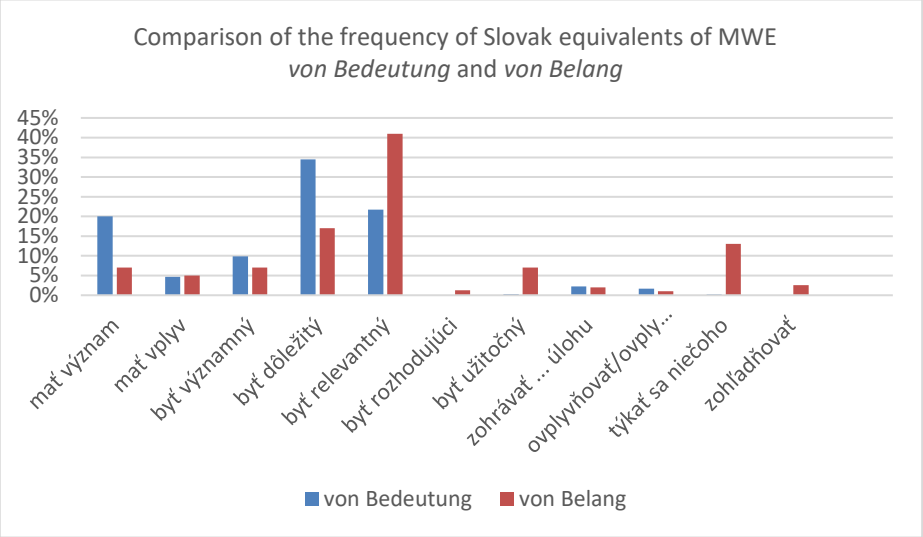
The primary analysis is oriented towards how often and in which contexts the expected prototypical equivalents "mať význam" (to have significance) and "byť významný" (to to be significant) occur, and whether certain regularities can be observed. The second step is to analyse the equivalence potential of this German construction in Slovak, i.e.

to collect data on the translation strategies for these equivalenceless construction in the target language.

I. Construction von + Sub$_{abstr}$ + sein

a) *von + Bedeutung + sein*
The analysis showed that the prototypical equivalents *mať význam* (to have significance) and *byť významný* (to be significant) are very often competed by other constructions.
The first group is formed by equivalents with possessive semantics: *mať rozhodujúcu úlohu* (to have a decisive role), *mať vplyv* (to have influence), *zohrať / zohrávať (významnú / dôležitú) úlohu* (to play a (significant/important) role), *byť výrazným príspevkom k niečomu* (to be a significant contribution to something), *byť dôležitým faktorom / byť nezanedbateľným faktorom* (to be an important factor / to be a non-negligible factor).
The second group of equivalents with qualitative semantics uses adjectives from the synonymic paradigm to the term important/significant: *byť dôležitý / ukázať sa ako dôležitý* (to be important / to appear important), *byť relevantný* (to be relevant), *byť pozoruhodný* (to be remarkable), *byť zaujímavý* (to be interesing), *byť užitočný* (to be useful).
A special case are the translations of the construction by autosemantic verbs with the general semantics "to stand in a relation to something, to be in relation to something", such as *týkať sa niečoho* (to relate to something), *znamenať niečo* (to mean something), *ovplyvňovať niečo* (to influence something), *zohľadňovať niečo* (to take account of something), *pomáhať niečomu* (to help something), *závisieť od niečoho* (to depend on something).

b) *von Belang sein*
Analogically differentiated equivalents are present also by this MWE. The first prototypical equivalent has possessive semantics: *mať význam* (to have significance); the second prototypical equivalent with qualitative semantics is expressed by cases such as: *byť relevantný pre niečo* (to be relevant to something), *byť dôležitý pre niečo* (to be important for something), *byť užitočný na niečo* (to be useful for something), *byť významný pre niečo* (to be significant for something), *zohrávať úlohu* (to play a role), *mať vplyv* (to have influence).
The second type of equivalence is realized by autosemantic verbs. The most frequent verbs are: *ovplyvniť niečo* (to influence something), *týkať sa niečoho* (to relate to something), *uvažovať o niečom* (to consider something), *zohľadňovať niečo* (to account for something).
The third type of equivalents uses qualitative adjectives with the general meaning "significant": *byť relevantný* (to be relevant), *byť rozhodujúci* (to be decisive).
The last type of equivalents are verbo-nominal phrases: *predstavovať dôležitý faktor v niečom* (to be an important factor in something), *byť dôležitým činiteľom pre niečo* (to be an important factor for something).

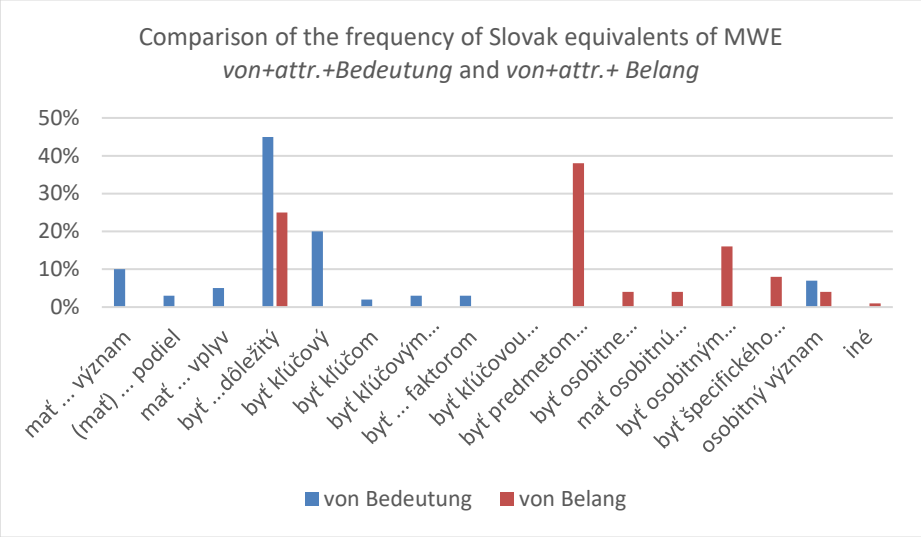


II. Construction von + Attribute + Sub$_{abstr}$ + sein

a) *von + Attribute + Bedeutung + sein*
This construction is very productive in German (299,466 hits in the corpus deTenTen13). In the attributive slot there occur numerous adjectives of strengthening/mitigating the meaning of the noun (*von allergrößter* (of utmost) / *von größter* (greatest) / *geringer* (less) / *geringerer* (lesser) / *höchster* (highest) *Bedeutung* (importance)), adjectives of qualification (*von allgemeiner* (of general) / *europäischer* (European) / *gemeinschaftlicher* (common) / *internationaler* (international) / *nationaler* (national) / *politischer* (political) / *strategischer* (strategic) / *systemischer* (systemic) / *unionsweiter* (Union-wide) / *untergeordneter* (subordinate) / *weltweiter Bedeutung* (worldwide importance), adjectives of intensification (*von außerordentlicher* (of extraordinary) / *ausschlaggebender* (crucial) / *besonderer* (special) / *beträchtlicher* (considerable) / *entscheidender* (decisive) / *erheblicher* (substantial) / *fundamentaler* (fundamental) / *grundlegender* (basic) / *grundsätzlicher* (fundamental) / *herausragender* (outstanding) / *maßgeblicher* (authoritative) / *überragender* (paramount) / *vorrangiger* (prior) / *wesentlicher* (essential) / *zentraler* Bedeutung (central importance)).
The analysis of the equivalence relations of the construction with the most frequent attribute *entscheidend* (decisive) has shown, firstly, that possessive semantics (= something has meaning) is much less represented than qualifying (= something is significant) and this is realised by various lexical and syntactic means. The second most important finding is that the attributive slot is translated by a wide range of synonyms, like *mať rozhodujúci / kľúčový / určujúci / zásadný význam / podiel / úlohu / vplyv* (to have a decisive / key / determining / essential importance / share / role / influence), *hrať / zohrávať dôležitú / kľúčovú / významnú úlohu* (to play an important / key / significant role), *byť (bytostne / veľmi / vysoko / mimoriadne / rozhodujúcim spôsobom / zásadne) dôležitý* (to be (essentially / very / highly / extremely / crucially / essentially) important,

byť kľúčový / byť kľúčom / kľúčovým prvkom / faktorom / kľúčovou zložkou (to be key / to be the key / the key element / the key factor / the key ingredient).

b) *von + Attribut + Belang sein*
The typical adjectives at least with the frequency of 10 hits in deTenTen13 in this construction model were the following: *groß* (large), *besonders* (specially), *gering* (small), *allgemein* (general), *wenig* (little), *öffentlich* (public), *erheblich* (considerable), *einig* (some), *hoch* (high), *keinerlei* (none), *sozial* (social), *höchst* (highly), *größer* (greater), *gesellschaftlich* (social), *öffentlich* (public), *persönlich* (personal), *international* (international), *finanziell* (financial), *technisch* (technical), *wirtschaftlich* (economic), *wesentlich* (substantial), *betrieblich* (operational), *gering* (minor), *weltlich* (mundane), *wirklich* (real), *gering* (low), *entscheidend* (decisive), *untergeordnet* (subordinate).
The bilingual corpus analysis of the MWE *von besonderem Belang sein* with the second most frequent adjective has brought wide range of equivalents with possessive and qualitative meaning, like *byť predmetom osobitného záujmu* (to be of particural interest), *byť osobitne zaujímavý* (to be especially interesting), *mať osobitnú dôležitosť* (to be of special importance), *byť osobitným záujmom niekoho* (to be of special interest to someone), *byť predmetom špecifického záujmu* (to be of specific interest), *mať osobitný význam* (to have a special significance).

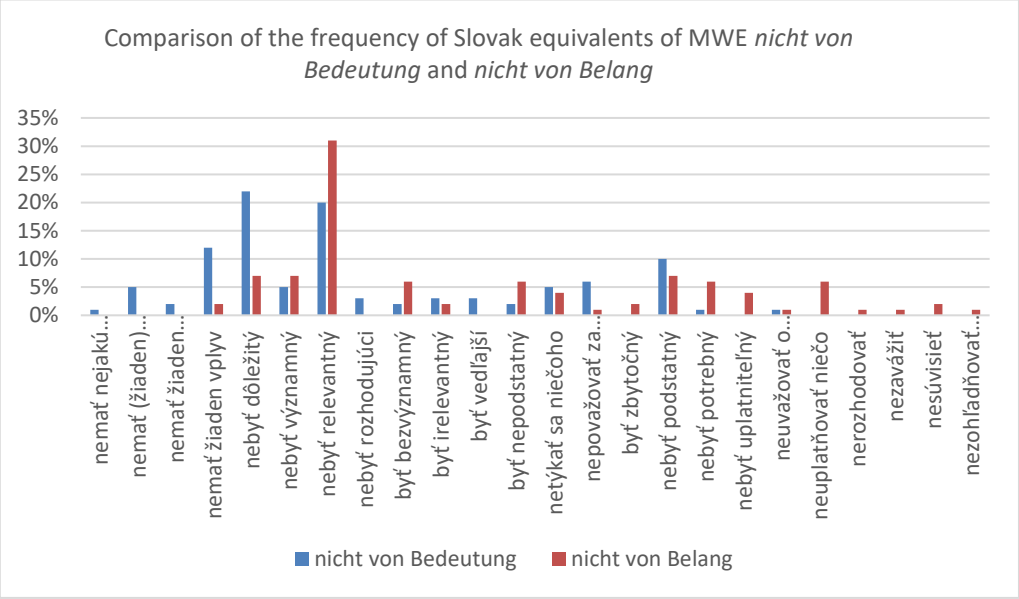


III. Construction nicht + von + Sub$_{abstr}$ + sein

a) *nicht von Bedeung sein*
The analysed construction with negation shows a wider range of possessive translation equivalents using verbo-nominal phrases: *nemať nejakú dôležitosť* (not to have any importance), *nemať (žiaden) význam* (to have no meaning), *nemať žiaden dosah* (to have no impact), *nemať žiaden vplyv* (to have no influence).

The qualitative semantics is expressed either with various synonymic adjectives: *dôležitý* (important), *významný* (significant), *relevantný* (relevant), *rozhodujúci* (deciding) with the verb *byť* (to be) in negative form *nebyť* (not to be), or by adjectives with negative semantics in the affirmative form of the verb *byť*: *bezvýznamný* (insignificant), *irelevantný* (irrelevant), *vedľajší* (side, subsidiary), *nepodstatný* (insignificant).
The third translation alternative is the use of negative verbs: *netýkať sa niečoho* (not to be related to something), *nepovažovať za niečo* (not to consider as something).

b) *nicht von Belang sein*
For this construction model is typical the qualitative semantics with predominant equivalents using adjectives with negative semantics: *byť zbytočný* (to be useless), *byť irelevantný* (to be irrelevant), *byť bezvýznamný* (to be insignificant), *byť nepodstatný* (to be unimportant).
Second type of equivalence uses negative form of the verb *byť – nebyť* (to be – not to be) with affirmative adjectives, like: *nebyť dôležitý* (not to be important), *nebyť podstatný* (not to be significant), *nebyť relevantný* (not to be relevant), *nebyť významný* (not to be significant), *nebyť potrebný* (not to be needed), *nebyť uplatniteľný* (not to be applicable).
This type uses also verbal equivalents with negative prefix: *neuvažovať o niečom* (not to consider something), *neuplatňovať niečo* (not to enforce something), *nerozhodovať* (not to decide), *nezávážiť* (do not matter), *netýkať* sa (not to concern), *nesúvisieť* (not to be related), *nezohladňovať niečo* (not to take account of something). The equivalent with possessive semantics *nemať žiadny vplyv* (to have no influence) is a marginal phenomenon.



4 Comparison

First of all, it should be noted that both units under comparison have a considerable common extensional range in their semantics, which is manifested in a wide variety of equivalents, and both analyzed MWEs show relatively high inclusion of their equivalents, which can be suprising when comparing very different semantics of the noun *Bedeutung* and *Belang*[6]. Only three meanings of *Bedeutung* and *Belang* are common: importance, interest, relevance. However, within a given construction, meanings that are common are obviously realized, and other meanings are not used in that construction. Common equivalents were indicated in the possessive as well in the qualitative meaning: *mať význam pre niečo* (to have relevance for something), *mať vplyv na niečo* (to have an impact on something), *mať veľký / kľúčový / podstatný / najpodstatnejší / rozhodujúci / zásadný význam* (to have major / key / essential / substantial / decisive / critical / essential importance), *byť významný* (to be significant), *byť dôležitý* (to be important), *byť relevantný* (to be relevant), *byť užitočný* (to be useful), *týkať sa niečoho* (to relate to something), *ovplyvniť niečo* (to influence something), *zohľadniť niečo* (to take into account something), *zohrávať významnú / dôležitú úlohu* (to play a significant/important role). This fact can be an argument for extensional semantic overlapping of compared MWEs.

Differences between both units in their equivalence were also indicated. Specific cases of *von Bedeutung sein* are: *znamenať niečo* (to mean something), *byť pozoruhodný* (to be remarkable), *byť zaujímavý* (to be interesting), *mať rozhodujúcu úlohu* (to have a decisive role).

Specific cases of *von Belang sein* are: *uvažovať o niečom* (to consider something), *zodpovedať niečomu* (to be relevant to something), *vplývať na niečo* (to affect something), *zohľadňovať niečo* (to account for something). These differences can be explained by the different semantic structure of the two polysemous nouns.

5 Conclusion

The present study focuses on the prepositional verbo-nominal construction *von + Bedeutung/Belang + sein"*, which has no analogous constructional counterpart in Slovak. Though the study is concentrated on its Slovak translation equivalents. The parallel corpus-based material analysis has shown that one has to reckon with a much wider range of translation strategies for equivalence-less constructions, both from a structural and lexical point of view. The German constructions *von + Bedeutung + sein"* and *von Belang sein"* corresponds to constructions with the auxiliary verb *haben"* in connection with nouns to the term *Wichtigkeit/Relevanz* (importance/relevance), predicative constructions with the auxiliary verb *sein"* in connection with adjectives to the term

[6] cf. Bedeutung: account, bearing, consequence, criticality, denotation, impact, Implication, importance, Intent, interest, meaning, prominence, relevance, repute, sense, seriousness, significance, signification, standing, weight [https://dict.leo.org/englisch-deutsch/Bedeutung]
Belang: concern, importance, interest, issues, matter, needs, procurement, relevance [https://dict.leo.org/englisch-deutsch/Belang]

wichtig/relevant (important/relevant). The third strategy is the translation by full verbs with the semantics “to stand in relation to something, to be in relation to something”.

References

1. Chomová, A.: Subjektivizácia vo vyjadrení posesívnych vzťahov. In: Philologia Juvenila Bohemica Olomucensia 2, 196–206 (2011).
2. Dobrovol’skij, D., Steyer, K.: Не то чтобы X- Nicht dass X. Konvergenz und Divergenz eines produktiven Musters. In: Gautier, L., Modicom, P.-Y., Vinckel-Roisin, H. (eds.), pp. 93–107. De Gruyter, Berlin, Boston (2018). https://doi.org/10.1515/9783110585292
3. Ďurčo, P.: Faktoren der konvergenten und divergenten Äquivalenz von präpositionalen Wortverbindungen. In: Steyer, K. (ed.) Sprachliche Verfestigung. Wortverbindungen, Muster, Phrasem-Konstruktionen, pp. 285–306. Narr Francke Attempo, Tübingen (2018a).
4. Ďurčo, P.: Lexikalisierte PWVs aus kontrastiver Sicht. In: Hornáček Banášová, M., Fraštíková, S. (eds.) AKTUELLE FRAGEN UND TRENDS DER FORSCHUNG IN DER SLOWAKISCHEN GERMANISTIK III., pp. 9–59. Kirsch, Nümbrecht (2018b).
5. Ďurčo, P.: Ustálené predložkovo-menné spojenia v nemčine. Philologia 30(2), 177–189 (2020). Fraštíkova, S.: Die korpusbasierte Untersuchung der lokalen” Präposition-Substantiv-Verbindung am Telefon aus kontrastiver Sicht. In: Hornáček Banášová, M., Fraštíková, S. (eds.) AKTUELLE FRAGEN UND TRENDS DER FORSCHUNG IN DER SLOWAKISCHEN GERMANISTIK III., pp. 60–106. Kirsch, Nümbrecht (2018).
6. Fraštíkova, S.: Die Präposition-Nomen-Wortverbindung lokalen Charakters. Eine kontrastive korpusbasierte Untersuchung am Beispiel von außer Sicht. In: Ďurčo, P., Tabačeková, J. (eds.) Präposition-Nomen-Verbindungen. Korpusstudien zu Gebrauch und Musterhaftigkeit phraseologischer Minimaleinheiten, pp. 145–165. Logos Verlag, Berlin (2019).
7. Fraštíková, S.: Rekurrente präpositionale Wortverbindungen im lokalen Bereich. Logos Verlag Berlin (2022).
8. Helbig, G., Buscha, J.: Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. Langenscheidt, Berlin/München (2001).
9. Hornáček Banášová, M.: Präpositionale Wortverbindungen mit modaler Bedeutung. In: Hornáček Banášová, M., Fraštíková, S. (eds.) AKTUELLE FRAGEN UND TRENDS DER FORSCHUNG IN DER SLOWAKISCHEN GERMANISTIK III., pp. 147–171. Kirsch, Nümbrecht (2018).
10. Hornáček Banášová, M.: Zum Autonomie-Status der präpositionalen Wortverbindungen. In: Ďurčo, P., Tabačeková, J. (eds) PRÄPOSITION-NOMEN-VERBINDUNGEN. KORPUSSTUDIEN ZU GEBRAUCH UND MUSTERHAFTIGKEIT PHRASEOLOGISCHER MINIMALEINHEITEN, pp. 125–143. Berlin: Logos Verlag (2019).
11. Jarošová, A.: Spracovanie adjektív v Slovníku súčasného slovenského jazyka s osobitným zreteľom na adjektíva vzťahové. In: Rangelova, A., Světlá, J., Jarošová, A. (eds.) Lexikografie v kontextu informační společnosti, pp. 57 – 72. Praha: Ústav pro jazyk český AV ČR (2008).
12. Laing, G. J.: The genitive value in Latin and other constructions with verbs of rating. The University of Chicago Press, Chicago (1920).
13. Lipavic Oštir, A.: Analytischer und synthetischer Genitiv im Deutschen - ein diachronischer Vergleich. Linguistica 30(2), 87–113 (1998).

14. Lohstoeter, L. O.: Der Kampf zwischen dem Genitiv und der Präposition von". Monatshefte für Deutschen Unterricht 23(8), 251–253 (1931).
15. Nábělková, M.: Vzťahové adjektíva v slovenčine. Funkčno-sémantická analýza desubstantívnych derivátov). Bratislava: VEDA (1993).
16. Ružička et al.: Morfológia slovenského jazyka. Vydavateľstvo Slovenskej akadémie vied, Bratislava (1966).
17. Schillová, A.: On corpus-driven research of complex adverbial prepositions with spatial meaning in Czech. Jazykovedný časopis 72(2), 425–433 (2021).
18. Schröder, J.: Lexikon deutscher Präpositionen. VEB Verlag Enzyklopädie Leipzig, Leipzig (1986).
19. Sokolová, J.: Sémantika kvalifikačných adjektív. Nitra: Univerzita Konštantína Filozofa (2003).
20. Steyer, K.: Sprachliche Verfestigung: Wortverbindungen, Muster, Phrasem-Konstruktionen. Narr Francke Attempto Verlag, Tübingen (2018).

Translation of Collocations in Seasonal Letting Agreements: A Corpus-driven Study

Luis Carlos Marín Navarro[1][0000-0003-4377-2177]

[1] University of Malaga, Malaga 29010, Spain
lmarin@uma.es

Abstract. In recent years, technology has entered the daily workflow of translators and (to a lesser extent) interpreters. In particular, language technologies have also shaped the discipline and provided a new paradigm for research. In this paper we will present a corpus-driven protocol to uncover collocational equivalents in legal genres, with special reference to seasonal letting agreements. In addition to the challenges posed by legal jargon, anisomorphisms, rhetorical conventions and so forth, collocations appear particularly problematic in the translation of legal texts. This paper will present corpora as an efficient and inexpensive means to uncover equivalents, find inspiration or help selecting among different options, reproduce the linguistic and cultural context of the target text, etc.

Keywords: Corpus, Phraseology, Collocation, Legal Translation, Seasonal Letting Agreements.

1 Introduction

Corpus linguistics has revolutionised translation and interpreting practices nowadays. It has also informed recent research in Translation Studies. A fairly comprehensive and generally accepted definition of corpus is offered by Bowker and Pearson (2002: 9): "a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria". In this paper we define corpus as a collection of texts that have to be representative of a language or part of it, gathered in electronic format and according to a series of specific criteria (Corpas Pastor, 2001 and 2008).

The main uses of corpora in translation are the following: (a) analysis of translations, (b) terminology study for specialised translations, (c) creation of translation memories, (d) creation of glossaries for translation and interpreting students and professionals, (e) automation of translational processes, (f) evaluation of translation quality, and (g) finding and checking translation equivalents (including discourse features, alternative translation choices, rhetorical features, phraseology, preferred wordings, etc.), among others.

In the case of interpreting, corpora are increasingly being used to prepare a given job, mainly for terminology work, phraseology, and key concepts of the domain.

However, corpus-based preparation of interpreting services is still in its infancy (cf. Corpas Pastor, 2021)

This study presents a protocolised procedure to translate phraseological units of seasonal letting agreements (holiday rentals) with the aid of corpora. We will be mainly concerned about challenges posed by collocations.

With the rise of short-term rentals and the rise of house-sharing holiday rentals through portals such as AirBnB, there have been big changes in the Tourism and Accommodation letting (De la Encarnación, 2016). Holiday rentals are not regulated specifically in Spain; they follow the legislation for leases in general. All rental agreements in Spain are regulated by Autonomous Regional Governments (Comunidades Autónomas). For long term rental (longer than 2 months) the agreements are subject to the Urban Tenancy Act (Ley de Arrendamientos Urbanos 29/1994 and Ley de Arrendamientos Urbanos 4/2013). Autonomous Communities have introduced their own additional requirements to regulate short-term rental (for example, the Andalusian Decree 28/2016). Many of the phraseological, discoursal and rethorical features of the target text type at hand stem from the regulations for holiday rentals in Spain.

The remaining of the paper is organised as follows. The first part presents linguistic and extralinguistic aspects that govern the translation of legal texts. The second part advances a protocol for translating seasonal letting agreements is presented, with special reference to phraseology and collocations. Finally, concluding remarks are drawn in the last section of the paper.

2 Translating Agreements

A multitude of texts from different legal systems circulate in our daily lives, from summons to letters rogatory (in the public sphere), as well as contracts, wills, and deeds (in the private sphere). Legal translation implies law, understood as a system that regulates the day-to-day life of citizens. Gutiérrez Arcones (2015) states that law is reflected through another language different from the current one, to our daily life. Legal texts function as vehicles of communication used by the Administration of Justice between the officials that form the body or the Administration itself and the citizens (both individuals and legal entities). This section will provide a brief overview of contractual agreements, as well as the main competencies required for the translation of this type of documents.

2.1 Key Features

According to Alcaraz Varó et al. (2006: 233), “the contract is defined as the promise or set of promises, legally binding, that oblige each of the parties to fulfil certain obligations in exchange for obtaining certain rights”. Roughly speaking, the intention to create a legal relationship is the cornerstone of this textual genre, followed by the

mutual acceptance of the parties, i.e., the agreement and consent expressed and signed in the contract. Other additional characteristics reflect the contractual capacity (referring to the person who can formalise a contract) and the performance, i.e., what the parties will receive in exchange for the fulfilment of the contract (Alcaraz Varó et al., 2006: 235-236).

Following Cuñado and Gámez (2014), the contractual genre, within the framework of legal-socioeconomic translation, is the one that tops the list of texts most demanded for translation. For the most part, contracts are a hybrid genre, sharing the characteristics of a legal text, but also those of a socio-economic text. Moreover, they are a textual genre that responds to an (almost) fossilised structure, which allows the translator to speed up his translation work. Like all legal documents, contracts tend to be worded in a hermetic, sophisticated, archaic, obscure, and complex language. Whereas lexical and semantic features index the special/specialised status of contracts, there are other syntactic and stylistic characteristics that single out this type of documents. The following is just a summary of the most prototypical features at various levels (cf. Gutiérrez Álvarez, 2010; Alcaraz and Hughes, 2009):

(a) Terminology: this type of documents makes use of a particular vocabulary. On the one hand, there are specialised terms, named entities and multiword units that are considered to be register-specific (*subarrendar, Ley de Arrendamientos Urbanos 29/1994, parte contratante*). On the other hand, there are own and unique meanings of common language concepts (for example, *celebrar, practicar*), together with words and expressions from the general language. Other outstanding features are a high frequency of prepositional expressions, massive use of clichés, formulas, couplets, relexicalisations, latinisms, anglicisms, etc. Some examples follow: *a tenor de lo previsto, exprese y materialice, salvo disposición en contrario, ejecutoriar, ex novo, trust*, etc.

(b) Syntax: although legal texts present a somewhat convoluted syntax, contracts, being a document of public nature to be signed by two persons who are possibly lay in the world of law, the syntax is assumed to be simple, with an abundance of simple and compound sentences (coordination and juxtaposition) by means of predefined structures. A typical pattern is the reiteration of syntactic structures, as well as syntactic ambiguity, deontic modality, plenty of passive structures and anaphoric elements, and so forth (cf. Bayo Delgado, 2000).

(c) Style and other conventions: As a genre framed within legal-administrative texts, contracts are full of archaisms (*cualesquiera, el que incumpliere*) and textual conventions. There is plentiful use of defined grammatical forms, like the tendency to use verbs in the present (*exponen, acuerdan, comparecen...*), and in the future (*deberá abonar, será reparado...*); plenty of gerunds and complex nominals, etc. Other preferences observed are the use of numbers instead of the corresponding word (e.g., *5* instead of *cinco*, for example) and monetary words instead of symbols (e.g., *euro* instead of *€*), etc. Contracts also tend to exhibit substandard or atypical orthotypography. For instance, there is a pervasive use of capitals for emphasis, which adds to the use of verbs in present, as in: *REUNIDOS de una parte_____, y de otra_____, INTERVIENEN.* In this example of the contract heading the verbs (in capitals) highlight the section that follows.

2.2 Translation Competencies

According to PACTE (2003: 44), translation competencies refer to the knowledge and abilities required within the translation process in order to produce a translation product. In the case of legal texts, Borja Albi (2005) has identified two spheres of competencies. From the list below, non-linguistic (or extralinguistic) competencies encompass (a) to (e), while (f) to (n) are considered linguistic competencies:

(a) Encyclopaedic knowledge of the world.
(b) Theoretical knowledge of translation.
(c) Knowledge of the law of both the source language and the target language (comparative law).
(d) Knowledge of international law.
(e) Participation in the experiences of the legal community.
(f) Knowledge of the typology of texts in the target language.
(g) Knowledge of the taxonomy of texts in source language.
(h) Knowledge of comparative legal textology.
(i) Formal aspects required by the legislation in each legal system.
(j) Function and legal efficacy of the genres in each legal system.
(k) Macrostructure of the different legal genres.
(l) Formal and stylistic aspects.
(m) Phraseology characteristic of each type of text.
(n) Terminology specific to each genre.

The aforementioned competencies are interrelated and constitute an indispensable requirement for quality work, including the need to abide by target micro- and macrostructural features (conventions) when translating legal texts. This paper will be particularly relevant for linguistic competencies (m) and (n).

3 A Corpus-driven Protocol to Uncover Collocational Equivalents

Due to space limitations, this study will focus on the phraseology of seasonal letting agreements, with special reference to the translation of collocations. Our main aim is to propose a protocolised procedure to extract collocational equivalents from comparable corpora as a means to improve the linguistic competencies of practitioners and student translators.

3.1 Corpus Compilation

In order to identify collocations and their equivalents in the target language, we have compiled an ad hoc corpus of seasonal letting agreements (cf. Torruella and Llisterri, 1999; Corpas Pastor, 2001 and 2008). The result we obtain is an unannotated and comparable corpus composed of 20 seasonal letting agreements, which in turn

is made up of two subcorpus of 10 texts in Spanish and 10 in English with a total of 48,686 words. The texts have been retrieved from different websites related to seasonal letting, both in Spain and in the United Kingdom. The corpus has been compiled according to the following criteria, proposed

(a) Textual genre

Although we are looking for a model for the translation of seasonal or holiday rental agreements, the texts that will make up our corpus will be these types of contracts. As we have seen in previous sections, a contract is a legal document that may be of a public nature (between the Administration and the citizen, for example) or private (such as a sale agreement, for example). We will therefore use model contracts taken from websites with a certain degree of reliability and whose professional activity is closely linked to property rental or civil law in general.

(b) Language

We will focus on two languages: Spanish and English. To do so, we will look for a large number of texts in this linguistic combination. However, the restriction we have to take into account is the one we have already seen in the section on textual genre, i.e., the diatopic restriction. To this end, we will use texts written in European Spanish and British.

(c) Comparable or parallel corpus

There are some websites where texts in the source language and their translation can be found. However, the agreement, being a textual typology with a larger scale within private law, it will be very difficult to find texts with their corresponding translations. Therefore, our mission here is to compile a comparable corpus and to be able to see in each language the different predominant linguistic structures in order to be able to compare them with each other and to propose a translation template based on the corpus.

(d) Complete or partial texts

The quality of our translations will depend on whether or not we have fragmented the text. For this type of translation, it is very important not to leave out any section, as they are all closely related to each other.

In the previous section we have examined the different competencies that a legal translator must have. From here on, in order to control all these competencies, we will delve into the genre that concerns the present work, as well as in compiling a corpus that meets the needs of the professional and provides us with a template with which to overcome the numerous linguistic and cultural barriers that arise in legal language.

3.2 Description of the Corpus

The result we obtain is an unannotated and comparable corpus composed of 20 seasonal letting agreements, which in turn is made up of two subcorpus of 10 texts in Spanish and 10 in English with a total of 48,686 words.

3.3 Procedural Steps in the Identification of Collocation Translation Equivalences

Before discussing the methodology, it is necessary to outline one of the spheres of phraseology. For the purposes of this research we are interested in collocations, which according to Firth (1957: 181), "collocations of a given word are statements of the habitual or customary places of that word in collocational order".

For the search of functional equivalents in our foreign language (English) we followed the corpus-driven methodology proposed by Tognini-Bonelli (2001: 87). Following the author, "the corpus-driven approach, on the other hand, fits well the definition in that it aims to derive linguistic categories systematically from the recurrent patterns and the frequency distributions that emerge from language in context". The first step consists of identifying a term or expression in L1 (Spanish) to be translated into L2 (English) and looking for its phraseological pattern, i.e., collocates. Based on the phraseological pattern of the term in question, step 2 consists of identifying equivalents in L2 prima facie based on the phraseological pattern identified in the first step. Finally, the third and last step would lead us to see the range of phraseological units in L2 from the equivalent identified prima facie in the second step. The following table shows in a more graphic way the methodology used:

Table 1. Procedural steps in the identification of translation equivalence (Tognini-Bonelli, 2001: 135)

COMPARABLE CORPUS(L1)	TRANSLATION CORPUS/TRANLSATOR'S EXPERIENCE (SL/TL)	COMPARABLE CORPUS (L2)
Step 1 from Formal Patterning/L1 to Function(s)/L1	Step 2 identify a prima facie translation equivalent for each function ➔ Function/L2	Step 3 from Function/L2 (as realised by a translation equivalent) to Formal Patterning/L2

The term *contrato* [111] according to the following image, could be placed with *arrentamiento* or *alquiler*. The first step is to identify the problematic collocate and study its phraseological pattern. Thus, *arrendamiento* or *alquiler* could go together with *contrato*. The second step, according to Tognini-Bonelli, would try, from the range of collocatives in L1, to find the translation equivalents in L2 (English). Finally, the third step would lead us to see the collocatives in L2 thanks to the equivalents found prima facie. The same occurs with the term *vivienda* [66], that could be placed with the verb *amueblar* (*amueblada*). The following process shows clearly, and in a more practical way, the process followed:

8

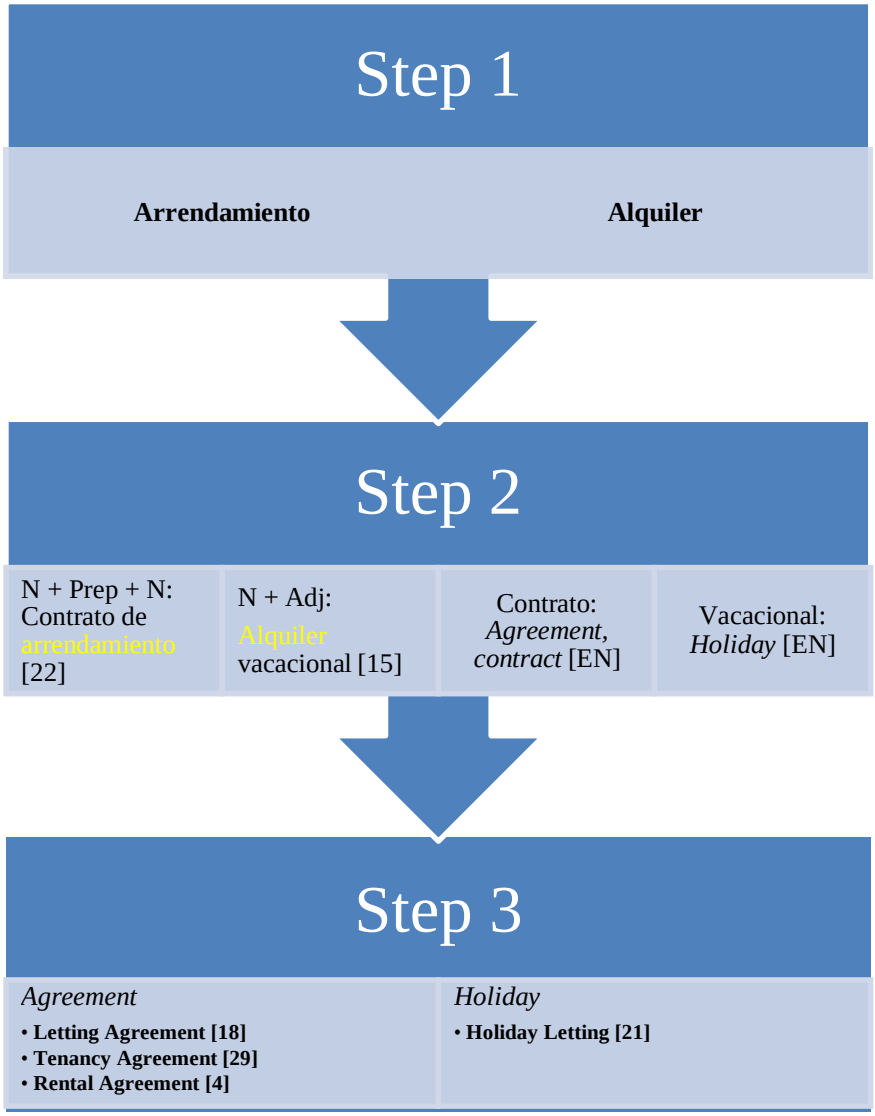


Fig. 1. Case I: *Arrendamiento/Alquiler*

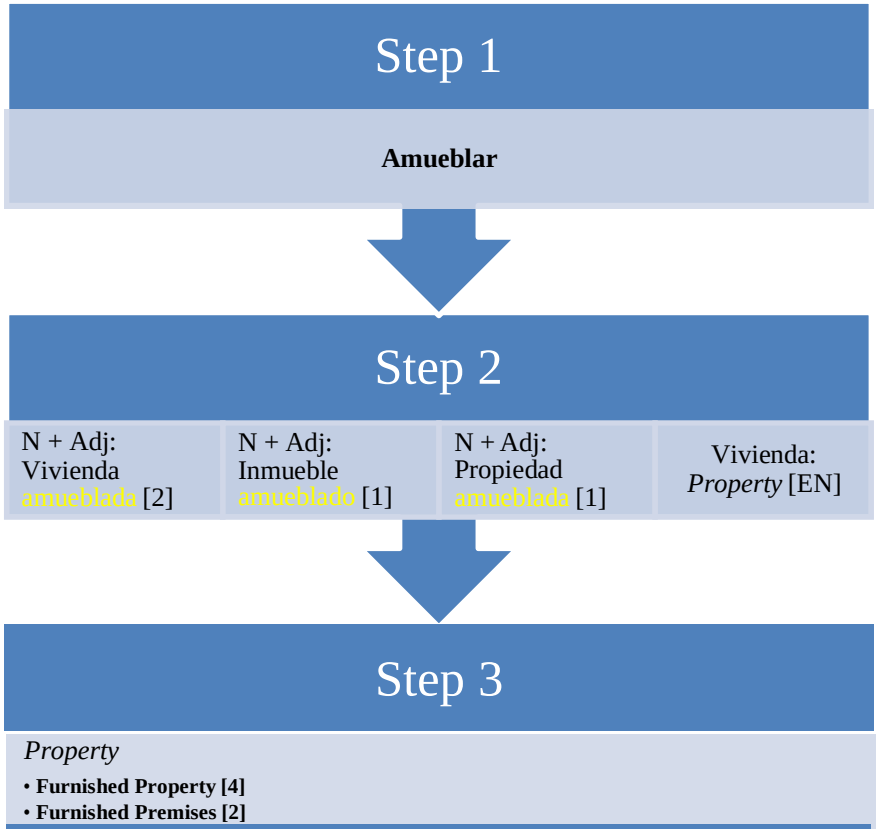


Fig. 2. Case II: *Amueblar*

However, it has been observed in the corpus that there are collocations in Spanish thatdo not have a correspondence in English, so their equivalent is a single term. The following list shows a list of collocations taken from the corpus together with their English equivalent, either a collocation or a single term.

Table 2. Equivalences

Collocation in L1	**Equivalences in L2 (collocation, if any)**	**Equivalences in L2 (single term, if any)**
Animal (N+Prep+N) Animal de compañía [2]		Pet [16] Animal [5]
Año (N+Prep+N) Año de construcción [1]	Year of construction [1]	
Antelación (Prep+N) Con antelación [8]	In advance [6]	

Causar (V+N) Causar molestias [1] Causar daños/deterioros [6]	Cause a disturbance [2] Cause a nuisance [3] Cause any damage [9] Arise a disturbance [1]	
Contrato (N+Prep+N) Objeto del contrato [12] Contrato de arrendamiento [22] **Contrato (V+N)** Rescindir el contrato [5] Finalizar el contrato [3] Otorgar el contrato [8]	Letting Agreement [18] Tenancy Agreement [29] Rental Agreement [4] Terminate the agreement [5] Cancellation of the agreement [40] Enter into the agreement [5]	
Deber (V+V)	Shall + V [147] Must + V [157]	
Fecha (N+N) Fecha de llegada [15] Fecha de salida [6]	Date of arrival [1] Arrival date [14] Departure day [2]	Arrival [9] Departure [6]
Importe (N+Adj) Importe total [2]	Total rental fee [2] Total rent payable [3]	
Incumplimiento (N+Prep+N) Incumplimiento de [8]	Breach of [23]	
Llaves (N+Prep+N) Entrega de llaves [5] Recepción de llaves [1]	Key collection [1]	
Pago (N+Prep+N) Forma de pago [12] **Pago (V+N)** Pactar el pago [2] Efectuar el pago [7] Pago total [1]	Rental payment [2]	Payment [19]
Prohibido (Adv+Adj) Expresamente prohibido [1] Estrictamente prohibido [1]	Strictly prohibited [7]	
Vacacional (N+Adj) Alquiler vacacional [15]	Holiday Letting [21]	
Vivienda (N+Adj) Vivienda amueblada [2]	Furnished Property [4] Furnished Premises [2]	

Thus, from the corpus we obtain the following results:

1. Out of the 14 groups of collocations in the table above, 11 out of 14 in L1 that have as equivalent another collocation in L2 (92%). 2 out of 14 have as equiv-

alent, in addition to a collocation (16.7%), a single term. Finally, only one group has no collocation as equivalent, but a single term (8.3%)

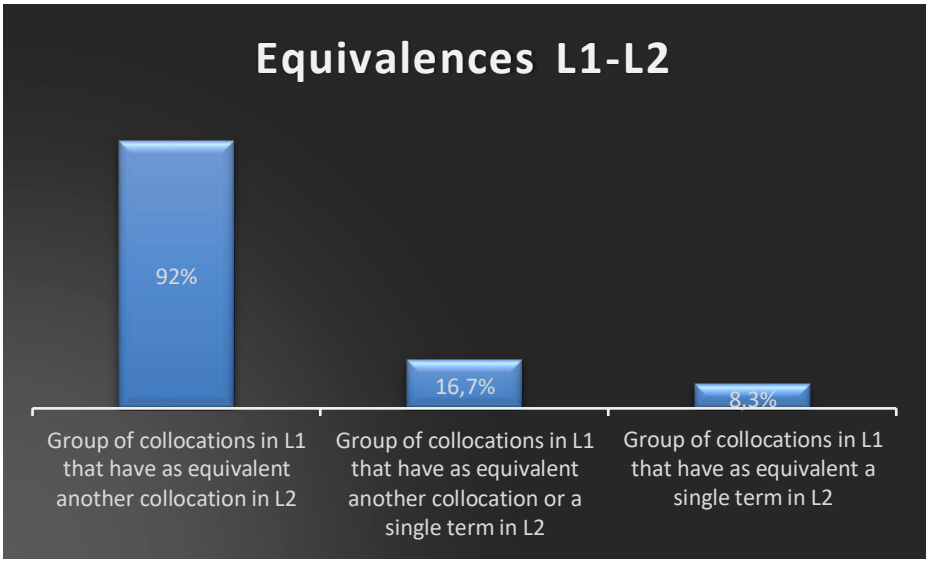


Fig. 3. Equivalences L1-L2

2. Among the 26 collocations observed in L2, 12 out of them maintain the same syntactic structure as in L1 (46%), while 14 do not follow the same syntactic pattern (54%).

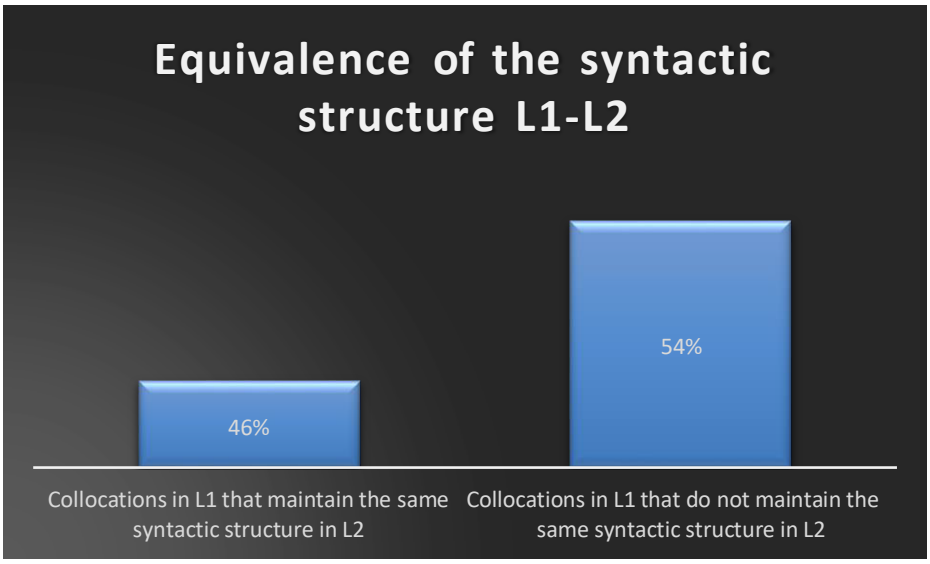


Fig. 4. Equivalence of the syntactic structure L1-L2

4 Conclusions

The quality of translation and interpreting work can be greatly reduced if accurate documentation is not carried out. As we have seen, this work is greatly enriched thanks to corpus methodology. The study of examples taken from real language helps to approach specialised languages in a more methodological, precise and intuitive way, and to study the subject matter as if we were professionals in that specialised field. Consequently, the work will gain in quality and fidelity.

This work has sought to demonstrate two aspects. Firstly, that through the use of corpora we can observe the use of the different constructions that characterise legal language. As we have observed in previous sections, the contractual genre responds to a fixed structure with very specific terminology. The main disadvantage is that of locating the collocations that shape seasonal letting agreements.

Thus, taking into account the objectives proposed at the beginning of this research and after observing the whole practical aspects of it, we can only affirm that linguistic technologies, and more specifically, the use of corpora in translation, represent a major advance, as well as a great incentive given their versatility. This will make it much easier for translating work to be almost automated, although we must always bear in mind that the translator is the human figure who must have the final say at all times.

References

1. Alcaraz Varó, E., Campos Pardillos, M.A. Miguélez, C.: El inglés jurídico norteamericano, 3rd edn. Ariel, Barcelona (2006).
2. Algar, R.: Collaborative consumption. Leisure Report, pp. 16-17. Retrieved from https://www.oxygen-consulting.co.uk/insights/collaborative-consumption/ (2007), last accessed 2021/10/01.
3. Baker, M.: Corpus Linguistics and Translation Studies: Implications and Applications. In Baker, M., Francis, G., Tognini-Bonelli, E. (eds.) Text and Technology: In Honour of John Sinclair, pp. 233-250. John Benjamins, Amsterdam/Philadelphia (1993).
4. Bayo Delgado, J..: El lenguaje forense: estructura y estilo. Estudios de derecho judicial 32, 35-76 (2000).
5. Bowker, L., Pearson, J.: Working with Specialized Language: A practical guide to using corpora. Routledge, London/New York (2002).
6. Cabré, M.T.: La terminología, representación y comunicación: elementos para una teoría de base comunicativa y otros artículos. Insitut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona (1999).
7. Corpas Pastor, G.: Compilación de un corpus ad hoc para la enseñanza de traducción inversa especializada. Trans 5, 155-184 (2001).
8. Corpas Pastor, G.: Investigar con corpus en traducción: los retos de un nuevo paradigma. Peter Lang, Frankfurt am Main (2008).
9. Corpas Pastor, G.: Technology Solutions for Interpreters: The VIP System. Hermēneus, Revista de Traducción e Interpretación 23, 91-123 (2021).
10. Cuñado, F., Gámez, R.: Aprendiendo sobre la traducción de contratos EN>ES. La linterna del traductor 9, 61-67 (2014).

11. De la Encarnación, A.M.: El alojamiento colaborativo: Viviendas de uso turístico y plataformas virtuales. REALA. Nueva Época 5, 30-55 (2016).
12. European Union.: EU consumer rules: The European Commission and EU consumer authorities push Airbnb to comply. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/IP_18_4453 (2018), last accessed 2022/01/25.
13. Firth, J.R.: Papers in linguistics, 1934-1951. Oxford University Press, London/New York (1957).
14. Gutiérrez Álvarez, J.M.: El español jurídico: propuesta didáctica orientada a la acción como base para un curso. MarcoELE: Revista de didáctica español como lengua extranjera 11, (2010).
15. Gutiérrez Arcones, D.: Estudios sobre el texto jurídico y su traducción: características de la traducción jurídica, jurada y judicial. Miscelánea Comillas, 73(142), 141-175 (2015).
16. Idealista: La patronal de los pisos turísticos acude a Europa para frenar las normas contra el alquiler vacacional. Retrieved from https://www.idealista.com/news/inmobiliario/vivienda/2016/11/24/744400-la-patronal-de-los-pisos-turisticos-acude-a-europa-para-frenar-las-normas-contra-el (2016), last accessed 2021/10/24.
17. McEnery, T., Wilson, A.: Corpus Linguistics. Edinburgh University Press, Edinburgh (1996).
18. PACTE: Building a Translation Competence Model. In Alves, F. (ed.) Triangulating Translation: Perspectives in process oriented research, pp. 43-66. John Benjamins, Amsterdam (2003).
19. Tognini-Bonelli, E.: Corpus Linguistics at Work. John Benjamins, Amsterdam/Philadelphia (2001).

BERT(s) to Detect Multiword Expressions

Damith Premasiri and Tharindu Ranasinghe

University of Wolverhampton, UK
{damith.premasiri,tharindu.ranasinghe}@wlv.ac.uk

Abstract. Multiword expressions (MWEs) present groups of words in which the meaning of the whole is not derived from the meaning of its parts. The task of processing MWEs is crucial in many natural language processing (NLP) applications, including machine translation and terminology extraction. Therefore, detecting MWEs is a popular research theme. In this paper, we explore state-of-the-art neural transformers in the task of detecting MWEs. We empirically evaluate several transformer models in the dataset for SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). We show that transformer models outperform the previous neural models based on long short-term memory (LSTM). The code and pre-trained model will be made freely available to the community.

Keywords: Multiword Expressions · Transformers · Deep Learning.

1 Introduction

The term "multiword expressions" (MWEs) denotes a group of words that act as a morphologic, syntactic and semantic unit in linguistic analysis; however, their meaning cannot be inferred from the meaning of their components [4]. For example, the MWE *"by and large"* have a meaning equivalent to *"on the whole"*. But none of the words in the MWE imply this [3]. MWEs can be categorised in to different categories such as lexicalised phrases and institutionalised phrases; however the basic definition remains same in all the categories. MWEs appear in almost all languages and is a common method of expressing ideas.

Apart from the difficulty of deriving meaning from individual components, which is known as non-compositionality in phraseology, MWEs have several challenges when processing them computationally [8]. 1. MWEs are non-substitutable, which means that the components of MWE cannot be replaced by synonyms (*e.g., by and big*). 2. MWEs and non-MWEs can be ambiguous (*e.g., by and large, we agree vs he walked by and large tractors passed him*). These unique challenges in MWEs raise several fundamental problems with many NLP applications. For example, parsing and machine translation (MT) [17, 16], which depends on a clear distinction between word tokens and phrases, has to be rethought to accommodate MWEs [8, 29]. The usual approach in these applications is to identify MWEs first, and then treat them accordingly. Therefore, detecting MWEs is a key research area in NLP.

In recent years, the identification of MWEs has been modelled as a supervised machine learning task where the machine learning models are trained on an annotated dataset. As we explain in Section 2, several datasets have been released to train these machine learning models. Furthermore shared tasks such as SemEval-2016 Task 10 [28] and PARSEME [27] have contributed to develop datasets. In recent years, neural network-based models, and in particular architectures incorporating Recurrent Neural Networks (RNNs) such as Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNNs) have achieved state-of-the-art performance in MWE identification tasks [27]. Usually, these models utilise pre-trained word embedding models such as word2vec [15] and glove [22]. We describe these models in Section 2. However, these traditional word embeddings provide the same embedding for polysemous words [21] [20]. Therefore, non-substitutability and the ambiguous nature of the MWEs can cause complications with traditional word embeddings.

A possible solution is to utilise neural architectures such as transformers that incorporate context more into the learning process. However, as far as we know, there has not been any research done to compare the performance of different transformer models in the MWE identification task. In this research, we empirically evaluate several transformer models in detecting MWEs to fill this gap. The findings of this research can be beneficial for many NLP applications that require detecting MWEs.

The main contributions of this study are,

1. We empirically evaluate eight different transformer models in the task of detecting multiword expressions using a recent dataset released for SemEval-2016 Task 10 [28].
2. We show that transformer-based models to identify multiword expressions outperform previous neural models based on LSTMs.
3. We provide important resources to the community: the code as an open-source framework, as well as the pre-trained models will be freely available to the community on HuggingFace [30] model hub[1]. Furthermore, we have created a docker image of the experiments adhering to the ACL reproducibility criteria[2].

2 Related Work

As mentioned before, a clear majority of the recent research to detect MWEs are neural based models. Usually the MWE detection is modelled as a token classification task where the model predicts whether a certain token belongs to a MWE or not. Therefore this task is similar to a named entity recognition

[1] The public GitHub repository is available on https://github.com/DamithDR/MultiwordExpressions and the pre-trained models are available on https://huggingface.co/Damith/mwe-xlm-roberta-base

[2] The docker image is available on https://hub.docker.com/r/damithpremasiri/transformer-based-mwe

(NER) task and the models that were used for NER has been used in MWE detection too [25].

The most popular method to detect MWEs are based on recurrent neural network variants such as LSTMs and gated recurrent units (GRUs) [25]. [18] use a LSTM model with Conditional random field (CRF) to detect MWEs. Furthermore, they incorporate dependency parse information to improve the results. Graph convolutional neural networks (GCNs) [13] have also been applied to MWE identification. [25] incorporate multi-head self-attention to improve the performance of GCN in MWE detection. Transformers have also been used to detect MWEs[5],[12]; however, the research has been limited to a few transformer models. Therefore, in this research, we fill this gap by empirically evaluating multiple transformers in the task of MWE identification.

3 Data

The dataset we used was from the 2016 SemEval shared task 10[3] [28]. The shared task was designed to predict both minimal semantic units and semantic classes (supersenses). The training data combines and harmonises three data-sets, the STREUSLE 2.1[4] corpus of web reviews, as well as the Ritter and Lowlands Twitter datasets[5]. The Ritter and Lowlands datasets have been reannotated for MWEs and supersenses to improve their quality and to more closely follow the conventions used in the STREUSLE annotations. The DiMSUM data files have tab-separated columns in the spirit of CoNLL, with blank lines to separate sentences. Each row contained nine columns : token offset, word, lowercase lemma, POS, MWE tag, offset of parent token (i.e. previous token in the same MWE), strength level encoded in the tag, supersense label and sentence ID. In this research we used only the word and the MWE tag. There are multiple MWE tagging formats such as IOB and IOB2. The dataset contains the IOB format where I - Inside, O - Outside, B - Beginning of a MWE. The I- prefix indicates that the tag is inside a chunk. An O indicates that a token belongs to no chunk. The B- prefix indicates that the tag is the beginning of a chunk that immediately follows another chunk without O tags between them.

The data composition is shown in the table 1. In the initial test dataset, there were 16500 words with 1000 sentences, however we had to remove one sentence from the test set due to encoding issues faced with the Python libraries.

4 Methodology

The main motivation behind the methodology is the state-of-art results produced by transformers in multiple different NLP tasks such as question answering [23],

[3] SemEval 2016 shared task description: http://dimsum16.github.io/

[4] The STREUSLE 2.1 is available on : http://www.cs.cmu.edu/ ark/LexSem/

[5] Twitter dataset is available on : https://github.com/coastalcph/supersense-data-twitter

Dataset	No of Words	No of Sentences
Train	73826	4800
Test	16400	999

Table 1. Datasets composition

machine translation quality estimation [24], cyber bullying [19] [26], language identification [11] and named entity recognition[2]. We experiment with two types of models, which we explain in the following sections.

Transformer models such as BERT [9] have been trained using masked language modelling objective and then can be fine-tuned for multiple different tasks [1]. This research uses the pre-trained transformer models for a token classification task. As shown in Figure 1, we added a token level classifier on top of the transformer model. The token-level classifier is a linear layer that takes the last hidden state of the sequence as the input and produces a label for each token as the output. In this case, each token can have three labels; B, I and O.

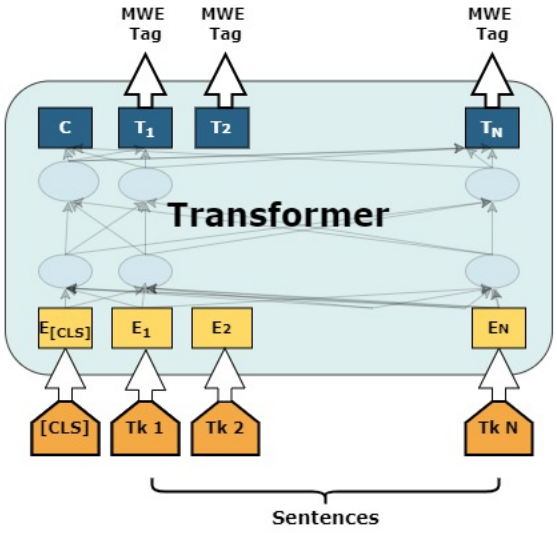


Fig. 1. Transformer Architecture for MWE classifier

We experimented with several popular, widely used transformer models to detect MWEs. Namely they are BERT [9], RoBERTa [32], XLNet [31], XLM-RoBERTa [7] and Electra [6]. For BERT we used several variations such as bert-base-cased, bert-base-uncased, bert-base-multilingual-cased and bert-base-multilingual-uncased while for other transformer model we only used the avail-

able base model. All the transformer-based methods were experimented using batch size 32, Adam optimiser with learning rate 4e-5. They were trained for 3 epochs with linear learning rate warm-up over 10% of the training data. These experiments were done in an NVIDIA GeForce RTX 2070 GPU.

BiLSTM-CRF is another token classification architecture which provided state-of-the-art results before transformers [10] . Bidirectional LSTM (BiLSTM) is capable of learning contextual information both forwards and backwards in time compared to conventional LSTMs. In this study, we used the Bi-LSTM architecture given this bidirectional ability to model temporal dependencies. CRFs [14] are a statistical model that are capable of incorporating context information and are highly used for sequence labeling tasks. A CRF connected to the top of the Bi-LSTM model provides a powerful way to model relationships between consecutive outputs (across time) and provides a means to efficiently utilise past and future tag information to predict the current tag of word. For the experiments, we used a learning rate 1.5e-1 and the model was trained for 50 epochs. BiLSTM-CRF experiments were conducted on a CPU.

5 Results

In this section, we report and compare the results of our experiments using standard evaluation metrics Weighted Recall, Weighted Precision, Weighted F1 and Macro F1 for the MWEs detection task. As shown in the Table 2 it is clear that the transformer-based models outperform the BiLSTM-CRF method with clear margins. The BiLSTM-CRF could achieve only 0.8253 and 0.3135 for Weighted F1 and Macro F1 scores, respectively, while all the transformer models we experimented outperform that. A clear observation is that even though BiLSTM-CRF has a fairly high Weighted F1 score, the Macro F1 score is very low. Since the Macro F1 score is sensitive to class imbalance, we hypothesise that this model is struggling to predict some specific label(s). On the other hand, transformer models achieve a high Macro F1 score suggesting that they can predict all the classes equally.

Results of transformers based neural methods have similar performance with slight differences from one model to another. It is clear that the best performer is the xlm-roberta-base model, which could achieve the best performance for both Weighted F1 and Macro F1 over all other models by achieving scores of 0.9169, 07366 accordingly. This is followed by the xlnet-base-cased model with a Macro F1 score of 0.7317, showing the competitiveness of the transformer models in MWE detection tasks. Interestingly, a multilingual model such as xlm-roberta-base could outperform language-specific transformer models on MWEs detecting task on this dataset.

Another interesting observation is that the cased models outperform the uncased models. This is similar in both bert and bert-multilingual models, where the cased models slightly outperform the uncased models. We believe that cased models can perform better in detecting MWEs than uncased models according to this dataset.

Model	Weighted Recall	Weighted Precision	Weighted F1	Macro F1
roberta-base	0.9125	0.9087	0.9103	0.7170
xlm-roberta-base	0.9194	0.9152	**0.9169**	**0.7366**
xlnet-base-cased	0.9179	0.9135	0.9152	0.7317
bert-base-multilingual-cased	0.9086	0.9042	0.9061	0.7049
bert-base-multilingual-uncased	0.9056	0.8990	0.9016	0.6863
bert-base-cased	0.9169	0.9122	0.9140	0.7290
bert-base-uncased	0.9087	0.9026	0.9050	0.6975
electra-base-discriminator	0.9125	0.9062	0.9085	0.7071
BiLSTM-CRF	0.8807	0.8059	0.8253	0.3135

Table 2. Results for different methods for multiword expression detection

Overall, transformers based neural methods perform higher than BiLSTM-CRF. It is clear that the results of all the transformer-based methods varied between 0.6863 - 0.7366 of Macro F1, showing their strong and competitive performance in MWE detection tasks.

6 Conclusion

MWE detection is an important research area for many NLP applications. In this paper, we empirically evaluate several neural transformer models in the MWE detection task using a recent dataset released for SemEval- 2016 Task 10 and show that all the transformer models outperform the LSTM based method. From the experimented transformer models, xlm-roberta-base provided the best results outperforming other transformer models. We can conclude that transformer models can handle the challenges presented by MWEs better than the previous LSTM based methods.

In the future, we would like to explore the cross-lingual capabilities of the transformer models in the MWE detection task. Cross-lingual transformer models such as xlm-roberta can be used to transfer knowledge between languages so that a model can be trained only on English data but can be used to predict on other languages. Since the xlm-roberta-based performed best in this study, we believe that this model can be further explored to detect MWEs in different languages.

References

1. Alloatti, F., Di Caro, L., Sportelli, G.: Real life application of a question answering system using BERT language model. In: Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue. pp. 250–253. Association for Computational Linguistics, Stockholm, Sweden (Sep 2019). https://doi.org/10.18653/v1/W19-5930, https://aclanthology.org/W19-5930

2. Arkhipov, M., Trofimova, M., Kuratov, Y., Sorokin, A.: Tuning multilingual transformers for language-specific named entity recognition. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. pp. 89–93. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-3712, https://aclanthology.org/W19-3712
3. Baldwin, T., Bannard, C., Tanaka, T., Widdows, D.: An empirical model of multiword expression decomposability. In: Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18. p. 89–96. MWE '03, Association for Computational Linguistics, USA (2003). https://doi.org/10.3115/1119282.1119294, https://doi.org/10.3115/1119282.1119294
4. Boros, T., Pipa, S., Barbu Mititelu, V., Tufis, D.: A data-driven approach to verbal multiword expression detection. PARSEME shared task system description paper. In: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017). pp. 121–126. Association for Computational Linguistics, Valencia, Spain (Apr 2017). https://doi.org/10.18653/v1/W17-1716, https://aclanthology.org/W17-1716
5. Chakraborty, S., Cougias, D., Piliero, S.: Identification of multiword expressions using transformers (2020)
6. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training text encoders as discriminators rather than generators. In: ICLR (2020), https://openreview.net/pdf?id=r1xMH1BtvB
7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747
8. Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Survey: Multiword expression processing: A Survey. Computational Linguistics **43**(4), 837–892 (Dec 2017). https://doi.org/10.1162/COLI_a_00302, https://aclanthology.org/J17-4005
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
10. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. ArXiv **abs/1508.01991** (2015)
11. Jauhiainen, T., Ranasinghe, T., Zampieri, M.: Comparing approaches to Dravidian language identification. In: Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects. pp. 120–127. Association for Computational Linguistics, Kiyv, Ukraine (Apr 2021), https://aclanthology.org/2021.vardial-1.14
12. Kanclerz, K., Piasecki, M.: Deep neural representations for multiword expressions detection. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. pp. 444–453 (2022)
13. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

14. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. p. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013), https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf
16. Mitkov, R.: Computational phraseology light: automatic translation of multiword expressions without translation resources. Yearbook of Phraseology **7**(1), 149–166 (2016)
17. Mitkov, R., Monti, J., Pastor, G.C., Seretan, V.: Multiword units in machine translation and translation technology, vol. 341. John Benjamins Publishing Company (2018)
18. Moreau, E., Alsulaimani, A., Maldonado, A., Vogel, C.: CRF-seq and CRF-DepTree at PARSEME shared task 2018: Detecting verbal MWEs using sequential and dependency-based approaches. In: Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018). pp. 241–247. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), https://aclanthology.org/W18-4926
19. Morgan, S., Ranasinghe, T., Zampieri, M.: WLV-RIT at GermEval 2021: Multitask learning with transformers to detect toxic, engaging, and fact-claiming comments. In: Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. pp. 32–38. Association for Computational Linguistics, Duesseldorf, Germany (Sep 2021), https://aclanthology.org/2021.germeval-1.5
20. Pathirana, N., Seneviratne, S., Samarawickrama, R., Wolff, S., Chitraranjan, C., Thayasivam, U., Ranasinghe, T.: Knowledge building via optimally clustered word embedding with hierarchical clustering. In: 15th International Conference on Natural Language Processing. p. 69 (2018)
21. Pathirana, N., Seneviratne, S., Samarawickrama, R., Wolff, S., Chitraranjan, C., Thayasivam, U., Ranasinghe, T.: Concept discovery through information extraction in restaurant domain. Computación y Sistemas **23**(3), 741–749 (2019)
22. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1162, https://aclanthology.org/D14-1162
23. Premasiri, D., Ranasinghe, T., Zaghouani, W., Mitkov, R.: Dtw at qur'an qa 2022: Utilising transfer learning with transformers for question answering in a low-resource domain. In: Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5). (2022)
24. Ranasinghe, T., Orasan, C., Mitkov, R.: An exploratory analysis of multilingual word-level quality estimation with cross-lingual transformers. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 434–440. Association for Computational

Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-short.55, https://aclanthology.org/2021.acl-short.55
25. Rohanian, O., Taslimipoor, S., Kouchaki, S., Ha, L.A., Mitkov, R.: Bridging the gap: Attending to discontinuity in identification of multiword expressions. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 2692–2698. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1275, https://aclanthology.org/N19-1275
26. Sarkar, D., Zampieri, M., Ranasinghe, T., Ororbia, A.: fBERT: A neural transformer for identifying offensive content. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 1792–1798. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.findings-emnlp.154, https://aclanthology.org/2021.findings-emnlp.154
27. Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., Doucet, A.: The PARSEME shared task on automatic identification of verbal multiword expressions. In: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017). pp. 31–47. Association for Computational Linguistics, Valencia, Spain (Apr 2017). https://doi.org/10.18653/v1/W17-1704, https://aclanthology.org/W17-1704
28. Schneider, N., Hovy, D., Johannsen, A., Carpuat, M.: SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 546–559. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/S16-1084, https://aclanthology.org/S16-1084
29. Taslimipoor, S., Mitkov, R., Pastor, G.C.: Using cross-lingual contexts to extract translation equivalents for multiword expressions from parallel corpora. In: Nuevos horizontes en los Estudios de Traducción e Interpretación (Comunicaciones completas): Conferencia AIETI7 (29 al 31 de enero de 2015 en Málaga). pp. 174–180. Editions Tradulex (2015)
30. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6
31. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf
32. Zhuang, L., Wayne, L., Ya, S., Jun, Z.: A robustly optimized BERT pre-training approach with post-training. In: Proceedings of the 20th Chinese National Conference on Computational Linguistics. pp. 1218–1227. Chinese Information Processing Society of China, Huhhot, China (Aug 2021), https://aclanthology.org/2021.ccl-1.108

Transformer-based Language models for the Identification of Idiomatic Expressions

Isuri Anuradha Nanomi Arachchige[1], Sachith Suraweera[2], and Dulip Herath[3]

[1] University of Wolverhampton, UK
[2] Informatics Institute of Technology, Sri Lanka
[3] University of Southern Queensland, Australia

Abstract. Idiomatic expressions are part of everyday speech in many languages and text genres. However, identifying idiomatic expressions can be problematic for different NLP tasks as their meaning cannot be easily inferred from their constituting words. Although the automatic identification and understanding of idiomatic expressions are essential for Natural Language Understanding tasks, they are still largely under-investigated. Until recently, the complex nature and lack of large datasets have prevented the development of machine learning approaches for identifying the idiomatic expressions.
With the advancement of machine learning techniques, transformer-based language models such as DistilBERT, RoBERTa, and their variants have shown state of the art performance by capturing the compositionality of the textual representations many NLP tasks. In spite of the progress, these vector representations fail to identify the multiword expressions (MWEs) such as idioms. This study demonstrate that transformer-based language models using contextual embeddings perform much better than existing approaches when applied to texts associated with depression. The core of this article presents two subtasks: (a) binary classification or idiomatic expression identification using BERT sentence embeddings and (b) task based on the analysis of the idiomatic expression usage in depressive textual contents.

Keywords: Idiomatic Expression · Language Models · Deep Learning

1 Introduction

An idiomatic expression is defined as a "constituent or series of constituents for which the semantic interpretation is not a compositional function of the formatives of which it is composed"[9]. Idiomaticity is a common linguistic feature of some research and studies in the areas of linguistics, psycholinguistics, developmental psychology, and neuropsychology [5].

Idioms pose challenges when performing tasks such as Word Sense Disambiguation (WSD), Semantic Role Labelling [6] and Semantic Parsing [4], Information Retrieval (IR), and Machine Translation (MT) [15, 2], Question Answering [18] and Text Summarisation. According to recent studies, very little

attention has been paid to identifying idiomatic expressions, i.e. multiword expressions (MWEs) with an established meaning unrelated to the meanings of the individual components. However, since idiomaticity is a common linguistic feature in all languages, idiomatic expressions play an important role in Natural Language Understanding. One of the main challenges in Natural Language Processing (NLP) is to embed the meaning of a piece of raw text (e.g. a word or a sentence) in a dense vector. With the advent of pre-trained language models, which exploit contextual information and assume the compositionality of word representations, significant improvements have been made in this direction [22, 7].

Although in the past, some statistical and rule-based models based on contexts have been used to classify idiomatic expressions, deep learning-based computational models have rarely been employed by researchers until very recently. Transformer is a prominent deep learning model which has been widely adopted in various fields, such as NLP, computer vision (CV) and speech processing. Though originally transformer was proposed as a sequence-to-sequence model for machine translation, later transformer based pre-trained models achieved state-of the-art performances in various downstream tasks such as text classification, question and answering name entity recognition. The transformer is a model based on deep neural network: it adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. As a recent advancement of transformer architecture, BERT is one of the most evident in its success [28, 7]. Some extensions of BERT model such as ALBERT, RoBERTa, DistilBERT [14] were proposed to address the drawbacks that exist in original BERT model. In this study we employ the original BERT model and some of its extensions to compare individual's performances.

The primary goal of research is to identify the use of idiomatic expressions and identify idiomatic sentences in texts describing states of depression in order to establish how people try to convey their thoughts using language when they are being depressed. Social stigma is a common social phenomenon [30] and try to hide their real feelings and emotions from others. As a result, they try to convey what is inside their minds using different language interpretations without directly expressing it to others[25]. Idiomatic expressions are a common linguistic feature used by them to their ideas and expressions. In order to find the usage of idiomatic expression in depressive contents, we fine-tune BERT models for depression specific domain datasets on depression and make use of human justification for added performance in our models. We determine that the BERT model is effective and significantly improves the performance of the above-discussed tasks.

The main contributions of this paper consist of:

- proposing predictive ensembling method to classify idiomatic expression from texts using BERT and extensions of the BERT model and
- adapting the transformer based models to identify how idiomatic expressions are used in posts which contains the phases or words related to depression disorder in DSM-V.

The rest of the paper is organised as follows: section 2 discusses the related work to the study, and section 3 describes the proposed transformer-based models and experimental setup for the classification of tokens as substrings of idiomatic expressions. Furthermore section 3 explains the results section while section 4 concludes the paper and offers suggestions for future work.

2 Related Works

With the appearance of the transformer-based architecture, the NLP domain witnessed the emergence of several pre-trained language models such as BERT and XLNet [19]. These pre-trained Language models proved good accuracy levels in both sentence-level [2] and token-level [18] techniques in different NLP tasks. However, Ensemble models are popular among the recent transformer-based architectural models.

Related Research has been reported automatically to classify idiomatic and literal expressions based on different datasets [21]. Another study was conducted on identifying the idiomatic expressions with the sentiment analysis [26]. Moreover, a neural network-based idiomatic expression recommendation process for essay writings was proposed by Liu [17]. Recently, with the advancement of transformer-based pre-trained architectural models, several studies have employed the BERT model to identify the idiomatic expressions which exists in the different contexts. Apart from those studies some have proposed a supervised learning approach to classify multiword expressions from literal ones using the VNC-Tokens (verb-noun construction) dataset [8].While different computational methods have been used so far, to the best of our knowledge ensemble approaches have not been employed for the identification of idiomatic expressions.

3 Methodology

In this section, we describe our pre-trained models and the experimental setup (Section 3.1) and the datasets we use to train and evaluate our idiomatic expression identification system (Section 3.2).

3.1 Datasets

The EPIE copus [24] was selected to train the different versions of the BERT model which contains the 25206 sentences labelled with lexical instance of 717 idiomatic expressions. The dataset is labelled according to the static and formal idioms.Table 1 shows the statistics of the EPIE dataset.

For the purpose of the testing the model we produced a data set by merging the following two datasets together.

– " SemEval All words" and " SemEval Lex sample" dataset contains 1143 sentences and 1423 sentences [15]. Following datasets were obtained by the

Table 1. Number of Sentences in the Dataset

Dataset	Train	Validation	Test	N
EPIE	15k	5k	2k	25206

of SemEval-2013 Task 5b Dataset. The dataset was created using entries available on Wiktionary under the label of English Idioms. Furthermore, sentences were annotated for figurative, literal, both or impossible to tell usage, using the CrowdFlower2 crowdsourcing annotation platform.
– PIE corpus [11] contains 2239 sentences with idiomatic expressions. The selected sentences were labelled according to the idiomatic usage ("y") and meaning literal usage ("n"). The PIE corpus created based on 4 different datasets (VNC-Tokens, Gigaword, IDIX and SemEval-2013 datasets).

Aligning to the contribution of the paper, depressive forum post dataset [19] were used for the evaluation of the models. The dataset was comprising forum posts collected from depression online support forums.

3.2 Pre-trained Transformer Models

The below mentioned pre-trained language models are applicable to many NLP tasks and performed well. Furthermore, they were employed in this study in order to build the ensemble model.

– BERT-Base
In 2018 BERT simple and powerful transformer based architecture was proposed with the purpose of train the bidirectional representations from the unlabeled dataset. BERT was able to deliver promising results on most of NLP tasks. BERT architecture uses a masked language model (MLM). In MLM's model, it makes the random words from input and then it predicts the ID of that word by utilising its context in both left and right directions contexts which enables training of the bidirectional model. BERT-base is affordable model which is comparatively smaller in its size and takes less computational time to process. The employment of the BERT model was able to give promising result on various NLP tasks [13, 27].
– DistilBERT
DistilBERT is extension of BERT model with the features of lighter, fast, smaller and cheap [23]. In DistilBERT, the size has reduced from 40% to 60% compared to the original BERT with more speed and 97% understanding of language capabilities. Most of research conducted based on the text classification and other NLP tasks [1] [12] have employed the DistilBERT model because more lighter than original BERT.
– ALBERT
Toyota technologies and Google research have jointly introduced the scalable and smaller successor of the original BERT model as the ALBERT [16]. In

the ALBERT model, two parameters were optimised to improve the training speed and to minimise the memory consumption of the original model. Since ALBERT has low number of parameters, according to recent studies been proved that ALBERT work well classification tasks [10, 29].

– Ensemble models
Noise, bias and variance are major causes to create errors in learning models. To overcome that problem Ensemble methods have been introduced to minimise the errors and to improve the stability and accuracy of machine learning algorithms [20]. Ensemble models consider meta-algorithms which combine several machine learning and deep learning classifiers into a predictive model to reduce the variance, and biases and improve accuracy.
From this study, we attempt to classify idiomatic expressions using the ensemble BERT-DistilBERT-ALBERT Combination. The above models are fine-tuned by training them with the EPIE dataset. The next section describes the proposed experimental setup for the classification of idiomatic expressions.

4 Experiments and Results

Due to the complex nature of the idiomatic expressions, it is quite obvious that the proposed model must have different aspects to precisely and accurately predict the idiomatic expressions from the depressive content. Fig 1 elaborates on the steps and architecture of the fine-tuned model.

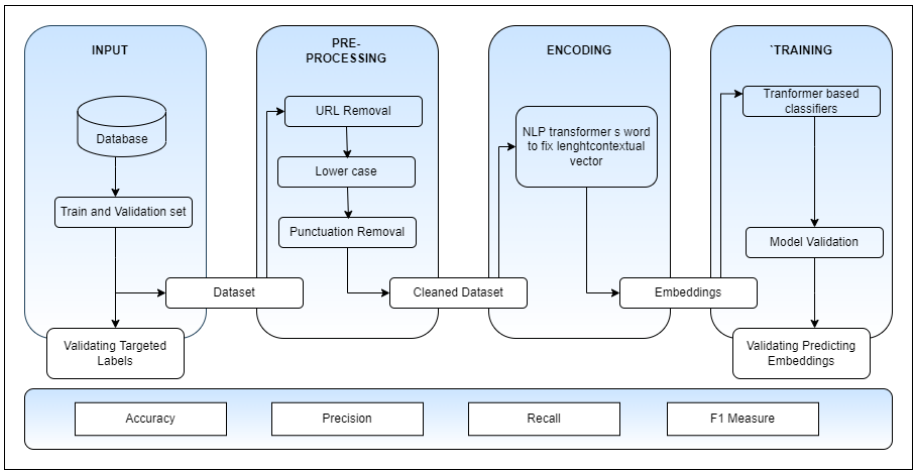


Fig. 1. Proposed experiment for the idiomatic expression identification

The pre-processing involves converting all upper case letters into lower case ones and removes all HTML tags, if any, though none was found as the data was

extracted manually and verified. Additionally, emojis and numbers were removed in the pre-processing stage. The training dataset is shuffled before training. The predictions of the three models are combined and the weighted average method is used to predict the final class labels according to the dataset.

Algorithm 1 Ensemble Learning Algorithm for Identify Idiomatic Sentences

i ← *Weights of the BERT-base Model (Bm)*
j ← *Weights of the ALBERT Model (ABm)*
k ← *Weights of the DistilBERT Model (Dm)*
L ← *All the sentences of the training set*
PBERT ← *Prediction from the BERT model range from 0 to 1*
PALBERT ← *Prediction from the ALBERT model range from 0 to 1*
PDistilBERT ← *Prediction from the DistilBERT model range from 0 to 1*

temp = 0
repeat
  repeat$x \in L$
    $temp \leftarrow$ (i * $PBERT_x$+j * $PALBERT_x$+k * $PDistilBERT_x$)

  until *all sentences in the dataset training set (L)*
  $final\_prediction \leftarrow temp/(i + j + k)$
  $output \leftarrow argmax(final\_prediction)$
until *apply for all the models*

In order to develop the ensemble model, we employed the grid search method to find the weights [3]. Grid search is able to identify a model's hyper-parameters, which results for accurate predictions. To calculate the ensemble model's predictions, the optimisation of simple averaging method that is known as the weighted average technique was used. In the last the predictions were combined from each model and take the average, often performing better, overall, than a single model. The weighted average in the ensemble is a machine learning strategy that aggregates predictions from several models, with each model's contribution weighted according to its competence or performance.

Given below are the results obtained from the experiments including the baseline models, BERT, ALBERT, DistilBERT. Moreover comparative analysis were preformed based on token level with two baseline models for the task of idiomatic expression classification. The models were evaluated using the Accuracy, Precision, Recall and F-score metrics.

In order to improve the performance and robustness we decide to use ensemble model over a baseline model. The proposed ensemble model is the reason to reduced variance and increased the overall results. The main reason is because the baseline models which the ensemble model was built, were fine-tuned with the three different datasets with idiomatic expressions and thus show robust results when tested with our own in depression text dataset with 5278 sentences.

Table 2. Results of the proposed models

Model	Accuracy	Precision	Recall	F measure
BERT-base	0.82	0.79	0.81	0.83
ALBERT	0.81	0.80	0.82	0.80
DistilBERT	0.83	0.79	0.77	0.84
Ensamble Model	0.84	0.81	0.79	0.85

Table 3 represent some test results obtained after applying the best model to the depressed text dataset.

Table 3. Results get from the ensemble model

Sentence No	Sentence	Idiomatic expressions
1	I hate getting so down in dumps to the point where my body completely shuts down, and all i can do is just lay there and think about all my problems.	Idiomatic sentence
2	I'm sorry for being annoying when I want to talk.. needy cause I miss you, emotional when I care and insecure because I'm afraid to lose you	Not Idiomatic sentence

5 Conclusion

This paper has presented an ensemble predictive model for the classification of idiomatic expressions in the depressive sentences combining two baseline models, BERT, ALBERT and DistilBERT. The proposed ensemble model was fine-tuned with the three datasets and tested with our own depression text dataset.

References

1. Abadeer, M.: Assessment of distilbert performance on named entity recognition task for the detection of protected health information and medical concepts. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. pp. 158–167 (2020)
2. Adewumi, T.P., Javed, S., Vadoodi, R., Tripathy, A., Nikolaidou, K., Liwicki, F., Liwicki, M.: Potential idiomatic expression (pie)-english: Corpus for classes of idioms. arXiv preprint arXiv:2105.03280 (2021)
3. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of machine learning research **13**(2) (2012)

4. Bevilacqua, M., Blloshmi, R., Navigli, R.: One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. In: Proceedings of AAAI (2021)
5. Cacciari, C.: The place of idioms in a literal and metaphorical world. Idioms: Processing, structure, and interpretation pp. 27–55 (1993)
6. Conia, S., Bacciu, A., Navigli, R.: Unifying cross-lingual semantic role labeling with heterogeneous linguistic resources. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 338–351 (2021)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Diab, M., Bhutada, P.: Verb noun construction mwe token classification. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009). pp. 17–22 (2009)
9. Fraser, B.: Idioms within a transformational grammar. Foundations of language pp. 22–42 (1970)
10. Gregory, H., Li, S., Mohammadi, P., Tarn, N., Draelos, R., Rudin, C.: A transformer approach to contextual sarcasm detection in twitter. In: Proceedings of the Second Workshop on Figurative Language Processing. pp. 270–275 (2020)
11. Haagsma, H., Nissim, M., Bos, J.: Casting a wide net: robust extraction of potentially idiomatic expressions. arXiv preprint arXiv:1911.08829 (2019)
12. Jayarao, P., Sharma, A.: Retraining distilbert for a voice shopping assistant by using universal dependencies. arXiv preprint arXiv:2103.15737 (2021)
13. Kaliyar, R.K., Goswami, A., Narang, P.: Fakebert: Fake news detection in social media with a bert-based deep learning approach. Multimedia tools and applications **80**(8), 11765–11788 (2021)
14. Kant, N., Puri, R., Yakovenko, N., Catanzaro, B.: Practical text classification with large pre-trained language models. arXiv preprint arXiv:1812.01207 (2018)
15. Korkontzelos, I., Zesch, T., Zanzotto, F.M., Biemann, C.: Semeval-2013 task 5: Evaluating phrasal semantics. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 39–47 (2013)
16. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
17. Liu, Y., Liu, B., Shan, L., Wang, X.: Modelling context with neural networks for recommending idioms in essay writing. Neurocomputing **275**, 2287–2293 (2018)
18. Mishra, A., Jain, S.K.: A survey on question answering systems with classification. Journal of King Saud University-Computer and Information Sciences **28**(3), 345–361 (2016)
19. Nanomi Arachchige, I.A., Jayasuriya, V.H., Weerasinghe, R.: A dataset for research on modelling depression severity in online forum data. In: Proceedings of the Student Research Workshop Associated with RANLP 2021. pp. 144–153. INCOMA Ltd., Online (Sep 2021), https://aclanthology.org/2021.ranlp-srw.20
20. Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of artificial intelligence research **11**, 169–198 (1999)
21. Peng, J., Feldman, A., Vylomova, E.: Classifying idiomatic and literal expressions using topic models and intensity of emotions. arXiv preprint arXiv:1802.09961 (2018)

22. Peters, M.E., Neumann, M., Zettlemoyer, L., Yih, W.t.: Dissecting contextual word embeddings: Architecture and representation. arXiv preprint arXiv:1808.08949 (2018)
23. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
24. Saxena, P., Paul, S.: Epie dataset: a corpus for possible idiomatic expressions. In: International Conference on Text, Speech, and Dialogue. pp. 87–94. Springer (2020)
25. Smith, R.A., Applegate, A.: Mental health stigma and communication and their intersections with education. Communication Education **67**(3), 382–393 (2018)
26. Spasić, I., Williams, L., Buerki, A.: Idiom-based features in sentiment analysis: Cutting the gordian knot. IEEE Transactions on Affective Computing **11**(2), 189–199 (2017)
27. Sun, C., Huang, L., Qiu, X.: Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. arXiv preprint arXiv:1903.09588 (2019)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
29. Vijjali, R., Potluri, P., Kumar, S., Teki, S.: Two stage transformer model for covid-19 fake news detection and fact checking. arXiv preprint arXiv:2011.13253 (2020)
30. Yokoya, S., Maeno, T., Sakamoto, N., Goto, R., Maeno, T.: A brief survey of public knowledge and stigma towards depression. Journal of clinical medicine research **10**(3), 202 (2018)

The phraseology of 'frontline' in the Covid-19 pandemic

Emma Franklin[1] ✉ [0000-0002-6434-9821] and Kathryn Spicksley[2] [0000-0002-1101-3726]

[1]Research Group in Computational Linguistics
[2] Institute for Community Research and Development
University of Wolverhampton, Wolverhampton, UK
{emma.franklin;k.spicksley}@wlv.ac.uk

Abstract. Since the outbreak of Covid-19, we have seen a surge in the use of the term 'frontline' to refer to key workers responding to the virus. Using the News on the Web (NOW) Corpus, we present a corpus pattern analysis of the adjective 'frontline' in British English news texts before and during the Covid-19 pandemic. Aside from the biopolitical implications of the Covid-as-war metaphor and its focus on health work, this new phraseological patterning raises questions about the impact of the pandemic on the meaning of 'frontline' itself.

Keywords: frontline, corpus pattern analysis, phraseology, Covid-19.

1 Introduction

War metaphors abound in the language surrounding Covid-19. The virus has been personified as an *enemy* and an *invader*, both in the British media and by political leaders around the world. In 2020, Boris Johnson referred to the pandemic as a *fight*, in which "each and every one of us is enlisted" [1], while Pedro Sanchez, for example, has called the virus *un enemigo al que aún estamos conociendo* 'an enemy we are still getting acquainted with' [2]. Elsewhere, healthcare workers have been framed as soldiers, with statements such as *Sie stehen für uns in diesem Kampf in der vordersten Linie* 'You are on the front line for us in this fight' [3] and *Вы сейчас на переднем крае защиты страны* 'You are now at the forefront of defending the country' [4, 5].

One of the outcomes of this pervasive metaphor in public discourse is the widespread construal of Covid-19 as a straightforward battle with two opposing sides, and thus the reduction of our various responses to the pandemic to a single 'front', or 'frontline' [6]. Another concerning consequence is the negative psychological impact this narrative has had on critical workers of all kinds [7]: health workers are consistently framed as *frontline* workers, ergo 'heroes' or accessories to war, while workers in other high-risk roles are typically excluded from this kind of valorising language [8, 9]. But what effect has this narrative had on the term *frontline* itself? In this paper, we examine the phraseological profile of the adjective *frontline*[1] before and during the Covid-19 pandemic via analysis of samples of the News on the Web (NOW) corpus.

[1] In this paper, *frontline* is used to represent all orthographic realisations of the term, that is, *frontline*, *front-line*, and *front line*.

2 Context

2.1 The *frontline* of the pandemic

The proverbial *frontline* and *frontline workers* of the Covid-19 pandemic have already come under some scrutiny in the scholarly literature, notably from sociological perspectives (e.g. [11]), but also linguistically, in terms of war metaphors [12, 13, 14] and discursive constructions of 'frontline' occupations [15, 16]. The narrative of war and its associated *frontline* has been found to co-opt health workers, in particular, into the unwilling role of 'hero' or 'soldier' [6, 17], while the risks and sacrifices of other keyworker roles, e.g. teachers, social workers and utilities workers, have largely been excluded from such discourse, with negative consequences for wellbeing [7, 8, 9, 18].

In our recent contribution to this discussion [16], we find that the frequency and behaviour of *frontline* in British news texts changes in line with the numbers of Covid-19-related deaths (according to the Office for National Statistics [19]) and also varies with the distinct 'waves' of the virus in the UK. At different points before and during the pandemic, *frontline* appears to be associated with different kinds of 'critical' job roles, and there is a clear preference for healthcare-related occupations in the news texts published following the outbreak of Covid-19. What we did not establish in that study, and what we focus on here, was the effect of the pandemic on the overall meaning of the word *frontline*, and whether it might be considered to now hold a different semantic profile as a result of the UK media reportage during the pandemic.

2.2 Corpus Pattern Analysis (CPA)

Especially useful for the semantic profiling of words, most of all verbs, is Hanks' [10, 20] Corpus Pattern Analysis (CPA) technique. Doing CPA involves taking a random concordance line sample for a particular word – usually a verb, and typically 250 lines – and carrying out an iterative sorting process that results in each line of the sample being tagged with a number corresponding to a syntagmatic pattern [10]. Patterns, built on the Sinclairian premise of form-function interdependence (cf. [21]), encode valency as well as preferred collocations, represented in CPA by *semantic types* (cf. [22]). These are logical constructs organised in a hierarchical 'is-a' ontology; for example, a *Fiat Panda* is classifiable as a *CAR*, which is a *ROAD VEHICLE*, which is a *VEHICLE* (is a *MACHINE*, is an *ARTIFACT ...)*. Grouping lexical items under these semantic types makes it easier to generalise the selectional preferences of words.

CPA has predominantly been carried out on verbs for lexicographical purposes, as in the *Pattern Dictionary of English Verbs* (http://pdev.org.uk), but it has also been carried out, minimally and experimentally, on nouns [23], prepositions [24] and modifiers [25], and has been applied to a range of problems, including machine translation evaluation [26], semantic processing [27] and discourse analysis [28]. The theory of norms and exploitations (TNE) that underpins CPA proposes a "double helix" representation of meaning, with norms on one side and exploitations of those norms on the other. The helix "is bidirectional, i.e. if on one hand norms are used to generate new semantic, figurative and syntactic exploitations, the latter can also turn into norms through frequent and continuous use over an extended period of time" [29, p. 12].

3 Methodology

The data used in this study (see Table 1) is sampled from the News on the Web (NOW) Corpus [30], a monitor corpus of some 10 billion words of online news made available via *English-Corpora.org*. One might ask why we chose not to use the Coronavirus Corpus [31], a thematic subcorpus of the NOW Corpus specifically designed for studying the language of Covid-19; in this study we have sought to compare the behaviour of *frontline* before and after the outbreak of Covid-19, for which the Coronavirus Corpus is too restrictive in both timeframe and scope. For the purposes of this study, we arbitrarily mark the outbreak of Covid-19 in the UK as 1 January 2020. We acknowledge that there are phraseological limitations to this genre of online news, as well as analytical limitations associated with the *English-Corpora.org* platform.

Table 1. Description of the corpus data used.

Dataset	Description	Texts	Tokens
NOW Corpus GB *frontline* sample, pre-Covid-19	Sample of 8,000 (max. limit) UK news texts that mention *front*line* between 1 January 2010 and 31 December 2019.	8,000	8,197,101
NOW Corpus GB *frontline* sample, during Covid-19	Sample of 8,000 (max. limit) UK news texts that mention *front*line* between 1 January 2020 and 31 December 2021.	8,000	8,839,718

From each dataset (pre- and during Covid-19), 300 adjectival instances of *frontline* were sampled at random and exported as concordance lines to Microsoft Excel for annotation. For each concordance line, the head noun modified by the attributive adjective *frontline* was assigned a CPA semantic type[2], e.g. *frontline carer* was assigned the semantic type *HUMAN*, and *frontline policing* was assigned *ACTIVITY*. The "hybrid" semantic type of *RESOURCE/ACTIVITY* was used to label *frontline service(s).*

4 Findings

Table 2 reports the semantic types assigned to adjectival instances of *frontline*, as well as some additional fine-grained classifications for the semantic type of *HUMAN*, with raw frequencies in brackets. As we can see, the dominant semantic type in both samples is *HUMAN*, but with very different proportions (n=97 pre-2020; n=194 2020 onwards), and with different distributions of fine-grained sub-types. In the first sample, the *HUMAN* sub-types are more evenly mixed, while the second sample shows a strong bias towards health workers (e.g. *doctor*, *carer*, *NHS staff*). Since the start of 2020, *frontline* is twice as likely to be associated with a *HUMAN*, as opposed to e.g. a *RESPONSIBILITY* (e.g. *frontline military role*) or *LOCATION* (e.g. *frontline state*) than it was before. Specifically, our results suggest that *frontline* is now six times as likely to describe health workers than it was before the outbreak of Covid-19 in 2020.

[2] CPA semantic types are listed on the PDEV website under 'Ontology': https://pdev.org.uk/

Table 2. CPA-inspired semantic types assigned to head nouns modified by *frontline*.

NOW GB (pre-2020) sample (300 lines)	NOW GB (2020/21) sample (300 lines)
▪ ***HUMAN (97)*** *[employee (30); health worker (19); police (18); sportsperson (10); social worker (3); soldier (3); unspecified human (3); activist (2); adviser (2); fire-fighter (2); prison officer (2); journalist (1); politician (1); royal (1)]* ▪ ***RESOURCE/ACTIVITY (51)*** ▪ ***ACTIVITY (40)*** ▪ ***LOCATION (16)*** ▪ ***RESPONSIBILITY (13)*** ▪ ***HUMAN GROUP (9)*** ▪ ***INSTITUTION (6)*** ▪ ***ASSET (4)*** ▪ ***RESOURCE (4)*** ▪ ***VEHICLE (3)*** ▪ ***EVENT (3)*** ▪ ***MONEY VALUE (1)*** ▪ ***STATE OF AFFAIRS (1)*** ▪ ***INFORMATION SOURCE (1)*** ▪ ***CONCEPT (1)***	▪ ***HUMAN (194)*** *[health worker (113); employee (56); police (9); transport staff (4); sportsperson (3); adviser (2); un-specified human (2); activist (1); animal rescuer (1); royal (1); social worker (1); soldier (1)]* ▪ ***RESOURCE/ACTIVITY (22)*** ▪ ***ACTIVITY (13)*** ▪ ***RESPONSIBILITY (6)*** ▪ ***INSTITUTION (5)*** ▪ ***LOCATION (4)*** ▪ ***HUMAN GROUP (3)*** ▪ ***EVENT (1)*** ▪ ***STATE OF AFFAIRS (1)*** ▪ ***RESOURCE (1)***

These findings are aligned with those of [16]: that *frontline* workers are now typically understood to be healthcare workers, and that *frontline services* – encoded here as *RESOURCE/ACTIVITY* – is a far less salient collocation now than it was pre-2020. We also see an overall narrowing, or restriction, of the semantic types of nouns modified by *frontline*: in the pre-2020 sample we identified 15 distinct semantic types, compared with 10 following the outbreak of Covid-19. Furthermore, the frequencies of non-*HUMAN* instances are significantly lower than they were before; *frontline* now appears to be predominantly a modifier of humans as opposed to activities or services.

This comparison also reveals a reduction in the normative semantic field of *frontline* typically seen before Covid-19: that of war, policing, sports, and other highly physical or literal defence domains. The now-heightened association of *frontline* with nouns denoting health workers strengthens the ongoing Covid-as-war metaphor by employing the phraseological norms of *frontline* to semantically coerce these workers into actors of war and defence. Conversely, turning to Hanks' TNE, we might also expect that this new and persistent use of *frontline* could lead to a new set of *frontline* norms "through frequent and continuous use over an extended period of time" [29].

5 Conclusion

In this short paper, we have presented findings from a small-scale analysis of the language of the adjective *frontline* pre- and post-2020 in the British press. While this is a limited and preliminary study, it raises important questions about the impact of current events on phraseology, and also represents a pilot application of CPA to adjectives. It remains to be seen whether this new phraseological patterning of *frontline* will survive beyond active media discussion of Covid-19. Given the deep and irreversible societal impact of Covid-19, we can only speculate that it will last some time.

References

1. Johnson, B. (2020, March 23) Prime Minister's statement on coronavirus (COVID-19): 23 March 2020. Available online at: https://www.gov.uk/government/speeches/pm-address-to-the-nation-on-coronavirus-23-march-2020, last accessed 12/05/2022.
2. El Español (2020, March 21). Largo monólogo de Sánchez para anunciar que "viene la ola más dañina". *El Español.* Available at: https://www.elespanol.com/espana/politica/20200321/largo-monologo-sanchez-anunciar-viene-ola-danina/476453146_0.html, last accessed 20/05/2022.
3. Die Bundesregierung (2020, March 18) Fernsehansprache von Bundeskanzlerin Angela Merkel. Available at: https://www.bundesregierung.de/breg-de/aktuelles/fernsehansprache-von-bundeskanzlerin-angela-merkel-1732134, last accessed 20/05/2022.
4. Президент России (2020, March 25) Обращение к гражданам России. Available at: http://kremlin.ru/events/president/news/63061, last accessed 20/05/2022.
5. Islentyeva, A. (2020). On the Front Line in the Fight against the Virus: Conceptual Framing and War Patterns in Political Discourse. *Yearbook of the German Cognitive Linguistics Association, 8*(1), 157-180.
6. Walker, I. F. (2020). Beyond the military metaphor. *Medicine Anthropology Theory, 7*(2), 261-272.
7. Bu, F., Mak, H. W., Fancourt, D. & Paul, E. (2022) Comparing the mental health trajectories of four different types of keyworkers with non-keyworkers: 12-month follow-up observational study of 21,874 adults in England during the COVID-19 pandemic. *The British Journal of Psychiatry, 1*(8).
8. Beames, J. R., Christensen, H. & Werner-Seidler, A. (2021) School Teachers: the forgotten frontline workers of Covid-19. *Australasian Psychiatry, 29*(4), 420-422.
9. Paul, T. J., Bruin, M. & Taylor, T. (2020) Recasting social workers as frontline in a socially accountable COVID-19 response. *International Social Work, 63*(6), 786-789.
10. Hanks, P. (2013) *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.
11. De Camargo, C. R. & Whiley, L. A. (2020) The mythologisation of key workers: occupational prestige gained, sustained...and lost? *International Journal of Sociology & Social Policy, 40*(9/10), 849-859.
12. Castro Seixas, E. (2021). War metaphors in political communication on COVID-19. *Frontiers in Sociology, 112*.
13. Semino, E. (2021) "Not Soldiers but Fire-fighters" – Metaphors and Covid-19. *Health Communication,* 36(1), 50-58.
14. Wicke, P., & Bolognesi, M. M. (2020). Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *PloS one, 15*(9), e0240010.
15. Farris, S., Yuval-Davis, N. & Rottenberg, C. (2021) The Frontline as Performative Frame: An Analysis of the UK Covid Crisis. *State Crime Journal, 10*(2), 284-303.
16. Spicksley, K. and Franklin, E. (forthcoming, 2022) Who works on the 'frontline'? Comparing constructions of frontline work before and during the Covid-19 pandemic. Applied Corpus Linguistics (Special Issue on 'Corpus Linguistics and the Language of COVID-19: Applications and Outcomes').
17. Lohmeyer, B. A. & Taylor, N. (2021) War, Heroes and Sacrifice: Masking Neoliberal Violence During the COVID-19 Pandemic *Critical Sociology, 47*(4-5), 625-639.
18. Sumner, R. C. & Kinsella, E. L. (2021) '"It's Like a Kick in the Teeth": The Emergence of Novel Predictors of Burnout in Frontline Workers during Covid-19. *Frontiers in Psychology 12,* 1875.

19. ONS (2021) Deaths involving COVID-19 by month of registration, UK, March 2020 to November 2021. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/deathsinvolvingcovid19bymonthofregistrationuk, last accessed 20/05/2022.
20. Hanks, P. (2004). Corpus Pattern Analysis. In G. Williams and S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (EURALEX 2004) (pp. 87-97). Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.
21. Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
22. Pustejovsky, J.: The Generative Lexicon. *Computational Linguistics 17*(4), 409-441 (1991).
23. Hanks, P. (2012) How people use words to make meanings: Semantic types meet valencies. In *Input, Process and Product: Developments in Teaching and Language Corpora* (pp. 54-69). Masarykova univerzita.
24. Litkowski, K. (2012). Corpus pattern analysis of prepositions. Technical report, Damascus, MD: CL Research.
25. Janssen, M. (2014). Half and other unique words: Corpus patterns and lexicalist syntax. In Simone, R. and Masini, F. (eds.), *Word Classes* (pp. 263-282). John Benjamins.
26. Béchara, H., Moze, S., El-Maarouf, I., Orasan, C., Hanks, P., & Mitkov, R. (2015) The Role of Corpus Pattern Analysis in Machine Translation Evaluation.
27. Popescu, O., Hanks, P., Ježek, E., & Kawahara, D. (2015, July). Corpus patterns for semantic processing. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: Tutorial Abstracts (pp. 12-15).
28. Franklin, E. (2017) Towards a Corpus-Lexicographical Discourse Analysis. In *Proceedings of EUROPHRAS 2017*, pp. 190-196.
29. Hanks, P., & Može, S. (2019) The way to analyse 'way': A case study in word-specific local grammar. *International Journal of Lexicography, 32*(3), 247-269.
30. Davies, M. (2017, July). The new 4.3 billion word NOW corpus, with 4–5 million words of data added every day. In The 9th International Corpus Linguistics Conference.
31. Davies, M. (2021) 'The Coronavirus Corpus: Design, construction and use.' International Journal of Corpus Linguistics, 26(4), 583-598.

PHRASEOLOGISCHE EINHEITEN UNTER BERÜCKSICHTIGUNG DES HISTORISCHEN UND KULTURELLEN RAHMENS FÜR DIE AKADEMISCHE ENTWICKLUNG UND DAS SPRACHLICHE ENGAGEMENT DER DAF-STUDENTEN

Marina Rueda Martín

Estudiante de primer año de doctorado en la Universidad Pablo de Olavide (Sevilla)

INTRODUCCIÓN

Toda asociación lingüística combina las palabras de forma que se crean enunciados, se repiten e interiorizan. Estas construcciones de estructura fija se llaman unidades fraseológicas y se caracterizan por representar el pensamiento concreto de una comunidad de hablantes (Velázquez, 2015). La disciplina que se encarga de su estudio se conoce como «fraseología» y sus unidades se diferencian por la composición en dos o más palabras significativas, holísticas acorde al valor y la estructura estable.

Penadés (2012, p.6) aclara que locuciones, significantes o conexivas, los proverbios y refranes tienden a ser cortos, llamativos y peculiares, pues están cargados de significado, siglos de historia e identidad social. Sin embargo, su origen es algo que se desconoce: se aprenden de forma oral e incuestionable debido a la fijación de tales estructuras previamente inexistentes.

En una primera aproximación, las lexías complejas con significado idiomático se determinan por «el principio de no composicionalidad», es decir, el sentido general de la unidad fraseológica es independiente del significado individual de sus componentes (Zuluaga, 1980) aunque el autor añade que la dificultad reside en su propia naturaleza: carga idiomática y fijación aunque no exenta de variación.

Zurdo (1993) continúa señalando la definición de Wandruszka (1969) como “Unterschiede des Wortschatzes spiegeln die unterschiedliche Erlebniswelt der Völker” (p.10): léxico como medio para aproximarse a la etimología de los refranes. La fusión historia-traducción no es más que la clave para acercar dos ámbitos interlingüísticos, lejos de compartir genética y zona geográfica (Wotjak, 1984 citado en Zurdo, 1993). Ahora bien, los fundamentos del análisis no pretenden

recorrer los parámetros actuales sino poner el ojo en épocas pretéritas: actos de habla “determinados por factores espaciales y temporales concretos, pero todavía no aceptados ni absorbidos como unidades estables por parte de esa misma comunidad” (Lázaro, 1980, p. 214).

Por otro lado, la camarada de expertos asegura que el desarrollo tecnológico ha ocasionado una transformación rápida y profunda de las estructuras socioeconómicas, culturales y de la conciencia histórica. Este fenómeno trae consigo un fuerte desapego de las tradiciones del pasado mientras la sociedad responde obediente a los cambios debilitando la “memoria colectiva", garante de la conservación y transmisión de los refranes (Lázaro, 1980, p.224). Como consecuencia, los recursos expresivos se entienden pero no generan; es decir, se desplazan al “nivel de la competencia multilectal pasiva” (Schlieben-Lange, 1977, p.1191) con claro reflejo en el mundo de la enseñanza.

Quizá el modesto abordaje sea el preámbulo de futuros trabajos en la misma línea temática. Desde el punto de vista didáctico-lexicográfico se examinan los repertorios fraseológicos y las referencias etimológicas (Celeste Frana, 2019) mientras que desde la perspectiva pedagógica, se motiva la mejora en el campo de instrucción. “En definitiva, el tema abordado es un gran disparador de planteamientos para el área de la enseñanza y el aprendizaje de lenguas, la lingüística y la traductología” (Zurdo, 1993).

HIPÓTESIS

Resulta sorprendente cómo los modismos, dichos o paremias siguen teniendo tanta fuerza sin llegar a caer en desuso, pues en el mejor de los casos, el repertorio fraseológico en los niveles superiores (C1-C2) es parvo debido a la falta de inclusión en las obras curriculares dentro del Marco Común Europeo de Referencia (MCER). Tal fenómeno conlleva a un desconocimiento de la parcela léxica que podría ralentizar la construcción de la interlengua y la capacidad comunicativa del alumnado (Velázquez, 2015, p.2). La autora continúa hablando sobre las dificultades reales de comunicación entre las lenguas alemán y español debido a la falta de dominio fraseológico generalizado, más aún en la actualidad, y especialmente en Europa, donde se ha visto un aumento significativo de la distancia subjetiva. Este fenómeno ha favorecido el aumento de la interconectividad a través de un transporte más rápido, medios de comunicación como Internet y la apertura de las fronteras dentro de la Unión Europea; además, los flujos migratorios entre el norte y el sur de Europa han provocado un entorno multicultural (Clyne, & Clyne, 1995; De Swaan, 2010).

Ejemplo de ello son los millones de turistas procedentes del norte del continente que llegan cada año a España o los estudiantes españoles que se dirigen hacia el norte de Europa en busca de oportunidades laborales.

En Europa, encontramos un creciente protagonismo del componente cultural en los planteamientos educativos: desde el documento del Consejo de Europa (Byram y Zárate, 1994) que describe la competencia sociocultural en términos de saber, saber hacer, saber comprender, saber aprender, saber implicarse y saber ser, hasta el más reciente Marco Común Europeo de Referencia (Consejo de Europa, 2002), que además de recoger el desarrollo de una competencia sociolingüística como parte de la competencia comunicativa, añade el desarrollo de otras competencias generales entre las que se encuentran los saberes interculturales.

El estudio sistemático de las unidades fraseológicas del alemán al español intentará disminuir la brecha lingüístico-cultural a la vez que se aprovecha de la relevancia de los datos para ofrecer oportunidades formativas creando usuarios competentes y autosuficientes. Seguidamente, se pretende ver cómo la anécdota de los relatos solidifica el conocimiento de las expresiones y dota al alumnado de valor para enfrentarse a situaciones reales de comunicación. Asimismo, se aspira a un avance significativo de la competencia interlingüsística en el estudiante hispanohablante y, sobre todo, a reavivar la historia del alemán mediante las curiosidades históricas, la cultura y las costumbres del pasado. La indiferencia quizá propicie el distanciamiento progresivo en los ámbitos que conforman el idioma y el bagaje académico de los hablantes nativos en la actualidad.

El proyecto podría ser cuanto menos curiosos para la contribución de las investigaciones fraseológicas desde el trasfondo cultural del alemán y la búsqueda de correlatos en español. De tal modo, comenta Areizaga que la traducción crea las similitudes interlingüísticas y conceptuales entre las lenguas y que “el significado de las palabras adquiere sentido cuando los interlocutores las ponen en relación con el contexto de la situación comunicativa en la que están y con su conocimiento previo del mundo” (Areizaga, 2005). Aquí mismo se asume que el conocimiento de los datos de origen de las unidades fraseológicas empodera a los estudiantes de herramientas comunicativas para crear seguridad en ellos mismos. Además, de nutrirse de curiosidades históricas que le ayuden al avance y sumersión en el idioma de estudio.

Desde mi perspectiva, el estudio de la lengua es un proyecto progresivo. La práctica del saber marca un antes y un después en la efectividad del conocimiento; pero, ¿hasta qué punto viene

motivado?, ¿entra en juego la afección?, ¿depende de los recursos didácticos o de una necesidad personal?

Estas cuestiones se responderán durante la investigación pero, sin lugar a duda, evocan al aprendizaje emocional donde la simpatía lingüística promueve el interés, la indagación y el recuerdo; algo similar al deporte donde la liberación de endorfinas y los beneficios físicos nos hacen partícipes de saludables estilos de vida. De hecho, inconscientemente se relacionan ideas conmemorando la naturaleza del aprendizaje humano: mecanismo de escucha, experiencias y "relaciones socioafectivas" para iniciarse en el habla (Mateo, 1995, p.126). Entretanto, el autor deja ver que las lenguas no son solo materias sino herramientas de comunicación social y que de tal modo, el aprendizaje resultaría más fácil, puesto que las lecciones resultan prácticas y eficaces a lo largo del tiempo. Habríamos de evitar la indiferencia en este campo de estudio, pues sus beneficios completan las tareas de instrucción en DaF.

OBJETIVOS

El propósito de mi investigación busca una mejora de la competencia lingüístico-cultural mediante la selección y análisis de unidades fraseológicas alemanas con carácter histórico. Su origen puede ser curioso para conformar el bagaje académico y personal del aprendiz de habla hispana, a la par que se trabajan áreas como la sintaxis, la morfología o el léxico (Ministerio de Educación y Ciencia, 1989). Entonces, la investigación aspira a influir beneficiosamente en la construcción de la interlengua y la didáctica de DaF. Con tal enfoque, se pretende que el alumnado, como cometa Altmayer (2004), aplique "los resultados de la investigación de los estudios culturales". Este pragmatismo en la política educativa puede impulsar la transdiciplinariedad de modalidades y asignaturas en un sistema educativo globalizado.

Considerando estas premisas, se pretende adoptar una posición epistémica-didáctica que se base en los siguientes objetivos:

- Conocer y reavivar la cultura que conforma la lengua alemana.
- Contribuir a los métodos de enseñanza fraseológica.
- Contribuir al resto de competencias lingüísticas.

- Dar la traducción y/o traducciones más relevantes para estudiantes extranjeros de habla hispana en situaciones reales de comunicación interlingüísticas.
- Despertar interés y sumergirse en los diferentes ámbitos de DaF.
- Esparcir y motivar la enseñanza de la lengua.
- Identificar los refranes con traducción literal en español y alemán.
- Facilitar y amenizar el aprendizaje autónomo.
- Observar si las traducciones literales comparten origen.
- Observar si hay algún punto en común entre ambas lenguas.
- Presentar información clara para el estudio lingüístico-fraseológico.
- Seleccionar unidades desde un punto de vista histórico-costumbrista.
- Sumergir a los estudiantes en antiguas tradiciones del pueblo germano.

Los presentes objetivos son lo suficientemente específicos para emprender la tarea de investigación: es un proyecto realizable gracias al apoyo de recursos físicos y digitales así como ilimitado para completar con nuevas entradas o adicción de información.

METODOLOGÍA

A propósito de cubrir las exigencias docentes, se diseña un corpus fraseológico especializado en estudios lingüístico-culturales (alemán-español). El bilingüismo garantiza la divulgación del conocimiento disponible para cualquier público que se preste a consultarlo. Estará dispuesto de forma clara y sencilla para hacer búsquedas avanzadas y recuperar datos cuanto menos recurrentes.

Comprendemos un amplia lista de proverbios, adagios de forma atractiva y con didáctica eficiente. Nos basamos, a primera instancia, en el origen etimológico con observación, quizá, de las evoluciones morfológicas y sus respectivas variaciones. Seguidamente, buscamos uno o varios correlato en español: “la traducción sería la búsqueda de equivalencia e interpretación de signos lingüísticos en códigos diferentes para respetar el mismo designado”(Del Blanco Mateos) mediante la comparación lingüística que relaciona símbolos y agiliza los procesos de aprendizaje.

Dado que los trabajos realizados en el ámbito de la fraseología contractiva muestran la existencia de paralelismos formales entre las UFs alemanas y españolas, se intuye que a mayor analogía formal interlingüística, mayor será el éxito en la comprensión de locuciones verbales somáticas desconocidas (Velázquez, 2015).

El contraste léxico activa el anclaje y la recuperación de datos (Velázquez, 2015), a la vez que ayuda a los estudiantes a que aseguren el reconocimiento de aquello que les resulte familiar (Ainciburu, 2011). Paralelamente, cabe la posibilidad de centrarse en detalles más puramente lingüísticos, dedicando un par de líneas al análisis sintáctico. Sin embargo, mi intención no es dar lecciones de gramática, más bien clarificar las explicaciones para reforzar ya no solo la lengua materna sino la distensión del aprendizaje en el nuevo idioma.

Cierto es que quienes carezcan de nivel lingüístico, el seguimiento de los contenidos quizá les resulte arduo; empero no descarto que la cultura sea el puente con el que iniciarse en el estudio de alemán.

Con todo mi pesar, nos ceñimos a las acepciones más útiles de las expresiones seleccionadas, excluyendo las traducciones secundarias para que el interesado las rescate fácilmente, analice y emplee con seguridad. De este modo, los elementos (léxico), sus correlatos (traducción) y los datos las conforman (historia) fomentan el éxito comunicativo en los discursos de alemán.

Partiendo de tal premisa, conseguimos unidades de naturaleza histórico-germana, exentas de implicar lo corpóreo y, por ende, de mayor envergadura para la propia comprensión. Tomaremos como referencia el manual de fraseología *Handbuch der Phraseologie* de los autores Burger, Buhofer y Sialm, en los diccionarios de *Redewendungen y Sprichtwörter* de la editorial Duden, más publicaciones de autores pioneros en el conocimiento de la materia. En cuanto a las muestras se extraen del diccionario alemán monolingüe en línea que cubre giros idiomáticos, refranes, proverbios y lenguaje coloquial *Redensartenindex,* complementado con traducciones ofrecidas por la *Hispanoteca* online con sus más de 433.000 entradas en español. Además, se acude al libro *Lexikon der Redensarten* para terminar de seleccionar expresiones propias de las siguientes categorías:

1. *Acontecimientos históricos*: nacen de sucesos e historias verídicas propias de una época.

2. *Concepciones del mundo*: atiende a los modos de entender y percibir la realidad. Dependiendo de la zona, el clima o los fenómenos naturales habría que dar designio a ciertos elementos.

3. *Costumbres y tradiciones*: repetidas actuaciones de comportamientos sucedidos por un colectivo, vinculados al quehacer cotidiano y duración en el tiempo: eventos, nupcias, comidas, bailes, relaciones interpersonales.

4. *Creencias*: abarca los aspectos culturales que guardan relación con la hechicería, la magia, los cuentos o leyendas.

5. *Personajes históricos:* realzados por su repercusión en acontecimientos históricos (1*).

6. *Vestimenta*: alude a prendas relacionadas con la moda, los vestuarios, tendencias o atuendos.

El análisis nos enseña que la mayor cantidad de locuciones hacen referencia a las costumbres y tradiciones del pueblo germano. De esta manera, la interpretación de la imagen o el nivel figurativo es el punto de partida para que el aprendiente sea capaz de elaborar redes asociativas y le permitan configurar un nexo entre la lectura literal y fraseológica (Timofeeva, 2013).

The conception of Glossomatic, a trilingual corpus-based glossary for the translation of manipulated idioms

Carlos Manuel Hidalgo Ternero[1][0000-0002-8338-2627]

[1] University of Malaga, Avda. Cervantes, 2. 29071 Málaga, Spain
cmhidalgo@uma.es

Phraseology plays a pivotal role in the translation task. Thus far, however, there remains a paucity of translation tools that can specifically assist the translator in finding appropriate phraseological equivalences for manipulated idioms in the source text, let alone for manipulated somatisms in particular (i.e. idioms containing terms that refer to human or animal body parts). Against such a background, this study aims to shed some light on how the trilingual (ES-EN-DE) corpus-based glossary of somatisms called Glossomatic was conceived. In this regard, we will first examine the diverse state-of-the-art technologies that have shaped its current form, in order to then analyse its potential as an essential tool in the translation of manipulated somatisms.

Keywords: Corpus-Based Glossary, Translation, Idiom Manipulation.

1 Introduction

In this paper, we will present how the corpus-based glossary of somatisms (i.e. idioms containing terms that refer to human or animal body parts) called Glossomatic was conceived. This tool was designed for the establishment of ad hoc phraseological equivalents in those cases in which phraseological manipulation in the source text together with the absence of biunivocal cross-linguistic correspondences could pose problems for the translation task. In this context, we will first analyse the different state-of-the-art technologies that have served as a reference in its design, in order to then review the functionalities of Glossomatic.

At this point, we are delving into the field of *Phraseography*, defined as follows by one of the main figures in the domain, Dr. Olímpio de Oliveira Silva (2007).

> La fraseografía es una disciplina lingüística que se ocupa, por una parte, de los principios teóricos y prácticos que rigen la inclusión de la fraseología en compilaciones léxicas (diccionarios, léxicos, vocabularios, glosarios, concordancias, etc.), tanto restringidas como generales, y, por otra, del estudio crítico y descriptivo de estas compilaciones, en lo que al tratamiento de la fraseología se refiere, lo que significa decir que el ámbito de interés de la fraseografía comprende desde la presentación tipográfica seguida en la obra hasta la adecuación a los usuarios. (p. 27)[1]

[1] Phraseography is a linguistic discipline that deals, on the one hand, with the theoretical and practical principles that govern the inclusion of phraseology in both restricted and general lexical compilations

In the following sections, we will focus more specifically on *practical phraseography*, which refers to "the activity of elaboration of phraseological dictionaries [as well as] the establishment of the technique or methodology of elaboration of phraseological dictionaries" (Olímpio de Oliveira Silva, 2007, p. 28).

2 The conception of Glossomatic

In the literature it is already possible to find some proposals for lexicographic resources that focus exclusively on the study of somatisms and that include Spanish among the addressed languages. One of them is that of Corrêa Rocha (2014), who, in *A elaboração de um repertório semibilíngue de somatismos phraseológicos do português brasileiro para aprendizes argentinos* ('The design of a semi-bilingual repertoire of Brazilian Portuguese idioms for Argentinian learners'), presents a semi-bilingual dictionary of Portuguese somatisms (diatopic variety of Brazil) with equivalents in Spanish (diatopic variety of Argentina). For each entry, the following lexicographical information is provided:

> Idiom: + periphrastic linguistic-conceptual definition; synonymic linguistic-conceptual definition. Example. # lexicultural information. = Equivalent in Argentinian Spanish (Corrêa Rocha, 2014, p. 139).

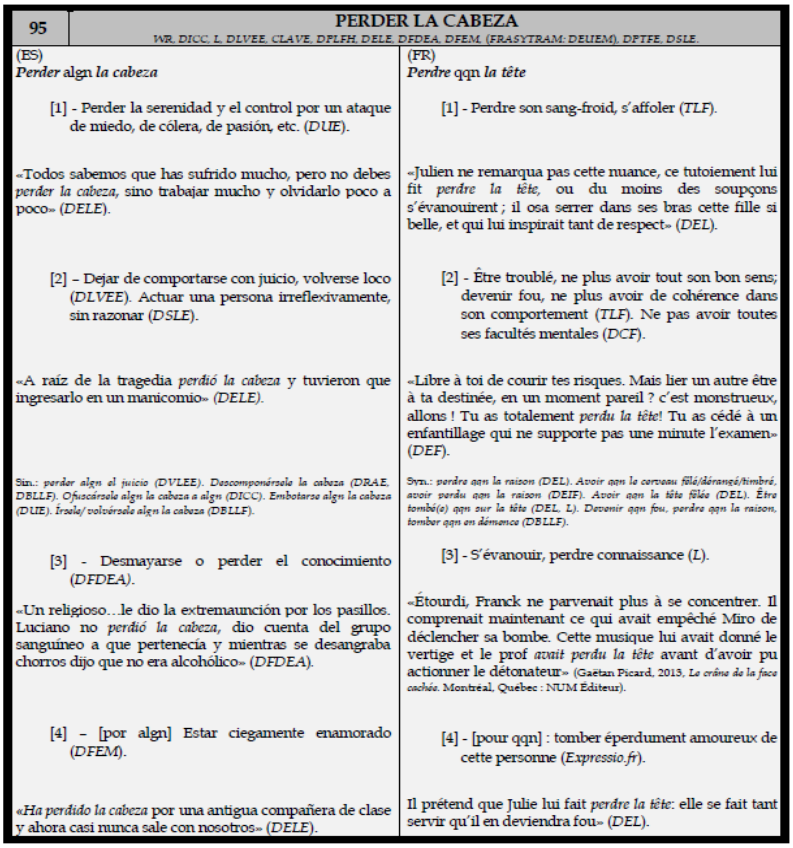
> Example:

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

In this way, following a semasiological approach, it is possible to consult each somatism in alphabetical order according to the first somatonym it includes (in the example, *cabeça* ['head']). Each entry includes a periphrasis explaining the meaning of the idiom, a synonym (if there is one), an example of use (taken from the web), lexicocultural information (with information on popular etymology, conceptual metaphor from which the somatism originates, etc.) and equivalent(s) in the lexical level in Argentinian Spanish.

(dictionaries, lexicons, vocabularies, glossaries, concordances, etc.), and, on the other hand, with the critical and descriptive study of these compilations, as far as the treatment of phraseology is concerned. This means that the field of interest of Phraseography ranges from the typographical presentation in the work to the adaptation to the users. [The translation is ours]

Baran à Nkoum (2015), in *La cabeza en las locuciones verbales españolas. Locuciones somáticas y correspondencias francesas* ('The head in Spanish verbal idioms. Somatic idioms and French correspondences'), offers again a model of a semasiological dictionary, in which each entry, classified according to the somatonym, comprises the different possible meanings of the idiom, examples of use taken from several lexicographical works and synonyms. For each of these meanings, French equivalents are provided, also with examples of use and synonyms. An example of the entry "perder la cabeza" ('to lose one's mind') can be observed in Figure 1 below.



95	PERDER LA CABEZA *WR, DICC, L, DLVEE, CLAVE, DFLFH, DELE, DFDEA, DFEM, (FRASYTRAM: DEUEM), DFTFE, DSLE.*
(ES) ***Perder*** algn ***la cabeza***	(FR) ***Perdre*** qqn ***la tête***
[1] - Perder la serenidad y el control por un ataque de miedo, de cólera, de pasión, etc. (*DUE*).	[1] - Perdre son sang-froid, s'affoler (*TLF*).
«Todos sabemos que has sufrido mucho, pero no debes *perder la cabeza*, sino trabajar mucho y olvidarlo poco a poco» (*DELE*).	«Julien ne remarqua pas cette nuance, ce tutoiement lui fit *perdre la tête*, ou du moins des soupçons s'évanouirent ; il osa serrer dans ses bras cette fille si belle, et qui lui inspirait tant de respect» (*DEL*).
[2] – Dejar de comportarse con juicio, volverse loco (*DLVEE*). Actuar una persona irreflexivamente, sin razonar (*DSLE*).	[2] - Être troublé, ne plus avoir tout son bon sens; devenir fou, ne plus avoir de cohérence dans son comportement (*TLF*). Ne pas avoir toutes ses facultés mentales (*DCF*).
«A raíz de la tragedia *perdió la cabeza* y tuvieron que ingresarlo en un manicomio» *(DELE).*	«Libre à toi de courir tes risques. Mais lier un autre être à ta destinée, en un moment pareil ? c'est monstrueux, allons ! Tu as totalement *perdu la tête*! Tu as cédé à un enfantillage qui ne supporte pas une minute l'examen» (*DEF*).
Sin.: *perder algn el juicio (DVLEE). Descomponérsele la cabeza (DRAE, DBLLF). Ofuscársele algn la cabeza a algn (DICC). Embotarse algn la cabeza (DUE). Írsele/ volvérsele algn la cabeza (DBLLF).*	Syn.: *perdre qqn la raison (DEL). Avoir qqn le cerveau fêlé/dérangé/timbré, avoir perdu qqn la raison (DEIF). Avoir qqn la tête fêlée (DEL). Être tombé(e) qqn sur la tête (DEL, L). Devenir qqn fou, perdre qqn la raison, tomber qqn en démence (DBLLF).*
[3] - Desmayarse o perder el conocimiento (*DFDEA*).	[3] - S'évanouir, perdre connaissance (*L*).
«Un religioso…le dio la extremaunción por los pasillos. Luciano no *perdió la cabeza*, dio cuenta del grupo sanguíneo a que pertenecía y mientras se desangraba chorros dijo que no era alcohólico» (*DFDEA*).	«Étourdi, Franck ne parvenait plus à se concentrer. Il comprenait maintenant ce qui avait empêché Miro de déclencher sa bombe. Cette musique lui avait donné le vertige et le prof *avait perdu la tête* avant d'avoir pu actionner le détonateur» (Gaëtan Picard, 2013, *Le crâne de la face cachée*. Montréal, Québec : NUM Éditeur).
[4] – [por algn] Estar ciegamente enamorado (*DFEM*).	[4] - [pour qqn] : tomber éperdument amoureux de cette personne (*Expressio.fr*).
«*Ha perdido la cabeza* por una antigua compañera de clase y ahora casi nunca sale con nosotros» (*DELE*).	Il prétend que Julie lui fait *perdre la tête*: elle se fait tant servir qu'il en deviendra fou» (*DEL*).

Figure 1. *Example of the entry "perder la cabeza"*

Another lexicographic tool for somatisms can be found in Rayyan's (2014) PhD project *Fraseología y lingüística informatizada: elaboración de una base de datos electrónica contrastiva árabe-español/español-árabe de fraseologismos basados en partes del cuerpo* ('Computational Linguistics and Phraseology: development of an Arabic-Spanish/Spanish-Arabic contrastive electronic database of idioms based on body parts'). This database presents four types of somatic dictionaries: a Dictionary of Equivalences (which includes *total* and *partial equivalences* according to the terminology of Corpas

Pastor, 2003), a Dictionary of Incorrect Usage (which contains *apparent equivalences*), a Thematic Dictionary of Somatisms (ordered according to the target domain and sub-domain) and, finally, an Axiological Dictionary of Somatisms (according to the connotative value of the somatisms). In Figure 2, we show an example entry for "perder la cabeza" in the Thematic Dictionary of Somatisms, which also includes information on the type of equivalence. Following the author's terminology, this case is considered "full equivalence (B)" since both idioms coincide both in connotative and denotative meaning as well as in register, but the somatonym is different (in Spanish it is “cabeza” [‘head’] and in Arabic it is “mente” [‘mind]). In Figure 3, we also show an example of the Axiological Dictionary of Somatisms for different idioms with the somatonyms "ala" (‘wing’), and "alma" (‘soul’).

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

Figure 2. *Excerpt from the Thematic Dictionary of Somatisms*

DICCIONARIO AXIOLÓGICO ÁRABE - ESPAÑOL

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

Figure 3. *Excerpt from the Axiological Dictionary of Somatisms*

With regard to other general phraseological databases (not exclusive to somatisms) that have served as a reference for Glossomatic, we can highlight, on the one hand, the application created by the FRASYTRAM[2] research group, which includes fixed verbal constructions (FVC) (Mogorrón, 2008 and 2010) such as verbal idioms, verbal collocations, support verbs and comparative verbal constructions. This database includes FVCs in 10 different languages: Spanish (8642 FVCs), Catalan (1610 FVCs), French (2704 FVCs), Italian (470 FVCs), Arabic (449 FVCs), Polish (469 FVCs), Russian (330 FVCs), English (1585 FVCs), German (1879 FVCs) and Chinese (2 FVCs). In order to analyse the lexicographic information collected in each entry, we can observe Figure 4, with an entry for the idiom "perder la cabeza".

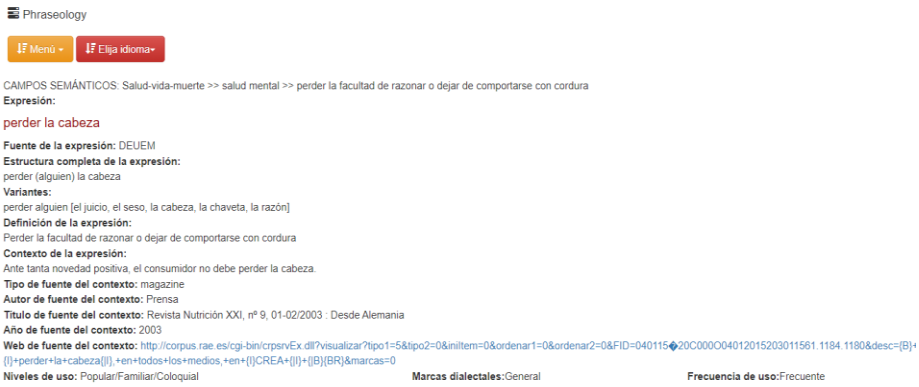


Figure 4. *Example of entry for the idiom "perder la cabeza"*

In this way, it is possible to find the idiom "perder la cabeza" within the semantic field "Salud-vida-muerte >> salud mental >> perder la facultad de razonar o dejar de comportarse con cordura" ('Health-life-death >> mental health >> to lose the faculty of reasoning or to stop behaving sanely'). Within this entry, different lexicographical information can be consulted, such as *source of the expression*, *complete structure of the idiom* (including here the possible actants), *variants*, *definition*, *context* (with information about it), *level of use* (popular/familiar/colloquial), *diatopic variety* and *frequency of use*. Additionally, in the subfield "perder la facultad de razonar o dejar de comportarse con cordura", it is possible to consult all the parasynonymous idioms in Spanish (*cruzarse los cables*, *perder la razón*, etc.), and equivalents in Catalan (*perdre el seny*, *beure's l'enteniment*, etc.), English (*not to be in one's right mind*, *to fly off the handle*...), and French (*[ne pas] avoir toute sa tête*, *perdre le nord*, etc.).

Another reference work was the *Diccionario de Locuciones Idiomáticas del Español Actual (DiLEA)* ('Dictionary of Idioms in Current Spanish'), designed by Penadés Martínez (2019), which mainly includes verbal idioms. In it, each entry contains different relevant lexicographic information, such as *category* (in the case of *perder la cabeza* is an intransitive verbal idiom), *diaphasic marking*, *frequency of use*, *idiom actants*

[2] The application of the multilingual database is available through the following link: http://84.127.230.137:6263/phraseology/

("[someone] loses his head"), *definition*, *examples* and *corpus with examples*, as we can see in Figure 5.

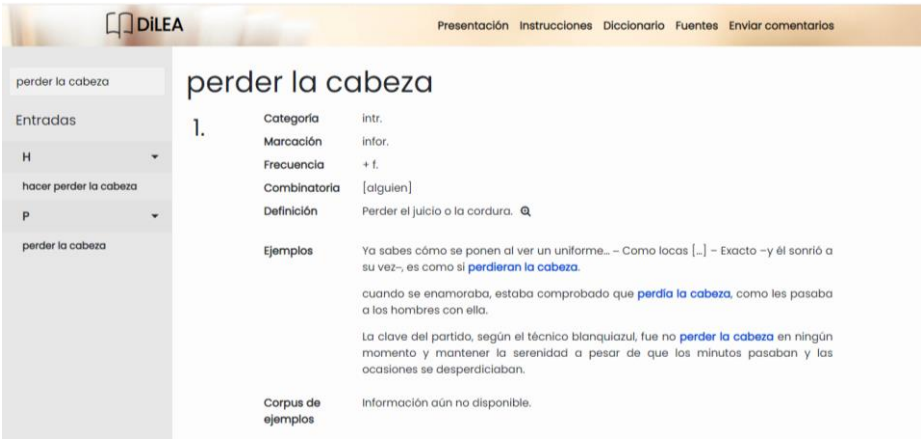


Figure 5. *Example of the entry "perder la cabeza" in the DiLEA dictionary*

PHRASEOTEXT (González Rey, 2017), a monolingual phraseodidactic dictionary with idioms in French extracted from a literary corpus of 18 contemporary authors (20th & 21st centuries) (González Rey, 2017, p. 33), also served as a basis for Glossomatic. Regarding its microstructure, as we can see in Figure 6, PHRASEOTEXT offers different lexicographic information for each entry, such as i*diom*, *context*, *definition*, *keyword*, *concept*, *author of the text*, *source* and *CEFR level*.

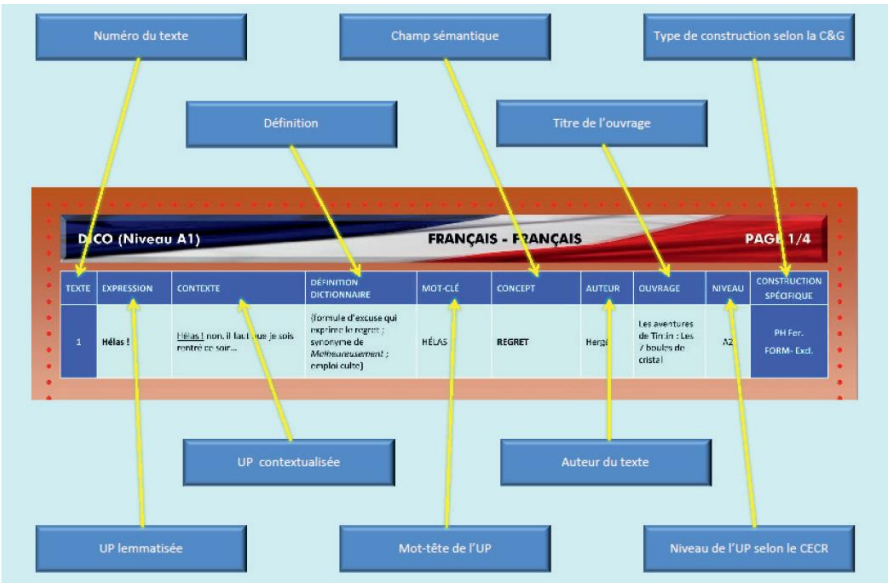


Figure 6. *Example of the entry "hélas" in the PHRASEOTEXT dictionary*

Finally, with regard to works that address, from a theoretical point of view, the design criteria for a phraseological database, we can highlight the chapter entitled *Diseño de una base de datos fraseológica para la traducción asistida por ordenador (TAO)* ('Design of a phraseological database for computer-assisted translation [CAT]) (Corpas Pastor, 2003). In it, the author mentions the information elements that a phraseological database for CAT should contain: *[1] language, [2] unit type, [3] neologism (yes/no), [4] unit identification, [5] semantic features, [6] syntactic features, [7] stylistic and expressive constraints, [8] pragmatic and discursive features, [9] Situational frames (for formulae), [10] Manipulations, [11] Related multiword units, [12] Related single-word units*, and *[13] Translation equivalents* (Corpas Pastor, 2003, pp. 194-196). Additionally, the author, as far as the design of the phraseological database is concerned, recommends the relational model as it "offers the necessary characteristics for a correct modelling that guarantees adequate levels in terms of redundant information, as well as information reuse and retrieval" (Corpas Pastor, 2003, p. 197). This was therefore the model we employed in the design of Glossomatic, as we will be able to observe in the following sections.

3 The design of Glossomatic

As described in Corpas Pastor et al. (2020), the glossary Glossomatic, designed in a relational database in Microsoft Access 2016[3], follows a dual onomasiological and semasiological approach, due to its very nature: since Glossomatic is designed for the creation of ad hoc phraseological equivalents in cases of phraseological manipulation in the source text, it is essential that not only the necessary information on a given idiom and its equivalents in other languages can be consulted, but that it is also possible to obtain idioms based on a series of parameters (register, semantic field, phono-stylistic characteristics, etc.).

In this way, for each entry, in addition to the equivalent(s) in English and German, relevant microstructural information is provided, such as *somatism type* (adjectival, verbal, nominal, prepositional or adverbial), *notions*, *collocations*, *register* (formal, neutral, informal or vulgar), *polarity* (positive, neutral or negative) and *crosslinguistic differences*. In the *notion* field, different concepts associated with an idiom (*anger*, *deceit*, *love*, etc.) are included, which allows easy retrieval of idioms within the same semantic field. For example, under the notion *thought* we find idioms such as *rondar por la cabeza* and its correspondences in English (*to run through someone's head*) and German (*jemandem im Kopf [herumspuken/herumgeistern/herumgehen]*). For those idioms with a different collocational spectrum in the different languages included, a *collocate* box has been created. For example, the Spanish idiom *rondar por la cabeza* can collocate with concepts such as *idea* or *project* (among others) but not with *music*, unlike *to run through someone's head* which can collocate with all three. These and other possible divergences among the idioms are summarised in the *differences* box, for a quick visualisation by the translator. Furthermore, the asterisk "*" has been used for all

[3] Microsoft Access 2016 is a Database Management System (DBMS) developed by Microsoft and belonging to the Microsoft Office suite of applications.

those correspondences which are an idiom but not a somatism and the double asterisk "**" in case of having employed a neutral term or explanatory periphrasis in the absence of a phraseological correspondence.

In addition to these features, the glossary also allows the search for idioms containing a certain combination of letters in case the translator wants to create a specific phono-stylistic effect in the target text. In this way, the search system allows the use of the most usual wildcards such as "*", which finds 0 or more characters; "?", for a specific number of characters; "[]", to detect characters entered in brackets; "¡", to exclude characters in brackets; "-", to search for a certain range of characters; and "#", to retrieve a numeric character.

In this context, the main view of Glossomatic is presented in Figure 7.

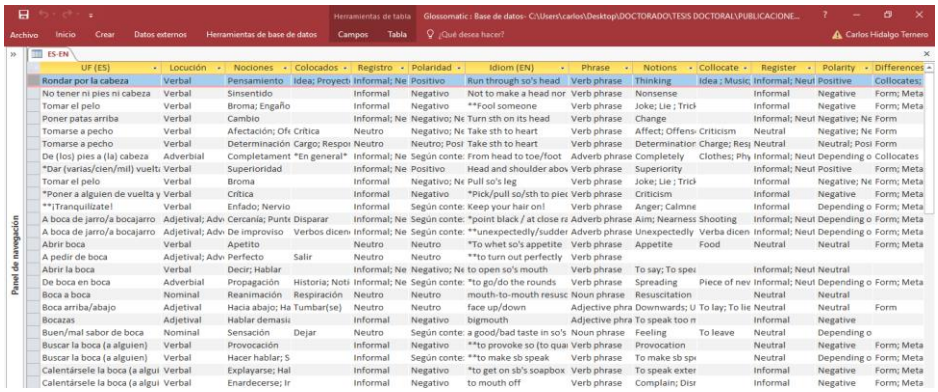


Figure 7. *Main view of Glossomatic*

4 Using Glossomatic for the translation of manipulated idioms

In order to observe how the glossary Glossomatic has been conceived for the translation of manipulated idioms, let us analyse the following example (cf. Corpas Pastor et al., 2020) with an article published in the Spanish newspaper El País (Figure 8).

Figure 8. *Article entitled Willian José, un nueve de los pies a la cabeza (Rodrigálvarez, 2017)*

Along this article the journalist seeks to highlight the fact that the football player Willian José is the best heading striker in the Spanish First Division. With this purpose, in the headline, the idiom *(un 'nueve') de los pies a la cabeza* is used, meaning Willian José is a 'total striker'. In order to emphasise this idea, the idiom includes the word *pie* ('foot') and *cabeza* ('head'); which hence activates both the literal and the figurative interpretation of the unit.

In the glossary, when searching for a primary correspondence in English for the idiom *de los pies a la cabeza*, there appears the expression *from head to toe/foot*, with similar form and metaphorical base. Nevertheless, when consulting all of its limitations, it is possible to ascertain the fact that *from head to toe/foot* con only co-occur with concrete and tangible concepts such as *being dressed or being covered (in a specific substance) from head to toe/foot*. In the BNC corpus (BNC Consortium, 2017), no concordance has been registered in which *from head to toe/foot* can also collocate with abstract concepts such as a profession or a skill, unlike its Spanish *counterpart*. Consequently, in this context both idioms are not textual equivalents.

In order to maintain the manipulation of the idiom in the target text, the translator could alternatively search the glossary for other English idioms containing the base *head*. Since no other idiom with *head* can be detected under the notion *completely* (analogously to *de los pies a la cabeza*), the translator could then explore other notions related to the article's main idea, e.g., *superiority*. Under those criteria, another idiom emerges: *head and shoulders above (someone)*, meaning *far superior to* (EOLD, 2022). A possible translation of the headline could hence be *Willian José, head and shoulders above the rest*, which would not only activate both the literal and the figurative interpretation of the idiom (similarly to the Spanish version) but it would also be reinforced by the main photo of the article, in which William José is literally jumping *head and shoulders above* another player.

When searching the glossary for a correspondence in German for the idiom *de los pies a la cabeza*, there appears *von Kopf bis Fuß*, which presents similarities not only in its form and metaphorical base, but also in its meaning, collocates and polarity. According to *Redensarten-Index*, this idiom means 'von oben bis unten; völlig; durch und durch' ('top to bottom; completely; through and through'). Seca & Wimmer (2013, p. 47) corroborate their correspondence: both in German and Spanish this expression can have a transparent meaning ('de arriba abajo', 'from top to bottom') and a more opaque one ('totalmente', 'totally'). To further verify our proposal, it was checked against corpus concordances from the *DWDS-Kernkorpus (1900–1999)* (cf. Geyken, 2007), where both senses could be detected in occurrences such as "Ein Gentleman von Kopf bis Fuß" or "Aber jeder Mensch kann sich täglich von Kopf bis Fuß waschen". This coincidence in form and meaning enables us to keep the manipulation detected in the original headline in Spanish. A possible translation into German could therefore be *Willian José, ein Stürmer von Kopf bis Fuß*, which maintains both the literal and figurative interpretation of the idiom, analogously to the Spanish version.

5 Conclusion

Translating idioms is no easy task. While looking for full correspondences it is necessary to take into account several factors regarding the source idiom and any target equivalents, viz., the specific denotative and connotative meanings, conventional implicatures, diasystematic restrictions, discursive, pragmatic and semantic load, among others. The process gets more complicated when searching for textual equivalents, and further more in the case of idiom manipulation. If, besides their cross-linguistic anisomorphism, these idioms undergo either internal or external manipulation in the source text, as it is often the case, chances are that some important facets and nuances will be lost along the path to the target text.

In this regard, along the present publication we have been able to observe an extensive variety of state-of-the-art lexicographic tools that can be of help in the translation of a particular type of idioms: the somatisms. Besides their potential in the translation task, these resources have also helped us design Glossomatic, a corpus-based glossary of somatisms created for the establishment of ad hoc phraseological equivalents in those cases in which phraseological manipulation in the source text together with the absence of biunivocal interlingual correspondences could pose problems for the translation task.

Given the current possibilities offered by Glossomatic, in future research we would like to escalate this system into a hybrid dictionary of somatisms, which, besides including all that useful lexicographic information (*definition, register, notions, correspondences in other languages*…), will also provide translators with direct access to quality corpora in order for them to be able to examine a somatism (and its variants) in contexts of real use. In this way, we will develop an eclectic approach that will combine the best of both worlds: the conciseness and clarity of dictionaries with the comprehensiveness and textual richness of corpora.

References

1. Baran À Nkoum, P. Estudio contrastivo español-francés de las locuciones verbales somáticas relativas a la cabeza [PhD Thesis, Universidad Complutense de Madrid] https://eprints.ucm.es/id/eprint/33799/ Last accessed 2022/06/13 (2015).
2. BNC Consortium, *The British National Corpus, XML Edition*, Oxford Text Archive, http://hdl.handle.net/20.500.12024/2554. (2007)
3. Corpas Pastor, G. Diez años de investigación en fraseología: análisis sintáctico-semánticos, contrastivos y traductológicos. Vervuert, Madrid (2003).
4. Corpas Pastor, G., Hidalgo Ternero, C. M., Bautista Zambrana, M. R. Teaching idioms for translation purposes: a trilingual corpus-based glossary applied to Phraseodidactics (ES/EN/DE). In: F. Mena Martínez y C. Strohschen (Eds.), Teaching Phraseology in the XXI Century: New Challenges (pp. 75-93). Peter Lang, Berlin (2020).
5. EOLD – Oxford University Press. *English Oxford 'Living' Dictionaries*. Oxford Dictionaries. http://www.oxforddictionaries.com/. Last accessed 2022/06/15 (2022).
6. Geyken, A. The DWDS corpus: A reference corpus for the German language of the 20th century. *Collocations and idioms: Linguistic, lexicographic, and computational aspects, 23*. Continuum Press. (2007)
7. González Rey, M. I. Le dictionnaire phraséodidactique: sa place dans la didactique de la phraséologie. Studii de lingvistică, 7, 27–44 (2017).
8. Mogorrón Huerta, P. Analyse du figement et de ses possibles variations dans les constructions verbales espagnoles, Linguisticae Investigationes, 33(1), 86–151 (2010).
9. Mogorrón Huerta, P. Traduction et compréhension des locutions verbales. Meta: journal des traducteurs/Meta: Translators' Journal, 53(2), 378–406 (2008).
10. Olímpio de Oliveira Silva, M. E. Fraseografía teórica y práctica. Peter Lang, Berlin (2007).
11. Penadés Martínez, I. Diccionario de locuciones idiomáticas del español actual. www.diccionariodilea.es Last accessed 2022/06/03 (2019).
12. Rayyan, M. S. Fraseología y lingüística informatizada. Elaboración de una base de datos electrónica contrastiva arabe-español / español-arabe de fraseologismos basados en partes del cuerpo [PhD Thesis, Universidad de Granada] (2014).
13. Rocha, C. M. C. A elaboração de um repertório semibilíngue de somatismos fraseológicos do português brasileiro para aprendizes argentinos [PhD thesis, Universidade Estadual Paulista Júlio de Mesquita Filho]. (2014).
14. Rodrigálvarez, E. Willian José, un 'nueve' de los pies a la cabeza. *El País* https://elpais.com/deportes/2017/04/28/actualidad/1493398636_157243.html Last accessed 2022/06/23 (2017, April 28)
15. Seca, J., Wimmer, S. *Das kannst du laut sagen*. Barcelona: Difusión. (2013)

The Role of Semi-productivity in Multiword Expression Identification: Why can BERT Capture novel MWEs?

Nicola Cirillo[1][0000−0002−2107−1313] and Antonietta Paone[2][0000−0001−5662−307X]

[1] University of Salerno, Italy
nicirillo@unisa.it
[2] "L'Orientale" University of Naples, Italy
apaone@unior.it

Abstract. In this paper we argue that multiword expression identification systems based on BERT are able to capture semi-productive patterns that generate multiword expressions. To test this hypothesis we analyzed the results obtained by MTLB-STRUCT on unseen multiword expressions during the edition 1.2 of the PARSEME shared task. We observed that MTLB-STRUCT discovers, in proportion, more light verb constructions and verb particle constructions than verbal idioms. Since light verb constructions and verb-particle constructions often result from semi-productive patterns, while verbal idioms are more idiosyncratic, the results corroborate our hypothesis.

Keywords: MWEs · BERT · NLP.

1 Introduction and State of the Art

Multiword expressions (MWEs) pose a significant challenge to NLP [19, 5]. [19] identify two main problems that arise from the incorrect treatment of MWEs: *idiomaticity* and *overgeneration*. *Idiomaticity* is the feature of those expressions whose meaning goes beyond the meaning of their parts (e.g. *spill the beans*). Without an adequate treatment, NLP systems are more likely to interpret these expressions literally. On the other hand *overgeneration* is about the generation of sequences that are perfectly formed from a syntactic and semantic perspective but are nonetheless unacceptable (e.g. **telephone closet* instead of *telephone box*). Furthermore, MWEs are pervasive across all languages and text genres thus their treatment is a crucial task in NLP.

[5] distinguish between two tasks concerning MWEs: *MWE discovery* and *MWE identification*. While the former aims at building MWE lexicons automatically or, at least, at assisting lexicographers in the job, the latter focuses on the identification and annotation of MWEs in running text. In recent years, the attention on MWE identification has grown significantly thanks to the Shared Task on Automatic Identification of Verbal Multiword Expressions proposed by the PARSEME European network in 2017 [21] which counts currently three editions [21, 17, 18].

The PARSEME shared task focuses on the identification of verbal multiword expressions (VMWEs). VMWEs are defined as those MWEs which have a verb as syntactic head. They are further divided into light verb constructions (LVCs), verbal idioms (VIDs), Inherently reflexive verbs (IRVs), verb-particle constructions (VPCs), multi verb constructions (MVCs), inherently adpositional verbs (IAVs), and inherently clitic verbs (ICVs).

Participants of the PARSEME shared task addressed the problem of MWE identification mostly via *linear supervised classifiers* such as CRF (Conditional Random Field) [12, 14] and SVM (Support Vector Machine) [1], or via *deep learning models* like CNN (Convolutional Neural Network) [22, 23], BDLSTM (BiDirectional Long-Short Term Memory) [8, 3, 23], and BERT (Bidirectional Encoder Representations from Transformers) [9, 23]. Rule-based models have also been employed with good results [16].

Even though MWE identification achieved remarkable results, especially via deep learning, there is still an open problem: even state-of-the-art classifiers show poor performances when dealing with those MWEs that do not appear in training data (*unseen MWEs*) [17, 18].

To solve this problem, it has been proposed to use large-coverage MWE lexicons and MWE discovery techniques [20]. However, such techniques have been proven effective only when dealing with very specific MWE classes [5]. In addition, it should be noted that the models that performed the best in the last two editions of the PARSEME shared task: SHOMA [24] and MTLB-STRUCT [23], both based on deep learning, achieved the best score also on unseen MWEs [17, 18].

Therefore, we believe that deep learning architectures like BERT, especially if pre-trained on large text corpora, may be able to capture some regularities that allow them to identify novel MWEs. In this respect, [15] showed that BERT is able to distinguish sentences that contains idioms from sentences that do not.

We argue that BERT is able to capture those novel MWEs that result from *semi-productive patterns* while being less able to identify novel MWEs with more idiosyncratic nature.

To test this hypothesis, we analyzed the results of MTLB-STRUCT [23], a model for MWE identification based on pretrained BERT [6] that obtained the best score at the edition 1.2 of the PARSEME shared task [18].

Section 2 gives an overview of semi-productive mechanisms that generate VMWEs. Section 3 describes the experimental setup. Section 4 provides the results of the experiment that are further discussed in Section 5. Finally, in Section 6 we report our conclusions.

2 Verbal multiword expressions and productivity

Even though MWEs are usually characterized by syntactic, semantic, and lexical idiosyncrasies, some combination of words that are generally considered MWEs are the result of semi-productive patterns. This happens with higher frequency in VMWE categories like verb-particle constructions (VPCs) and light verb constructions (LVCs) and more rarely among verbal idioms (VIDs).

Some authors [2, 13] compare VPCs to morphologically complex words, arguing that the particles are functionally similar to affixes. This phenomenon is more common in Germanic languages. In Dutch, for example, particles are used to form derived verbs [2].

In addition, some VPCs are composed of a particle and an open slot that can be filled with verbs with certain characteristics. It is the case, for example, of the Italian particle *via*, that can be combined with agentive verbs of removing to convey a resultative meaning (1) [13].

(1) $[\text{V}[\text{via}]_{\text{P}}]_{\text{VPC}}$
V = agentive verb of removing

a. raschiare via
scrape away
'to scrape away/off'

b. lavare via
wash away
'to remove by washing'

c. grattare via
scrape away
'to scrape away/off'

Also LVCs shows semi-productive patterns. [25] attempted to define a set of semantic constraints to predict which verb can appear in the *HAVE a V* frame. [4] analysed four constructions

in Spanish composed of a verb that means *become*, an animated subject and an adjective. They discovered that adjectives appearing in these constructions can be divided into semantic clusters centered around a high-frequency exemplar.

Even some VIDs are argued to be the result of productive, yet very restricted, patterns. For instance, in the Lakoff-Johnson theory of metaphor [11] some idiomatic expressions are derived from conventional metaphors [10]. For example, English idioms that mean *revealing a secret* (2) can be thought of as a product of the conventional metaphor *the mind is a container and ideas are entities* [7].

(2) a. spill the beans
b. let the cat out of the bag
c. blow the whistle
d. blow the lid off
e. loose lips

3 Experimental setup

Our hypothesis is that MTLB-STRUCT [23] is able to capture, to some extent, those novel VMWEs that results from semi-productive patterns already experienced by the model in the training data. Vice versa, we argue that it performs worse with truly idiosyncratic VMWEs.

In this experiment, we employed the 14 corpora provided in the edition 1.2 of the PARSEME shared task [18] and the system results of MTLB-STRUCT[3].

We modeled an *MWE type* as a multi-set of unordered lemmas. For example the idiom *take the bull by the horns* would be represented as:

```
{bull, by, horns, take, theX2}
```

To obtain our data, we extracted all the MWE types from the test corpus that were never seen by the model during training (UNS) i.e. those MWE types that were not in the training nor in the development corpus. Then, from this list, we kept only the *discovered* MWE types (DIS). We considered an unseen MWE to be *discovered* if it has been correctly identified by the algorithm at least one time.

To check our hypothesis, we compared the distribution of VMWE classes in UNS with the distribution in DIS. If our hypothesis is true, we expect VMWE classes that are more productive (VPCs and LVCs) to show a higher percentage in DIS. Conversely, we expect less productive VMWE classes (VIDs) to show a higher percentage in UNS.

Furthermore, since in different languages certain MWE types are more productive than others, results should vary according to the language. Therefore, we also analysed the distribution of VMWE classes across languages.

4 Results

In Table 1 we compared the distribution of VMWE classes in DIS with the distribution in UNS while Table 2 shows the percentage of discovered VMWE per class.

As we can observe, the percentage of LVCs is higher in DIS than in UNS for all the languages except Romanian and Swedish. The higher percentages of discovered LVCs belong to French (80%) and Hindi (72%) while Chinese (17%) and Hebrew (23%) have the lower rates. It is not

[3] https://gitlab.com/parseme/sharedtask-data/-/tree/master/1.2

		LVC	VID	IRV	VPC	MVC	IAV	LS.ICV	TOT.
DE	DIS	18%	43%	4%	45%	-	-	-	148
	UNS	15%	46%	5%	35%	-	-	-	296
EL	DIS	60%	37%	-	3%	-	-	-	144
	UNS	49%	49%	-	2%	-	-	-	298
EU	DIS	80%	21%	-	-	-	-	-	127
	UNS	69%	32%	-	-	-	-	-	257
FR	DIS	76%	15%	10%	-	0%	-	-	136
	UNS	61%	28%	11%	-	0%	-	-	291
GA	DIS	80%	3%	2%	2%	-	14%	-	59
	UNS	53%	18%	1%	9%	-	18%	-	290
HE	DIS	56%	37%	-	8%	-	-	-	52
	UNS	42%	47%	-	11%	-	-	-	304
HI	DIS	78%	1%	-	-	21%	-	-	165
	UNS	68%	12%	-	-	20%	-	-	261
IT	DIS	34%	43%	16%	2%	2%	2%	2%	67
	UNS	23%	57%	12%	1%	3%	3%	2%	286
PL	DIS	71%	9%	21%	-	-	-	-	126
	UNS	64%	21%	16%	-	-	-	-	291
PT	DIS	79%	14%	7%	-	0%	-	-	122
	UNS	65%	26%	8%	-	0%	-	-	298
RO	DIS	5%	28%	67%	-	-	-	-	121
	UNS	6%	43%	52%	-	-	-	-	212
SV	DIS	12%	13%	5%	70%	-	-	-	154
	UNS	18%	24%	5%	53%	-	-	-	282
TR	DIS	57%	43%	-	-	-	-	-	135
	UNS	45%	55%	-	-	-	-	-	289
ZH	DIS	22%	6%	-	34%	38%	-	-	145
	UNS	21%	18%	-	31%	29%	-	-	282

Table 1. VMWE distributions in discovered and unseen.

		DE	EL	EU	FR	GA	HE	HI	IT	PL	PT	RO	SV	TR	ZH
LVC	discov.	59%	59%	57%	**80**%	31%	23%	**72**%	35%	48%	49%	50%	37%	58%	17%
	TOT.	44	147	176	167	154	128	178	66	185	195	12	52	130	34
VID	discov..	36%	36%	32%	25%	4%	13%	7%	18%	18%	22%	**38**%	30%	37%	17%
	TOT.	135	146	81	80	53	144	31	163	60	77	90	67	159	52
IRV	discov.	40%	-	-	39%	33%	-	-	33%	57%	36%	**74**%	54%	-	60%
	TOT.	15	-	-	33	3	-	-	33	46	25	110	13	-	53
VPC	discov.	**66**%	80%	-	-	4%	13%	-	11%	-	-	-	**72**%	-	70%
	TOT.	102	5	-	-	27	32	-	3	-	-	-	150	-	58
MVC	discov.	-	-	-	0%	-	-	**67**%	14%	-	0%	-	-	-	-
	TOT.	-	-	-	1	-	-	52	9	-	1	-	-	-	-
IAV	discov.	-	-	-	-	15%	-	-	20%	-	-	-	-	-	-
	TOT.	-	-	-	-	53	-	-	7	-	-	-	-	-	-
LS.ICV	discov.	-	-	-	-	-	-	-	33%	-	-	-	-	-	-
	TOT.	-	-	-	-	-	-	-	5	-	-	-	-	-	-

Table 2. % of discovered VMWE for each class

possible to conclude that LVCs are more productive in French than in other languages because on a closer analysis many of the MWEs discovered are medical terms (e.g. *bénéficier de angioplastie*, *subir angiographie*). So this result may be due to a different definition of LVCs or to a peculiar corpus composition.

Also the percentage of VPCs is higher in DIS than in UNS for all the languages but Irish and Hebrew, and Germanic languages, in which VPCs are very productive, show the highest rate of discovered VPCs.[4] (72% for Swedish and 66% for German).

Conversely, the rate of VIDs is lower in DIS than in UNS for all the 14 languages. Moreover the percentage of discovered VIDs never exceed the 40%. The lowest rate of discovered VIDs is that of Irish (4%) while Romanian has the higher rate (38%).

Finally, the percentage of IRVs and MVCs tend to be higher in DIS, while IAVs are more present in UNS. However, the interpretation of these latter results is beyond the scope of this paper.

5 Discussion

As we expected, MTLB-STRUCT discovers more easily unseen LVCs and VPCs (that are generally more productive) while it struggles to discover unseen VIDs. Thus, the results obtained corroborate our initial hypothesis. However, it must be considered that the data is fairly noisy. At a closer examination, we found lemmatization errors and misspellings that led to unseen MWEs that were actually seen.

Another issue concerns differences in the annotation process among languages. For instance, the notable result obtained with French LVCs is probably due to a different definition of LVC. We found out that a lot of unseen LVCs in the French data are composed of a light verb and a disease or a surgery name while we were not able to find such constructions in Italian.

5.1 Italian and French verbal idioms

In order to acknowledge which kind of VIDs the model is able to discover and which not, we examined Italian and French data. Unfortunately a true evaluation is not feasible since the correctly discovered VIDs are too few (29 for French and 20 for Italian). However we still observed some patterns.

First of all, we observed that some discovered Italian VIDs (3b) are in fact just aspectual variants of seen VIDs (3a) and some others (3d) are lexical variants (3c). We argue that such constructions must not be considered unseen MWEs but just variants of seen MWEs in order to distinguish the ability of the system to generalise over variants from its ability to discover novel MWEs.

(3) a. mettere a disposizione
put at disposal
'to make available'

b. essere a disposizione
be at disposal
'to be available'

c. finire nel dimenticatoio
finish into oblivion
'to be forgotten'

[4] Greek has too few examples to be considered

d. andare nel dimenticatoio
go into oblivion
'to be forgotten'

The same happens for French, some discovered VIDs (4a) are aspectual variants of seen VIDs (4b).

(4) a. donner le jour
give the day
'give birth'

b. voir le jour
see the day
'be created'

Furthermore, we found also Italian discovered VIDs (5a) that were the product of conventional metaphors already experienced by the model (5b, 5c, 5d). For instance, the examples in (5) can be thought as product of the metaphor *time is a resource* [11].

(5) a. rubare il tempo
steal the time
'take up time'

b. avere tempo
have time
'to have time'

c. perdere tempo
lost time
'to waste time'

d. prendere tempo
take time
'to take time'

Finally, there were discovered VIDs for which we were unable to find related expression among seen VIDs (e.g. *essere sulla bocca di tutti*). Such expressions may have been identified thanks to the ability of BERT of capturing the idiomatic key [15].

6 Conclusions

We observed that MTLB-STRUCT tends to discover more LVCs and VPCs, than VIDs. Since a lot of LVCs and VPCs result from semi-productive patterns while VIDs are more idiosyncratic, the results suggest that MWE identification systems based on BERT are able to capture, to some extent, semi-productive patterns that generate MWEs. Vice versa, they find it harder to discover idiosyncratic MWEs.

However, the data used in the experiment is too noisy to make strong assertions. A further experiment might consist in testing the system on artificially selected sentences.

In this way it should be possible to have more control over intervening factors. For example, it could be tested whether the system suffers from overgeneration by submitting to it incorrect MWEs that are similar to seen MWEs.

References

1. Al Saied, H., Candito, M., Constant, M.: The atilf-llf system for parseme shared task: A transition-based verbal multiword expression tagger. In: The European Chapter of the Association for Computational Linguistics EACL 2017. pp. 127–132 (2017)
2. Booij, G.: Constructional idioms, morphology, and the dutch lexicon. Journal of Germanic linguistics **14**(4), 301–329 (2002)
3. Boroş, T., Burtica, R.: Gbd-ner at parseme shared task 2018: Multi-word expression detection using bidirectional long-short-term memory networks and graph-based decoding. In: Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018). pp. 254–260 (2018)
4. Bybee, J., Eddington, D.: A usage-based approach to spanish verbs of'becoming'. Language pp. 323–355 (2006)
5. Constant, M., Eryiğit, 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. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186, (2019)
7. Gibbs Jr, R.W., O'Brien, J.E.: Idioms and mental imagery: The metaphorical motivation for idiomatic meaning. Cognition **36**(1), 35–68 (1990)
8. Klyueva, N., Doucet, A., Straka, M.: Neural networks for multi-word expression detection. In: Proceedings of the 13th workshop on multiword expressions (MWE 2017). pp. 60–65. Association for Computational Linguistics (2017)
9. Kurfali, M.: Travis at parseme shared task 2020: How good is (m) bert at seeing the unseen? In: International Conference on Computational Linguistics (COLING), Barcelona, Spain (Online), December 13, 2020. pp. 136–141 (2020)
10. Lakoff, G.: Women, fire, and dangerous things: What categories reveal about the mind. University of Chicago press (1987)
11. Lakoff, G., Johnson, M.: The metaphorical structure of the human conceptual system. Cognitive science **4**(2), 195–208 (1980)
12. Maldonado, A., Han, L., Moreau, E., Alsulaimani, A., Chowdhury, K., Vogel, C., Liu, Q.: Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In: Proceedings of the 13th Workshop on multiword expressions (MWE 2017). pp. 114–120. Association for Computational Linguistics (2017)
13. Masini, F.: Multi-word expressions between syntax and the lexicon: The case of italian verb-particle constructions. SKY Journal of Linguistics **18**(2005), 145–173 (2005)
14. Moreau, E., Alsulaimani, A., Maldonado, A., Vogel, C.: Crf-seq and crf-deptree at parseme shared task 2018: Detecting verbal mwes using sequential and dependency-based approaches. In: Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018) at the 27th International Conference on Computational Linguistics (COLING 2018). pp. 241–247 (2018)
15. Nedumpozhimana, V., Kelleher, J.: Finding BERT's idiomatic key. In: Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021). pp. 57–62. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.mwe-1.7, https://aclanthology.org/2021.mwe-1.7
16. Pasquer, C., Savary, A., Ramisch, C., Antoine, J.Y.: Seen2unseen at parseme shared task 2020: All roads do not lead to unseen verb-noun vmwes. In: Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons. pp. 124–129 (2020)
17. Ramisch, C., Cordeiro, S., Savary, A., Vincze, V., Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., 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. Association for Computational Linguistics (2018)

18. Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Mititelu, V.B., Bhatia, A., Iñurrieta, U., Giouli, V., et al.: Edition 1.2 of the parseme shared task on semi-supervised identification of verbal multiword expressions. In: Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons. pp. 107–118 (2020)
19. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for nlp. In: International conference on intelligent text processing and computational linguistics. pp. 1–15. Springer (2002)
20. Savary, A., Cordeiro, S., Ramisch, C.: Without lexicons, multiword expression identification will never fly: A position statement. In: Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019) (2019)
21. Savary, A., Ramisch, C., Cordeiro, S.R., Sangati, F., Vincze, V., Qasemi Zadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., et al.: The parseme shared task on automatic identification of verbal multiword expressions. In: The 13th Workshop on Multiword Expression at EACL. pp. 31–47 (2017)
22. 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). pp. 268–274 (2018)
23. Taslimipoor, S., Bahaadini, S., Kochmar, E.: Mtlb-struct@ parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. arXiv preprint arXiv:2011.02541 (2020)
24. Taslimipoor, S., Rohanian, O.: Shoma at parseme shared task on automatic identification of vmwes: Neural multiword expression tagging with high generalisation. arXiv preprint arXiv:1809.03056 (2018)
25. Wierzbicka, A.: Why can you have a drink when you can't* have an eat? Language pp. 753–799 (1982)

A corpus-based analysis of mediation in EU multi-word organization names

Fernando Sánchez Rodas[1][0000-0003-4244-1835]

[1] Universidad de Málaga, Spain
fersanchez@uma.es

Abstract. This study aims at using Named Entity Recognition (NER) to extract a specific type of multi-word entity, that is, multi-word organization names (MWORGs), from an English-Spanish comparable corpus of European Parliament documents. Following a triadic, Peircean model of translation and grammar, we hypothesize that MWORGs are nominal constructions (or signs) which serve a semiotic function of mediation in EU translations (Stecconi 2009; Torres-Martínez 2022). Specific performance of the VIP-DeepPavlov NER system (Corpas Pastor 2021) with MWORGs is evaluated in terms of precision, recall, and F-1 scores. Relevant MWORGs are then annotated and analyzed from a contrastive, semi-constructional approach (Boas 2010) to determine how many of them are mediating, and under which schemata. Results predictably show that non-mediating constructions are prevalent in non-translated English (66 %), as mediating constructions are in translated Spanish (81 %). However, a surprising 34 % of the organization names in non-translated English are mediating; inversely, 19 % of the MWORGs in translated Spanish serve a non-mediating function. Seven different mediation schemes (blending, borrowing, translation, and further combinations of the three) where discovered among MWORGs, some of them language-preferent. This reinforces our belief that names are largely disregarded semiotic hubs, and indeed a crucial piece in the understanding of (non-)translations and (non-)interpretations as construction-based grammars with a specific number of *similar*, *different*, and *mediating* rules in each language and textual typology.

Keywords: Construction Grammar, mediation, multi-word entities, multi-word organization names, named entity recognition

1 Introduction

Multi-word expressions or units (MWUs) can be very roughly defined as "idiosyncratic interpretations that cross word boundaries (or spaces)" (Jackendoff 1997:156). Common properties of MWUs are cross-lingual variation, single-word paraphrasability, proverbiality, and prosody (Baldwin and Nam Kim 2010). The major NLP tasks relating to MWUs are (1) identifying and extracting MWUs from corpus data and disambiguating their internal syntax, and (2) interpreting MWUs. Steadily, the completion and evaluation of such tasks have fostered the celebration of dedicated workshops (Corpas Pastor et al. 2016) and the publication of subsequent volumes (Mitkov et al. 2018), and

they have been pipelined with parsers and applications such as word alignment (Hatami, Mitkov and Corpas 2021).

In terms of token frequency, type frequency, and their occurrence in the world's languages, nominal MWUs are one of the most common MWU types (Lieber and Štekauer 2009). Nominal MWUs usually include a head noun plus one or more modifiers, and their general form and function depends on the language (compare the English *computer science department* with the Spanish *vistas al mar*). For some authors such as Sag et al. (2002), proper names are often semi-fixed nominal MWUs with a highly idiosyncratic syntax. US sport team names, for example, are canonically made up of a place or organization name (possibly a MWU itself, such as *San Francisco*) and an appellation that locates the team uniquely within the sport (such as *49ers*).[1] Resorting again to common NLP terminology (cf. Nouvel, Ehrmann and Rouset 2016), we could refer to these units as multi-word (named) entities, or MWNEs.

Multi-word entities are crucial bearers of information and often provide answers to major text understanding questions; however, work focusing exclusively on MWNEs is of recent prominence in NLP literature. In the task of Named Entity Classification (NEC), Chesney et al. (2017) described an approach for the distantly supervised classification of millions of existing multi-word entities into thirteen category types. In a similar vein, Jacquet, Piskorski, and Chesney (2019) presented several experiments on the construction of fine-grained and out-of-context multi-word entity classification models, exploiting a large BabelNet-derived multilingual Named Entity corpus of 49 languages from 7 different scripts. Following a different direction but also relying on BabelNet, Jacquet et al. (2019a) addressed large-scale multilingual multi-word entity recognition and variant matching in 22 languages. Results were later re-used in Jacquet et al. (2019b) to propose a hybrid system for named-entity recognition, lemmatization, and cross-lingual linking in Slavic languages.

The present study aims at applying Named Entity Recognition (NER) to a specific type of MWNE, that is, multi-word organization names (MWORGs) extracted from an English-Spanish comparable corpora of European Parliament documents. Specific performance of the chosen NER system when dealing with MWORGs will be evaluated attending to the common metrics (precision, recall, and F-1 score). Then, the resulting MWORGs will be analyzed from a contrastive, semi-constructional approach to determine how many of them serve a mediating function in translation, and under which forms or schemata.

---

[1] Sag et al. (2002:5-6) add that such names pose several obstacles for NLP primarily in the form of elision (*the (San Francisco) 49ers)* and determiner selection (*an/the* [[*(Oakland) Raiders*] *player*] or coordination (*the* [*Raiders and 49ers*]), combined with alternative lexicalization (*the San Francisco 49ers, those San Francisco 49ers, San Francisco 49ers, the 49ers, those 49ers,* and *49ers*), internal modification (*the league-leading (San Francisco) 49ers*), and gross overgeneration *(the Oakland 49ers)*.

2 Methodology

2.1 Corpus Selection

We selected the non-translated English (N-T_EN) and translated Spanish (T_ES) subcorpora of PETIMOD 3.0 as the material for NER. PETIMOD 3.0 is an intermodal[2] English <> Spanish corpus of non-mediated (non-translated, non-interpreted) and mediated (translated, interpreted) texts from the Committee on Petitions of the European Parliament (the PETI Committee). Sections of this corpus (2,348,774 tokens and 2,194 documents in total) were successfully applied to previous NLP-enhanced, contrastive translation and interpreting analyses including shifts (Corpas Pastor and Sánchez Rodas, forthcoming), verbal patterns (Corpas Pastor and Sánchez Rodas, in press; Sánchez Rodas, in press) and argument-structure constructions (Sánchez Rodas and Corpas Pastor, in preparation). When confronted, PETIMOD 3.0 N-T_EN and T_ES form a bilingual comparable corpus, as text distribution is asymmetric (not every English text has an available Spanish translation, and the other way around). Table 1 summarizes the main features of the two subcorpora as measured by ReCor (Corpas Pastor and Seghiri 2007).

Table 1. Total size of PETIMOD N-T_EN and T_ES by no. of documents, types, and tokens

PETIMOD 3.0 corpus (N-T_EN + T_ES)			
Subcorpus	**Documents**	**Types**	**Tokens**
N-T_EN	880	29,273	1,159,248
T_ES	716	31,528	1,059,346
TOTAL	**1,596**	**60,801**	**2,218,594**

2.2 Multi-Word Named Entity Recognition

After uploading N-T_EN and T_ES to the VIP platform, we performed an automatic NER in each of the subcorpora, filtering the results by the label ORG (companies, agencies, institutions, etc.). Although VIP offers two libraries for NER (DeepPavlov and SpaCy), we chose DeepPavlov as it slightly outperformed the metrics of SpaCy in previous evaluations (see above-cited papers for details). Raw results were first exported to an Excel file, where multi-word candidates were filtered with the help of a word-

[2] See Bernardini (2016) for a detailed definition and an account on intermodal corpora.

counting function introduced with Kutools.[3] The numbers of single-word (SWORG) and multi-word (MWORG) organization candidates extracted from each subcorpus are presented in Table 2 below.

Table 2. Single-word (SWORG) and multi-word (MWORG) organization candidates extracted from the bilingual corpus

	N-T_EN	**T_ES**	**TOTAL**
Extracted SWORG candidates	147	133	207
Extracted MWORG candidates	475	420	659
Extracted ORG candidates	*622*	*553*	*866*

2.3 Precision, recall and F-1 scores

Specific precision, recall, and F-1 scores for the MWORG recognition performance of DeepPavlov in each subcorpus were then calculated. Precision was very fine-grained; errors comprised instances which were wrongly recognized as a MWNE (e.g., *Emergency Assistance*), and MWNEs wrongly classified as MWORGs (e.g., *Habitats Directive, Autonomous Emergency Braking Systems*). We corrected and deemed relevant MWORGs with slight transcription mistakes, as well as with undefined determiners (*a Court of Justice*).

Specific recall scores divided the automatically extracted relevant MWORGs between the automatically plus manually extracted relevant MWORGs in each corpus. For the manual MWORG extraction, we resorted to frequency lists generated with VIP for each subcorpus, accounting for the top 100 nouns in English and Spanish. This method, already tested with Sketch Engine in previous studies (Corpas Pastor and Sánchez Rodas, in press), allows for quick detection of key nominal headers (*association, comité*, etc.) and effective access to complete MWNEs by using simple corpus functionalities such as case sensitiveness. Overall scores for both corpora are displayed in Table 3.

[3] Kutools is an Excel add-ins collection with a downloadable 30-day free trial (https://www.extendoffice.com/download/kutools-for-excel.html). It includes a formula helper to automatically introduce complex functions for word counting. The formula deducts the total number of words in a cell by counting spaces, which sometimes require manual revision for an accurate result.

Table 3. Performance scores of MWORG recognition in N-T_EN and T_ES corpora

Subcorpus	Precision	Recall	F-1 score	Relevant entities
N-T_EN	0.40	0.79	0.53	243
T_ES	0.43	0.63	0.51	287

2.4 Annotation of relevant MWORGs

The last step prior to analysis was the semi-automatic annotation of the obtained MWORGs in separate Excel sheets. Four parameters were observed and catalogued: word counting (using Kutools again), ORG sub-labelling,[4] language(s)[5] and mediating function/scheme (if any). Even when this paper focuses on the last type of data and employs the other three auxiliary, we expect to take great advantage of these annotation techniques in further publications.

3 Mediation Analysis

The analysis in this paper follows a contrastive, semi-constructional approach (Boas 2010) with a semiotic twist. The analysis does not delve into full linguistic details of the form-meaning pairings posited by the construction of multi-word organization names; however, it does employ abstract schemes to quantify the third logico-semiotic condition of translation along with *similarity* and *difference*, that is, *mediation*.[6] In words of Stecconi (2009:263), "it is logically impossible to label as translation a text that is not perceived as speaking on behalf of another – i.e. that does not mediate between source and target environments." This triadic, Peircean assumption of translation can easily lean on Construction Grammar, provided that we also follow a model of construction as a sign, in which form and function are inextricably bound up with agency (Torres-Martínez 2022).

To classify the spotted mediating schemes, we have adopted consolidated denominations from both translation (*translation* and *borrowing*) and cognitive studies (*blending*). While the first two are self-explanatory, blended constructions intertwine native and foreign units, each one serving a specific function or nest in the constructional

[4] Each multi-word entity was manually sub-indexed as pertaining to a certain subtype of ORG, taking the description in VIP as a reference: companies (ORG_{CO}), agencies (ORG_{AGCY}) and institutions (ORG_{INST}).

[5] Languages in each MWNE were notated by strict syntactic order (e.g., *Greenalia Biomass Power Curtis-Teixeiro S.LU.* was entered as ENGLISH + GALICIAN + SPANISH). As it can be also noticed in Footnote 4, the general goal was to follow a system akin to Goldbergian Construction Grammar notation.

[6] To set apart a translation from a non-translation, similarity, difference, and mediation must be jointly met (*foundation*), along with given translation *events* and *norms* in the space-time continuum (Stecconi 2009).

scheme (e.g., *Bundestag alemán*). At the same time, the three basic categories can combine with each other, forming more complex constructions with different functions, normally signaled by brackets or commas in the texts. An example is *SALA (Sewerage Board of Limassol Amathountos)*, which brings together a Greek acronym plus its English translation. Finally, non-mediating MWORGs are also counted. This category includes entities which are already assimilated into the constructicon of the target corpus. Examples are long-established national institutions (e.g., *National Health Service*, *Comisión Nacional del Mercado de Valores de España*) and, in the case of non-translated English, fixed international and EU organization names which might be a result of standardization and/or repeated translation over time (e.g., *the International Civil Aviation Organisation*, *European Economic and Social Committee*). Table 4 lists all mediating and non-mediating constructional types with one example per corpus, while Table 5 presents the quantitative results of our analysis.

Table 4. Examples of mediating and non-mediating schemes in PETIMOD (N-T_EN + T_ES).

	N-T_EN	T_ES
BLENDING	*the 'Petón do Lobo' Environmental Association*	*AG de Frankfurt*
BLENDING + (BORROWING)	-	*la Conselleria de Agricultura, Desarrollo Rural, Emergencia Climática y Transición Ecológica (Generalitat Valenciana)*
BORROWING	*Sociedad de Agricultores de la Vega*	*Bulgarian Helsinki Committee*
BORROWING + BORROWING	*Air France-KLM*	*Pannon GSM*
BORROWING + (TRANSLATION)	*Associazione Autisti Soccorritori Italiani (Association of Italian Road Rescue)*	*Comitato volontario dei cittadini contro la discarica nell'ex cava Viti (Comité voluntario de ciudadanos contra el vertedero en la antigua cantera Viti)*
TRANSLATION	*Basque Department of Health*	*Parlamento de Baja Sajonia*
TRANSLATION + (BORROWING)	*Netherlands Institute for Human Rights (College voor de Rechten van de Mens)*	*banco del Estado de Luxemburgo (BCEE)*
NON-MEDIATING CONSTRUCTIONS	*British Airlines*	*el Tribunal Superior de Justicia de Andalucía*

Table 5. Quantified mediating and non-mediating schemes in PETIMOD (N-T_EN + T_ES).

	N-T_EN	T_ES
MEDIATING CONSTRUCTIONS	**83**	**231**
BLENDING	7	15
BLENDING + (BORROWING)	0	1
BORROWING	28	36

BORROWING + BORROWING	1	4
BORROWING + (TRANSLATION)	3	1
TRANSLATION	42	166
TRANSLATION + (BORROWING)	2	8
NON-MEDIATING CONSTRUCTIONS	**160**	**56**
TOTAL	***243***	***287***

The results predictably show that non-mediating constructions are dominant in the non-translated corpus, as it happens with mediating constructions in translated Spanish. Nevertheless, numbers also present an interesting 34 % (83 out of 243) of mediating constructions in N-T_EN. It does not function the other way around, as the percentage of non-mediating constructions in T_ES falls to 19 % (56 out of 287). This could indicate that, at least in the case of EU-related MWORGs, English is more successful in the integration of foreign MWNEs than Spanish.

As we look deeper into the mediation schemes of both languages, more differences emerge. Translated Spanish outnumbers in all categories because for obvious reasons (it has the highest amount of MWORGs in general and mediating MWORGs in particular), except for one. In BORROWING + (TRANSLATION), results are relatively better for English (3) than for Spanish (1). This allegedly means that English prefers to introduce the nominal construction with the foreign ORG, thus giving an extra topical relevance to it; consider for example *Audiencia Provincial (Court of appeal)*. On the contrary, Spanish favors TRANSLATION + (BORROWING), even when the involved foreign language is also Romance, as in *Ministerio de Medio Ambiente, Protección del Territorio y Medio Marino (Ministero dell'Ambiente e della Tutela del Territorio e del Mare)*. Again, this could imply that English is more diplomatic when dealing with such foreign, EU-related MWORGs than Spanish.

To wrap up, some other indicators are also worth a comment. TRANSLATION is the preferred mediation scheme in both languages, and it is surprising how the non-translated corpus contains an impressive 17 % of MWORGs (42 out of 243) that are clear English versions of foreign organization names, which often provide the same translation in their websites or documentation.[7] BORROWINGS such as *Tribunal Central Administrativo Sul* and *Radlberger Getränkegesellschaft* are second in importance, while BORROWING + BORROWING is a less common category which mingles relatively minor languages and is conformed almost exclusively by company names (*Air France-KLM, Pannon GSM*). Lastly, the absence of BLENDING + (BORROWING) instances in the English corpora contrasts with one single example in translated Spanish: *la Conselleria de Agricultura, Desarrollo Rural, Emergencia Climática y Transición Ecológica (Generalitat Valenciana)*. Here, the first MWNE is a blend of Spanish + Valencian (*la Conselleria*…), while the

[7] For example, *French Agency for Food, Environmental and Occupational Health & Safety* is the official translation of ANSES (*L'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail*) used in their English website (https://www.anses.fr/en/content/our-identity).

bracketed ORG is completely in Valencian.[8] Although in this case numbers are too low to extract solid conclusions, the sole presence of this category in Spanish seems to compensate its above-perceived "lack of inclusion" by diversifying its mediation schemes with Valencian, a co-official language of Spain.[9]

4 Conclusions

This paper has skimmed a handcrafted English-Spanish corpus of EU texts (2,218,594 tokens) in search for multi-word organization names (MWORGs) with the help of a pre-trained NER system. On the one hand, F-1 scores are above average in both cases (0.53 for English and 0.51 for Spanish). These numbers however show that there is still a long way to go for the recognition of multi-word entities, especially when they are class specific like MWORGs. On the other hand, their subsequent analysis has unveiled the existence of repeated schemes of mediation in the institutional usage of MWNEs, both in translated and non-translated texts. This reinforces our belief that names are largely disregarded semiotic hubs, and indeed a crucial piece in the understanding of (non-)translations and (non-)interpretations as construction-based grammars with a specific number of similar, different, and mediating rules in each language and textual typology.

In the future, more steps need to be taken in the refinement of NER systems for MWNE extraction. One possibility is to compare the performance of generally trained NER systems (e.g., VIP) with specific ones (e.g., JRC-Names), or to improve manual extraction by using domain-specific term lists instead of frequency noun lists. Such improvements, together with a deeper (not only schematic) construction-based analysis of multi-word entities with semi-automatically and/or manually annotated information, would make it possible to better answer another conundrum of this study; which semiotic rules govern EU (non-)translations and (non-)interpretations beyond mediation, or what is *similar* and what is *different* in institutional (non-)translations? With such a full picture, we could continue producing rich datasets reusable for training NLP applications like text generation and/or summarization systems and, naturally, machine translation and interpreting in EU settings.

Funding

This work was supported by the Spanish Ministry of Education and Professional Training under Grant FPU18/05803. It has also been carried out in the framework of the projects VIP II (PID2020-112818GB-I00), TRIAGE (UMA18-FEDERJA-067), E3/04/21 and MI4ALL (UMA-CEIATECH-04).

[8] The entire syntactic sequence would be SPANISH + VALENCIAN + SPANISH + (VALENCIAN).

[9] Galician was also annotated in some entries of both subcorpora (see Footnote 5 and Table 4).

References

Baldwin, T., Nam Kim, S.: Multiword expressions. In: Indurkhya, N. Damerau, F. J. (eds.) Handbook of Natural Language Processing, pp. 267-292. CRC Press, Boca Raton (2010).

Bernardini, S.: Intermodal corpora: A novel resource for descriptive and applied translation studies. In: Corpas Pastor, G., Seghiri, M. (eds.) Corpus-based Approaches to Translation and Interpreting: From Theory to Applications, pp. 129–148. Peter Lang, Frankfurt (2016).

Boas, Hans C. (ed.): Contrastive Studies in Construction Grammar. 1st edn. John Benjamins, Amsterdam/Philadelphia (2010).

Chesney, S., Jacquet, G., Steinberger, R., Piskorski, J. Multi-word Entity Classification in a Highly Multilingual Environment. In: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pp. 11–20. Association for Computational Linguistics, Stroudsburg PA (2017).

Corpas Pastor, G.: Technology Solutions for Interpreters: The VIP System. Hermēneus 23, 91-123 (2021).

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).

Corpas Pastor, G., Sánchez Rodas, F. EU Phraseological Verbal Patterns in the PETIMOD corpus: a NER-enhanced Approach. In: Biel, Ł., Kockaert, H. J. (eds.) Handbook of Legal Terminology. John Benjamins, Amsterdam/Philadelphia (In press).

Corpas Pastor, G., Sánchez Rodas, F. NLP-enhanced Shift Analysis of Named Entities in an English<>Spanish Intermodal Corpus of European Petitions. In: Kajzer-Wietrzny, M., Bernardini, S., Ferraresi, A., Ivaska, I. (eds.) Empirical Investigations into The Forms of Mediated Discourse at the European Parliament. Language Science Press, Berlin (Forthcoming).

Corpas Pastor, G., Monti, J., Seretan, V. Mitkov, R. (eds.): Workshop Proceedings Multi-Word Units in Machine Translation and Translation Technologies MUMTTT. Tradulex, Geneva (2016)

Hatami, A., Mitkov, R., Corpas Pastor, G.: Cross-Lingual Named Entity Recognition via FastAlign: a Case Study. In: Proceedings of the Translation and Interpreting Technology Online Conference, pp. 85–92. INCOMA Ltd., held online (2021).

Jackendoff, R.: The Architecture of the Language Faculty. MIT Press, Cambridge MA (1997).

Lieber, R., Štekauer, P. (eds.): The Oxford Handbook of Compounding. 1st edn. Oxford University Press, Oxford (2009).

Jacquet, G., Piskorski, J., Chesney, S. Out-of-context fine-grained multi-word entity classification: exploring token, character n-gram and NN-based models for multilingual entity classification. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), pp. 1001-1010. Association for Computing Machinery, New York (2019).

Jacquet, G., Ehrmann, M., Piskorski, J., Tanev, H., Steinberger, R.: Cross-lingual linking of multi-word entities and language-dependent learning of multi-word entity patterns. In: Parmentier, Y., Waszczuk, J. (eds.) Representation and parsing of multiword expressions: Current trends, pp. 269-297. Language Science Press (2019a).

Jacquet, G., Piskorski, J., Tanev, H., Steinberger, R.: JRC TMA-CC: Slavic named entity recognition and linking. participation in the BSNLP-2019 shared task. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pp. 100-104. Association for Computational Linguistics, Stroudsburg PA (2019b).

Mitkov, R., Monti, J., Corpas Pastor, G., Seretan, V. (eds.): Multiword Units in Machine Translation and Translation Technology. 1st edn. John Benjamins, Amsterdam/Philadelphia (2018).

Nouvel, D., Ehrmann, M., Rosset, S.: Named Entities for Computational Linguistics. Wiley, Hoboken NJ (2016).
Sag, I. A., 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 Intelligent Text Processing and Computational Linguistics (CICLing-2002), pp. 1–15. Mexico City (2002).
Sánchez Rodas, F.: Introducing the PETIMOD Corpus: A Resource for the Analysis of Personification in Mediated and Non-Mediated Discourse. In: Proceedings of 43rd Conference Translating and the Computer (TC43). Tradulex, Geneva (In press).
Sánchez Rodas, F., Corpas Pastor, G.: Transitive Constructions with Organization Names in the English Eurolect: The Case of [ORG + V + *that*] (In preparation).
Stecconi, U.: Semiotics. In: Baker, M. Saldanha, G. (eds.) Routledge Encyclopedia of Translation Studies, pp. 260-263. Routledge, London/New York (2009).
Torres-Martínez, S.: The role of semiotics in the unification of langue and parole: an *Agentive Cognitive Construction Grammar* approach to English modals. Semiotica 244, 195-225 (2022).

Transformer-based Detection of Multiword Expressions in Flower and Plant Names

[1]Damith Premasiri, [2]Amal Haddad Haddad, [1]Tharindu Ranasinghe, and [1]Ruslan Mitkov

[1]University of Wolverhampton, UK
[2]University of Granada, Spain
{damith.premasiri,tharindu.ranasinghe,R.Mitkov}@wlv.ac.uk
amalhaddad@ugr.es

Abstract. Multiword expression (MWE) is a sequence of words which collectively present a meaning which is not derived from its individual words. MWEs detection is an important topic in different natural language processing (NLP) applications, including machine translation. Therefore, detecting MWEs in different domains is an important research topic. With this study, we explore state-of-the-art neural transformers in the task of detecting MWEs in flower and plant names. We evaluate different transformer models on a dataset created from Encyclopedia of Plants and Flower. We empirically show that neural transformers could outperform previous neural models like long-short term memory (LSTM) even in specific domains such as flower and plant names.

Keywords: Multiword Expressions · Transformers · Deep Learning · Flowers · Plants

1 Introduction

The correct interpretation of Multiword Expressions (MWEs) is crucial to many natural language processing (NLP) applications but is challenging and complex. In recent years, the computational treatment of MWEs has received considerable attention, but there is much more to be done before one can claim that NLP and Machine Translation (MT) systems process MWEs successfully [12].

The study of multiword expressions in NLP has been gaining prominence, and in recent years the number of researchers and projects focusing on them has increased dramatically. The successful computational treatment of MWEs is essential for NLP, including MT and Translation Technology. The inability to detect MWEs automatically may result in incorrect (and even unfortunate) automatic translations and may jeopardise the performance of applications such as text summarisation and web search.

Multiword expressions do not only play a crucial role in the computational treatment of natural languages. Often terms are multiword expressions (and not single words), making them highly relevant to terminology. The requirement for correct rendering of MWEs in translation and interpretation highlights their

importance in these fields. Given the pervasive nature of MWEs, they play a crucial role in the work of lexicographers who study and describe both words and MWEs. Lastly, MWEs are vital in the study of language, which includes not only language learning, teaching and assessment but also more theoretical linguistic disciplines such as pragmatics, cognitive linguistics and construction grammar, which are nowadays aided by (and, in fact, often driven by) corpora. MWEs are very relevant for corpus linguists, too. As a result, MWEs provide an excellent basis for interdisciplinary research and for collaboration between researchers across different areas of study, which for the time being, is underexplored.

This study is concerned with developing and evaluating a methodology designed to identify multiword expressions among flower and plant names. Multiword expressions are common among the names of flowers and plants, as in the case of *Leontopodium alpinum, White Moonlight or Pink Shirley Alliance*. To the best of our knowledge, this is the first study covering this domain.

The rest of the paper is structured as follows. Section 2 outlines related work. Section 3 describes the dataset used for our experiments, while section 4 presents the methodology. Section 5 reports the evaluation results, and finally, section 6 summarises the conclusion of this study.

2 Related Work

Neural models are increasingly employed to detect MEWs. Since both MWEs detection and named entity recognition (NER) tasks are about token classification, they can be modelled using similar models. Therefore we are using a set of models which are used in NER for the MWE detection task, too [20].

LSTMs [9] and gated recurrent units (GRUs) [20] are the most popular deep learning methods which have been employed in MWEs detection task. There are methods where an LSTM network [9] is combined with a Conditional random field (CRF) for the same task. Furthermore, graph convolutional neural networks (GCNs) [10] have also been used for MWE identification. The performance of GCNs in MWEs detection has improved using multi-head self-attention [20]. Transformers have also been used in MWEs detection; [21, 24]; however, the research has been minimal and has focussed only on general domains. More specifically, there is limited research on MWEs detection tasks in specific fields, such as MWEs detection in flower and plant names. This paper tries to understand how the state-of-the-art transformer models work in MWEs detection in flower and plant names by empirically evaluating several transformer models.

3 Data

The subject of flowers and plants is of great interest to both professionals and the general public and is relevant to professionals in botany, phytotherapy, plant pharmacology, designers, etc. In addition, there is a lot of interest among the general public as many people dedicate their time to planting plants and growing them in their gardens and homes. Apart from that, the identification of names of

flowers and plants as terms is also relevant to terminologists and translators. The study of the names of plants as terms helps in laying the basis of term coining processes and gives insights into the underlying mechanisms of term creation. Translators also benefit from this information for its transfer between languages. For this reason, the automatic identification of names of flowers and plants is relevant to meeting all those needs.

The Encyclopaedia of Plants and Flowers[3] of The American Horticultural Society, edited by Christofer Brickel and published by Dorling Kindersley editorial, is used as the corpus for this study. This edition is available in a digitalised format in the online library of the Internet Archive. This encyclopaedia consists of 522,707 words. It contains a dictionary of names of flowers from around the world, with approximately 8000 terms referring to both scientific and common names and their origins, as well as 4000 images. It also contains descriptions of each flower, instructions on how to plant it and how to use the plants to design gardens. The dataset has been created by extracting the text from Encyclopedia of Plants and Flowers[3]. This encyclopedia consists of two main parts: the plant catalogue and the plant dictionary. The book describes the origins of plant names and basic gardening concepts apart from its main concerns. The plant catalogue describes different categories such as Trees, Shrubs, Roses, Perennials etc. In the remaining parts, the plants are subdivided into large, medium and small subdivisions Ex: large trees, and small trees. The plant dictionary is a dictionary which seeks to cover all possible plant names along with a short description and references to different sections in the plant catalogue accordingly. The flower and plant names in the dictionary are abbreviated similarly to an ordinary dictionary. The data was pre-processed by annotating the terms of proper names. The training and test data were created by combining both the plant catalogue and plant dictionary and pre-processing them and tagging the MWE according to the IOB format. The I, O, and B tags stand for Inside, Outside and Beginning, respectively, based on the MWE tag position of the word. The training set consists of 38,985 sentences, whilst the test set consists of 14,234 sentences.

4 Methodology

Transformers based models have produced state-of-the-art art results in many NLP tasks such as text classification [22, 18, 23], sense disambiguation [8, 7], question answering[15], machine translation[25, 16] and named entity recognition[17, 19]. Therefore, we employ transformers based models for MWEs detection task while comparing performance with Bidirectional LSTM model. We further discuss these models in the following sections.

Transformer models such as BERT [6] have been trained using masked language modelling objective and they can be fine-tuned for multiple different tasks [1]. Within this study we fine-tune number of transformer models for MWEs detection which is a token classification task. We modify the original BERT architecture by adding a token level classifier following the last hidden layer as shown

in Figure 1 to achieve the architecture of the MWE detection model. This is a linear classification layer which uses the last hidden state as input and output the relevant token per word such as B,I and O.

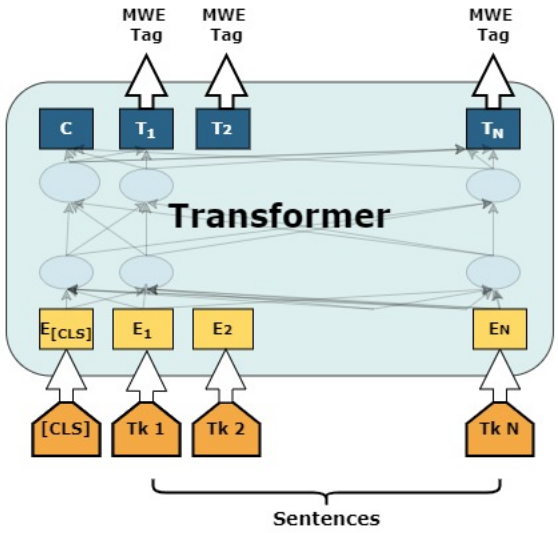


Fig. 1. MWEs token classification transformer architecture [14]

We conducted our experiments with many popular transformer models to detect MWEs such as BERT [6], RoBERTa [27], XLNet [26], XLM-RoBERTa [5] and Electra [4]. We evaluated several different BERT variations such as bert-base-cased, bert-base-uncased, bert-base-multilingual-cased, bert-base-multilingual-uncased. Furthermore we experimented sci-bert-cased[2] and sci-bert-uncased[2] which are pre-trained on scientific corpus as more specific variants of BERT[6]. The other transformer models are evaluated only on their base model. All the transformer-based methods made use of a batch size 32, Adam optimiser with learning rate 4e-5. They were trained for 3 epochs with linear learning rate warm-up over 10% of the training data. These experiments were carried out in an NVIDIA GeForce RTX 2070 GPU and in Google colab GPUs[1].

BiLSTM-CRF is another token classification architecture which provided state-of-the-art results before transformers [13] . Bidirectional LSTM (BiLSTM) is an improved version of conventional LSTMs which is capable of learning contextual information both forwards and backwards in time. Unlike standard LSTM, the input flows in both directions, and it's capable of utilizing information from both sides. It utilises an additional LSTM layer which reverses direction of the

[1] https://colab.research.google.com/

information flow. This study leverages the BiLSTM architecture given its' bidirectional ability to model temporal dependencies. CRFs [11] are a statistical model that are capable of incorporating context information and are highly used for sequence labelling tasks. The BiLSTM-CRF model provides a way of combining the relationships of consecutive outputs of the network and utilise both past and future tag information to for prediction. Since the BiLSTM has both of the past and future information, combining CRF on top of the BiLSTM provides powerful network for precisely predicting the tags. BiLSTM-CRF experiments were performed on a CPU with a learning rate of 1e-3 and the model was trained for 60 epochs.

5 Results

In this section, we report the results of conducted experiments using macro averaged F1 as it is widly used in classification tasks. As shown in Table 1 it is clear that the transformer-based models outperform the BiLSTM-CRF method with clear margins. The BiLSTM-CRF could achieve only 0.3320 Macro F1 score, while all the transformer models we experimented outperformed that. A noticeable observation is that best transformer model had nearly double the F1 score of BiLSTM-CRF model while all the other transformer models performed competitively.

Model	**Macro F1**
roberta-base	0.5039
xlm-roberta-base	0.5650
xlnet-base-cased	0.6312
bert-base-multilingual-cased	0.6086
bert-base-multilingual-uncased	**0.6422**
bert-base-cased	0.6393
bert-base-uncased	0.6227
electra-base-discriminator	0.5753
sci-bert-cased	0.6214
sci-bert-uncased	0.6307
BiLSTM-CRF	0.3320

Table 1. Results for multiword expression detection in flower and plant names

The clear winner is the bert-base-multilingual-uncased model with a Macro F1 score of 0.6422. This is followed by bert-base-cased and xlnet-base-cased models with Macro F1 scores of 0.6393 and 0.6312, respectively. In general, transformer models have better performance with slight margins among them. An

interesting observation is that multilingual-bert model outperforms sci-bert models, which are trained on a corpus featuring a high frequency of scientific terms. We conjecture that this could be due to a lack of the flower names and plant names related data in the sci-bert training set. Nevertheless, sci-bert-uncased model competitively performed with 0.6307 Macro F1 score, which is only a 0.0115 difference from the best performing model.

Another interesting observation was while the multi-lingual bert model was the best performer, the cross-lingual model; xlm-roberta-base did not do well with a F1 score of 0.5650. This score is very close to the least successful model among transformers which was roberta-base with Macro F1 of 0.5039. Yet these values are outperforming the BiLSTM-CRF model, which shows the powerful nature of transformers based models in MWE tasks over the other neural methods like BiLSTM.

Overall, transformers-based neural methods clearly perform better than BiLSTM-CRF. All the transformer-based methods performed above 0.5000 of Macro F1, showing their strong performance in MWE detection tasks.

6 Conclusion

MWE detection has significant importance in many NLP applications, especially in translation and terminology studies. In this paper, we focus on an empirical analysis of multiple neural transformer models in the MWE detection task using a flowers and plants dataset. We show that all transformer models outperform the LSTM-based method. Of the transformer models we experimented with, bert-base-multilingual-uncased reported the best results doing better than other transformer models. We can conclude that transformer models can handle the challenges presented by MWEs in local domains like plant names and flower names better than the previous neural methods, such as LSTM.

In the future, we would like to explore more specific domains similar to flower and plant names. It would be interesting to study how MWEs detection works in different languages with different flower and plant names. We are encouraged to explore cross-lingual models more in this regard to understand how well these models perform across languages on the MWEs detection task for similar datasets.

References

1. Alloatti, F., Di Caro, L., Sportelli, G.: Real life application of a question answering system using BERT language model. In: Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue. pp. 250–253. Association for Computational Linguistics, Stockholm, Sweden (Sep 2019). https://doi.org/10.18653/v1/W19-5930, https://aclanthology.org/W19-5930
2. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language

Processing (EMNLP-IJCNLP). pp. 3615–3620. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1371, https://aclanthology.org/D19-1371
3. Brickell, C.: Encyclopedia of plants and flowers. In: Encyclopedia of plants and flowers. Dorling Kindersley, Santa Fe, New Mexico, USA (2012)
4. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training text encoders as discriminators rather than generators. In: ICLR (2020), https://openreview.net/pdf?id=r1xMH1BtvB
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
7. Hettiarachchi, H., Ranasinghe, T.: BRUMS at SemEval-2020 task 3: Contextualised embeddings for predicting the (graded) effect of context in word similarity. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 142–149. International Committee for Computational Linguistics, Barcelona (online) (Dec 2020). https://doi.org/10.18653/v1/2020.semeval-1.16, https://aclanthology.org/2020.semeval-1.16
8. Hettiarachchi, H., Ranasinghe, T.: TransWiC at SemEval-2021 task 2: Transformer-based multilingual and cross-lingual word-in-context disambiguation. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). pp. 771–779. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.semeval-1.102, https://aclanthology.org/2021.semeval-1.102
9. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (11 1997). https://doi.org/10.1162/neco.1997.9.8.1735, https://doi.org/10.1162/neco.1997.9.8.1735
10. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
11. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. p. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
12. Monti, J., Seretan, V., Pastor, G.C., Mitkov, R.: Multiword expressions in machine translation and translation technology. In: Multiword Expressions in Machine Translation and Translation Technology. pp. 1–37. John Benjamin Publishers (2018)
13. Panchendrarajan, R., Amaresan, A.: Bidirectional LSTM-CRF for named entity recognition. In: Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation. Association for Computational Linguistics, Hong Kong (1–3 Dec 2018), https://aclanthology.org/Y18-1061

14. Premasiri, D., Ranasinghe, T.: Bert (s) to detect multiword expressions. In: International Conference 'Computational and Corpus-based Phraseology - Europhras 2022 (2022)
15. Premasiri, D., Ranasinghe, T., Zaghouani, W., Mitkov, R.: Dtw at qur'an qa 2022: Utilising transfer learning with transformers for question answering in a low-resource domain. In: Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5). (2022)
16. Ranasinghe, T., Orasan, C., Mitkov, R.: An exploratory analysis of multilingual word-level quality estimation with cross-lingual transformers. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 434–440. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-short.55, https://aclanthology.org/2021.acl-short.55
17. Ranasinghe, T., Sarkar, D., Zampieri, M., Ororbia, A.: WLV-RIT at SemEval-2021 task 5: A neural transformer framework for detecting toxic spans. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). pp. 833–840. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.semeval-1.111, https://aclanthology.org/2021.semeval-1.111
18. Ranasinghe, T., Zampieri, M.: Multilingual offensive language identification with cross-lingual embeddings. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5838–5844. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.470, https://aclanthology.org/2020.emnlp-main.470
19. Ranasinghe, T., Zampieri, M.: MUDES: Multilingual detection of offensive spans. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations. pp. 144–152. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-demos.17, https://aclanthology.org/2021.naacl-demos.17
20. Rohanian, O., Taslimipoor, S., Kouchaki, S., Ha, L.A., Mitkov, R.: Bridging the gap: Attending to discontinuity in identification of multiword expressions. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 2692–2698. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1275, https://aclanthology.org/N19-1275
21. Taslimipoor, S., Bahaadini, S., Kochmar, E.: MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In: Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons. pp. 142–148. Association for Computational Linguistics, online (Dec 2020), https://aclanthology.org/2020.mwe-1.19
22. Uyangodage, L., Ranasinghe, T., Hettiarachchi, H.: Can multilingual transformers fight the COVID-19 infodemic? In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). pp. 1432–1437. INCOMA Ltd., Held Online (Sep 2021), https://aclanthology.org/2021.ranlp-1.160
23. Uyangodage, L., Ranasinghe, T., Hettiarachchi, H.: Transformers to fight the COVID-19 infodemic. In: Proceedings of the Fourth Work-

shop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda. pp. 130–135. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.nlp4if-1.20, https://aclanthology.org/2021.nlp4if-1.20

24. Walsh, A., Lynn, T., Foster, J.: A bert's eye view: Identification of irish multiword expressions using pre-trained language models. In: Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022). European Language Resources Association (ELRA), Marseille, France (Jun 2022)
25. Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D.F., Chao, L.S.: Learning deep transformer models for machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1810–1822. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1176, https://aclanthology.org/P19-1176
26. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf
27. Zhuang, L., Wayne, L., Ya, S., Jun, Z.: A robustly optimized BERT pre-training approach with post-training. In: Proceedings of the 20th Chinese National Conference on Computational Linguistics. pp. 1218–1227. Chinese Information Processing Society of China, Huhhot, China (Aug 2021), https://aclanthology.org/2021.ccl-1.108

Searching for the Linguistically Indefinable: Automatic Extraction of Pragmatemes

Judyta Mężyk[1][0000-0003-1073-9119]

[1] Paris-East Créteil University, France, University of Silesia in Katowice, Poland
judyta.mezyk@us.edu.pl

Abstract. Although challenging, the research on multi-word units (MWUs) has grown significantly since Firth's famous 1950s publications on collocations. Since then, MWUs such as idioms, proverbs, lexical bundles, metaphors, and many others have been subjected to careful examination. However, pragmatemes (also called routine formulas, among other terms) seemed to be given much less attention until recently (Mel'čuk 2012). This paper aims to summarize the nature of pragmatemes in order to suggest a method of their automatic extraction from corpora. This is done on the example of a textual corpus of French captions. To extract the units, a list of pragmatemes in French presented by Blanco and Mejri (2018) is used along with Unitex, a corpus processing suite. The results are gathered and discussed with the potential search errors in mind. The research presented in this paper and the analysis of its strong and weak points provide a good starting point for the development of a reliable, automatic method of extracting MWUs as indefinable as pragmatemes.

Keywords: pragmatemes, multi-word units extraction, corpus linguistics

1 Introduction

The idea of examining the meaning of a word not in isolation but in the company of words that surround it is said to date back to the 1950s, when Firth first described the meanings created by associations of multiple, rather than individual, words (Firth, 1951, 1957). Later examinations done by means of corpus linguistics (among other subfields) have proved that words, in fact, do "often have a preference for what they combine with" (Greaves, Warren, 2022:212). Having realized the importance of multi-word units (MWUs, or multi-word expressions – MWEs, often used interchangeably), the research on their nature has long been a topic of interest of linguists (and not only) in many subfields.

According to Sag et al. (2002:1), MWEs are "idiosyncratic interpretations that cross word boundaries (or spaces)," while Finkbeiner and Schlücker (2019:3) describe them as "syntactic patterns consisting of at least two words, the combination of which may be more or less fixed, more or less idiomatic, and more or less productive." Although definitions may vary and go from broad to narrow, researchers generally agree that all MWUs can be characterized by their frequent co-occurrence patterns, lexical status similar to the one of single words (some scholars, e.g., Masini 2009, go as far as calling

MWUs "phrasal lexemes"), and prefabrication (Maziarz et al., 2022; Wray and Perkins, 2000). The main areas of focus in the MWUs research contain the examination of units such as idioms, lexical phrases, collocations, metaphors, lexical bundles, phrasal verbs, concgrams, proverbs, n-grams, and compounds (Wood, 2015).

However, the topic of the present paper, i.e., pragmatemes, is not listed by Wood (2015) as one of the main focus areas in the study of MWUs. Yet, as it will be presented, they are one of the types of MWUs, and therefore, the ability to use them in a spontaneous speech is a vital factor in language fluency (ibid). That is why, to shed more light on the issue, this paper will examine how pragmatemes can be studied using corpus linguistic software, specifically whether and how they can be automatically extracted from corpora. Before that, a closer explanation of this type of MWUs is provided in the following subsection.

1.1 How to define pragmatemes

Pragmatemes, according to Blanco (2013), are phrases (or, more rarely, a single lexemes) the meaning of which is limited by the communicative situation in which they are used. The first systematic research to call the linguistic phenomenon of using fixed expressions in particular situations of communication 'pragmatemes' is believed to be the one proposed by Mel'čuk in 1995. Mel'čuk describes the main unit of his theory, a phraseme, or 'phraseological expression', as a non-free phrase, meaning that "at least one of its lexical components L_i is selected by the speaker in a linguistically constrained way – that is, as a function of the lexical identity of other component(s)" (2012:33). Then, he divides phrasemes into lexical, e.g. *kick the bucket*, and semantic-lexical (clichés). According to Mel'čuk (2012:34), the latter type, in which both the components of its lexical expression and the components of its meaning "are selected by the Speaker in a constrained way," for instance *Happy birthday*, is the most frequent in the language and therefore is the most difficult to study, and the least researched so far (even "seriously understudied"). Semantic-lexical phrasemes are further divided into pragmatically non-constrained clichés (proverbs and complex proper names, e.g. *A watched pot never boils* and *Red Planet*) and pragmatically-constrained clichés, i.e., pragmatemes (such as the beforementioned *Happy Birthday*). Nevertheless, while Mel'čuk is believed to be the first researcher to give this linguistic phenomenon a new name, its existence was acknowledged long before his theory (e.g., Bally's 1909 description of fixed, situationally constrained units, and Coulmas's 1979 article on routine formulae, among others). In linguistic research, numerous terms are used to name language units of more or less similar definitions, such as situational phrases, situation-bound utterances, stereotyped language acts, pragmatic phraseologisms, and formulas, among others (Barnas 2017). While numerous researchers (e.g., Coulmas 1979, Bardovi-Harlig 2012, Chlebda 1993) acknowledge the existence of prefabricated utterances connected to specific situations of communication, there seems to be no agreement on the detailed characteristics of these units. For example, Kauffer (2019) argues that these units are non-compositional, while according to Blanco and Mejri (2018), it is the opposite, and pragmatemes are most often semantically compositional. In the visible plurality of definitions and the abundance of different terms, Fléchon, Frassi, and Polguère (2012)

even ask in the title of their paper *Les pragmatèmes, ont-ils un charme indéfinissable ?* ("Do pragmatemes have the charm of being indefinable?"[1]).

While it is necessary to acknowledge the existence of numerous theories on the topic, for the purpose of the present study, the definition of a pragmateme by Blanco and Mejri (2018) will be adopted. According to these researchers, a pragmateme is a polylexical, autonomous, semantically compositional utterance which is constrained by the situation of communication in which it is produced. Nonetheless, it seems crucial to stress that this definition only refers to "prototypical pragmatemes" (Blanco and Mejri, 2018). In their 2018 book devoted entirely to pragmatemes, Blanco and Mejri do not dismiss one-word pragmatemes such as *Bonjour* (Eng. "Hello"), although they stress that they are less common, or semantically idiomatic pragmatemes such as *Chaud devant* (lit. "Hot in the front," a French idiomatic phrase used by a waiter to warn others that they are carrying a hot dish).

2 Automatic extraction of pragmatemes from a corpus of captions

With numerous theories on the subject and without the possibility of a general linguistic deconstruction of pragmatemes (for instance, one can never say that pragmatemes are always composed of pronouns and verbs), their automatic extraction from corpora may cause problems for researchers. Yet, it is important, as pragmatemes, being present in typical situations of communication, constitute a big part of everyday language. That is why their careful examination is not to be disregarded, even if it may seem problematic. Corpus linguistics has provided researchers with tools that can answer questions of frequency, the scope of language use, and patterns. All of this information can be further used in, for instance, creating learning materials that would be based on real-world use of a language (Rogers et al., 2021). That is why an automatic method of extracting pragmatemes in corpora is suggested, with the example of a corpus of captions.

2.1 Methods

The corpus used for this study consists of French captions (i.e., intralingual subtitles, with French being the source language) extracted from five different TV series available on Netflix (i.e., *Plan Cœur, Mytho, Dix Pour Cent, Family Business, Lupin*) with the use of Language Reactor tool[2] and then converted into a .txt file. Each TV series was chosen so that it would represent everyday language without being stylized to fit a specific historical period or a particular language environment. Overall, 30 episodes from the said TV series constitute the corpus, which makes it the size of 120 429 tokens.

Due to the beforementioned fact that pragmatemes do not have a fixed syntactic structure, a different basis for the search has to be implemented. In this study, a list of 865 French pragmatemes by Blanco and Mejri (2018) is used. To this day (and to my

[1] If not stated otherwise, all translations in this paper are made by the author.
[2] Available at: https://www.languagereactor.com/. Last accessed 2022/7/14.

best knowledge), it is the only published piece presenting a substantial index of situational units, which are pragmatemes.

When both the corpus and the pragmatemes list are compiled as two separate files, the automatic search can be done. For this purpose, Unitex[3], a corpus processing suite, is used. Unitex offers many different tools essential for corpus analysis. The one used in this study is the "Locate Pattern" function: it helps to find a pattern either through the form of a regular expression (one particular sequence, for instance *bad weather*, or one pattern such as every adjective followed by a noun, i.e., <*A*><*N*>) or through a graph created in Unitex, giving all the information needed for the search. In this study, in order to find multiple sequences at once, the created graph contains the copy-pasted units from Blanco and Mejri's 2018 list (see: Fig. 1).

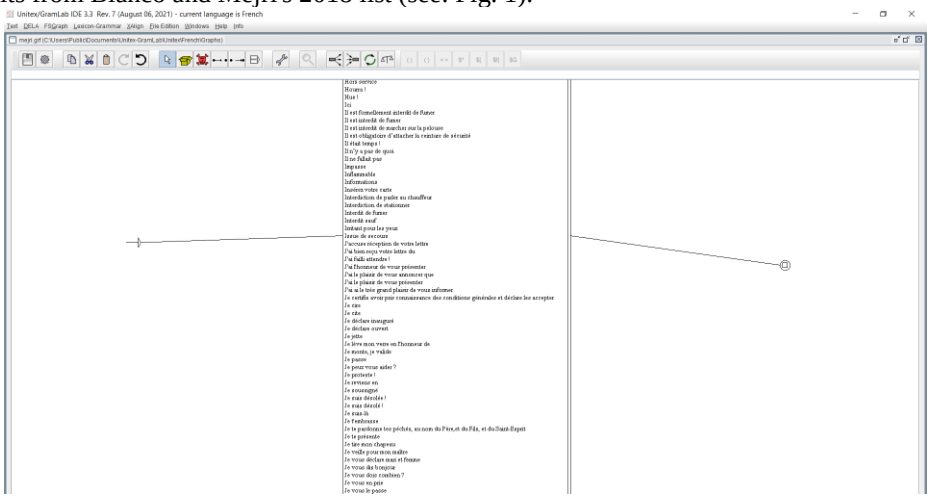


Fig. 1. Screenshot of a part of the graph created in Unitex containing the list of pragmatemes found in Blanco and Mejri (2018)

As a result, 1253 occurrences of searched units were found: 26 multi-word and 22 single-word pragmatemes (single-word pragmatemes were not dismissed in this study as their equivalents in different languages may consist of multiple words, for example, French *Désolé* and Polish *Przykro mi*; the results may therefore serve for further translational or comparative analysis). Table 1. presents the found multi-word units along with their frequency.

Table 1. Multi-word pragmatemes found in the captions corpus and their frequency

Pragmateme	Frequency	Pragmateme	Frequency
À bientôt.	6	Ça va ?	72
À demain !	3	Comment ça va ?	3
À l'aide !	3	De rien. / , [NAME].	3
[NAME], à l'appareil.	2	Dis-moi.	6

[3] Available at: https://unitexgramlab.org/fr. Last accessed 2022/7/14.

À plus tard.	6	Et vous ?	3
À tout de suite. / , [NAME].	2	Je peux vous aider ?	2
Attention !	3	Je t'embrasse.	2
Au secours !	4	Je te présente [NAME].	5
Au voleur !	3	Je vous en prie.	9
Bonne journée. / ! / , [NAME].	14	Je vous remercie.	3
Bonne nuit. / ! / , [NAME].	11	S'il vous plaît.	3
C'est pas grave.	9	Très bien.	15
Ça va.	19	Tu sais quoi ?	6

As it can be observed from the Table 1., the most frequent pragmatemes were: *Ça va ?*, *Très bien.*, and *Bonne journée*, while the least frequent were: *[NAME], à l'appareil.*, *À tout de suite. / , [NAME].*, *Je peux vous aider ?*, and *Je t'embrasse*. These results can serve for further research, for instance a comparative one that would examine the nature of pragmatemes in different languages, a translational one that would discuss the difficulties in translating pragmatemes, or a didactic one, among others.

The next subsection provides a qualitative point of view to the results.

2.2 Suggestions of improvement

While the search can be considered successful given the small size of the corpus, a few improvements to the method can be implemented in order to obtain more accurate results.

Adjusting the list of units

When using corpus software for the search of fixed units, their spelling is particularly important as it can make an impactful difference.

Firstly, Unitex takes into account all capital letters in the graph. For instance, when one searches for *Ça va.*¸they will not find *Oui, ça va*. which allows for a slot at the beginning of the unit and is just a variant of the original pragmateme. However, if one was to use the search with the use of lowercase letters only, the occurrences would show both the units starting with the uppercase and the lowercase. Blanco and Mejri's 2018 list of pragmatemes contains only seven pragmatemes that start with a lowercase so it would be advisable to adjust the rest in further studies, which can be easily done, e.g. with the application of the regular expression "=LOWER(CELL)" in Excel.

Then, some inconsistencies can be found regarding prosody markers used with the units prepared by Blanco and Mejri (2018), which can severely affect the automatic corpus search. Some units are marked as exclamations while other similar units are not, for example, *Bon voyage !* and *Bon séjour*. While it cannot be disputed that prosody is important in the study of pragmatemes (for further information on this topic, see: e.g., Banyś 2020), the addition of punctuation marks at the end of a unit can visibly alter the automatic search (i.e., if one searches for *Bon voyage !*, units such as *Bon voyage.* would

not be found). Therefore, the deletion of exclamation marks and periods would make the search more accurate.

Adding more sources

At the time of writing this article, Blanco and Mejri's 2018 *L'index de pragmatèmes* seems to be the only available extensive list of pragmatemes per se. However, it should not prevent any researchers from preparing their own pragmatemes lists. Having strictly defined their understanding of the term, a researcher interested in the topic can then collect pragmatemes from various resources, such as dictionaries of phrases, conversation guides, language textbooks, and articles published on the subject. To fully represent a modern-day language, one would also have to consider new pragmatemes present in the language of the youth (often unavailable in standard dictionaries).

Searching for patterns

Even though it was claimed at the beginning of this paper that pragmatemes do not have a fixed syntactic structure, some regular patterns can be distinguished. Nonetheless, an extensive list of pragmatemes is still needed to do so, as these regularities can only be distinguished by observation of pragmatemes already present in the list. Here are a few examples based on pragmatemes from Blanco and Mejri's 2018 list:

- bon + NOUN (e.g. *Bon voyage !*, *Bon retour !*, *Bon vol !*)
- À + TIME (e.g. *À demain !*, *À lundi prochain !*, *À mardi !*)
- En cas de + NOUN[,] IMPERATIVE VERB (e.g. *En cas d'incendie briser la vitre*, *En cas d'affluence ne pas utiliser les strapontins*, *En cas` de malaise, consulter un médecin*)

Such a search may result in finding pragmatemes that would not be found in any other sources and may add new value to the research on said units. However, in researching regular patterns, one would have to examine whether a pattern always results in a pragmateme, or whether it can produce a sequence that would not be considered a fixed unit. That is why a great amount of data should be provided for such research.

3 Conclusions

This paper aimed to explore an automatic method of extracting pragmatemes in a textual corpus. As it was presented, these units may cause a number of problems to researchers, having been presented in literature from many different points of view. Therefore, a crucial first step to any study on pragmatemes would be to define them in detail, either by adapting an already suggested theory (as it was done in this paper) or by proposing a new one, should the existing ones not be detailed enough or not fit for the particular study.

In terms of the automatic extraction of pragmatemes, it was shown that it is indeed possible; however, its preparation should be carefully managed. First, compiling a list

of well-documented pragmatemes in a language (or using an already existing one, as done in this study) is necessary in order to have a base for the automatic search. Then, for better results, adjusting the list according to the formal requirements of the used corpus software is advisable. Having done so, one is able to obtain valid occurrences from the corpus.

In closing, corpus studies on pragmatemes are essential for understanding our use of language in repeatable everyday situations. It provides a practical point of view with real-life examples and their frequency. Sinclair's dream was the creation of "a dictionary containing all the lexical items of language, each one in its canonical form with a list of possible variations," which would be "the ultimate dictionary" (Sinclair in Sinclair et al. 2004: XXIV). It cannot be achieved without an extensive examination of pragmatemes, the units accompanying us every day wherever we go. Such a dictionary may be then used for didactic purposes, as an aid for translators, among others. That is why it is worth studying pragmatemes, no matter how difficult it may seem.

References

1. Greaves, C., Warren, M.: What can a corpus tell us about multi-word units?. In: The Routledge Handbook of Corpus Linguistics, Routledge, pp. 212-226 (2022).
2. Rogers, J., Müller, A., Daulton, F.E., Dickinson, P., Florescu, C., Reid, G., Stoeckel, T.: The creation and application of a large-scale corpus-based academic multi-word unit list, In: English for Specific Purposes, Volume 62, pp. 142-157 (2021).
3. Banyś, W.: Pragmatèmes au pays de la prosodie, In: Neophilologica, Volume 32, pp. 89-116 (2020).
4. Firth, J. R.: Modes of meaning. In: Firth, J. R. (Ed.), Essays and studies, London: Oxford University Press, pp. 118-149 (1951).
5. Firth, J. R.: Papers in Linguistics 1934–1951. London: Oxford University Press (1957).
6. Bally, C.: Traité de stylistique française, Georg & cie, Paris (1909).
7. Blanco, X.: Les pragmatèmes : définition, typologie et traitement lexicographique. In: Verbum 4, pp. 17-25 (2013).
8. Blanco, X., Mejri, S.: Les Pragmatèmes, Classiques Garnier, Paris (2018).
9. Mel'čuk, I.: Phrasemes in language and phraseology in linguistics. In: Everaert, M., Van der Linden, E., Schenk, A., Schreuder, R. Hillsdale Idioms: Structural and Psychological perspectives, pp. 167–232 (1995).
10. Melčuk, I. A.: Phraseology in the Language, in the Dictionary, and in the Computer. In: Język i metoda 1, pp. 217-239 (2012).
11. Coulmas, F.: On the sociolinguistic relevance of routine formulae. In: Journal of Pragmatics, 3, pp. 239–266 (1979).
12. Barnas, M.: Les pragmatèmes dans les dialogues dans les romans de Marc Lévy. In: Sciences de l'Homme et Société (2017).
13. Bardovi-Harlig, K.: Formulas, routines, and conversational expressions in pragmatics research. In: Annual Review of Applied Linguistics, 32, pp. 206–227 (2012).
14. Chlebda, W.: Frazematyka. Encyklopedia kultury polskiej XX wieku, Bartmiński, J. Wrocław (1993).
15. Fléchon, G., Frassi, P., Polguère, A.: Les pragmatèmes ont-ils un charme indéfinissable ? In: Lexiques. Identité. Cultures, QuiEdit, pp. 81-104 (2012).

16. Kauffer, M.: Les « actes de langage stéréotypés » : essai de synthèse critique, in: Cahiers de lexicologie. Les phrases préfabriqués : Sens, fonctions, usages, Classique Garnier, pp. 149-172 (2019).
17. Language Reactor, https://www.languagereactor.com/, last accessed 2022/7/14.
18. Unitex/Gramlab Homepage, https://unitexgramlab.org/fr, last accessed 2022/7/14.
19. Sinclar, J., Jones, S., Daley, R.: English Collocation Studies: The OSTI Report. London: Continuum (2004)
20. Wood, D.: Fundamentals of Formulaic Language, Bloomsbury Academic (2015).
21. Wray, A., & Perkins, M. R.: The functions of formulaic language: An integrated model. Language & Communication, 20(1), pp. 1–28 (2000)..
22. Maziarz, M., Rudnicka, E., Grabowski, Ł.: Multi-word Lexical Units Recognition in WordNet (2022).
23. Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for NLP. In International conference on intelligent text processing and computational linguistics, Springer, Berlin, Heidelberg, pp. 1-15 (2002).
24. Finkbeiner, R., Schlücker, B: Compounds and multi-word expressions in the languages of Europe. In: Schlücker, B. (ed.) Complex Lexical Units: Compounds and Multi-Word Expressions, Berlin, Boston: De Gruyter, pp. 1-44 (2019).
25. Masini, F.: Phrasal lexemes, compounds and phrases: A constructionist perspective. Word Structure 2, 2. pp. 254–271 (2009).

Handling the study of Multi-Word Expressions from beginning to end via an online collaborative annotation platform: ACCOLÉ

Emmanuelle Esperança-Rodier[1] Fiorella Albasini[1] and Francis Brunet-Manquat[1]

1 Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
firstname.lastname@univ-grenoble-alpes.fr
fiorella.albasini@etu.univ-grenoble-alpes.fr

Abstract. "Multi-word expression" (MWE) refers to "linguistic phenomena that lead to a variety of constructions such as idioms, frozen expressions, compound nouns, verb-particle constructions" (Ramisch, 2012). MWEs make up a large part of our everyday language (Jackendoff, 1997) and their detection is often problematic in Natural Language Processing (NLP) (Bouamor, 2014). Nevertheless, to our knowledge, no tool offers to create as well as study MWE in NLP contexts and Machine Translation ones. This article presents the online collaborative annotation platform, ACCOLÉ, for the annotation of continuous and non-continuous MWE, in monolingual and multi-target corpora. Among several features, ACCOLÉ offers an MWE annotation proposal to be validated manually, along with manual annotations, and a discussion feature that ensures intersubjective annotations. After explaining the context of our research and the related works, we describe the main features of ACCOLÉ. We will conclude with further works to be achieved and the ones already performed.

Keywords: Multi-Word Expressions, annotation, Machine Translation.

1 Introduction

Multi-word expressions (MWE) make up a large part of our everyday language (Jackendoff, 1997). Similarly, Sag et al. (2002) estimate that their use is equivalent to that of single words in the language. These constructions are mastered unconsciously by native speakers (Hausmann, 1997). Thus, MWEs can be defined as an often arbitrary combination of words where the meaning in their totality cannot be deduced from its individual components (Sag et al., 2002)

* Institute of Engineering, Univ. Grenoble Alpes

2

As pointed out by Ramisch (2012), MWEs appear as different morphosyntactic structures. We use the term "multi-word expression" to refer to "linguistic phenomena that lead to a variety of constructions such as idioms, frozen expressions, compound nouns, verb-particle constructions" (Ramisch, 2012) among them we also find collocations, i.e., "combinations of words that have affinities and tend to appear together, but not necessarily continuously" (Tutin and Grossmann, 2002): as well as proper names, proverbs, and routines.

MWEs are easily recognized by humans, however, their detection is often problematic in Natural Language Processing (NLP) (Bouamor, 2014). In machine translation (MT), failing to recognize an MWE is one of the main sources of error. (Constant et al., 2011).

Consequently, the study of how MWEs appear and behave in documents, or are translated, is of high interest among the NLP community. To achieve this goal, a lot of corpora have been created. As we mainly work on French, we will just list the main ones for this language. First, the French Treebank (Abeillé et al., 2003) contains various MWEs, including verbal ones, but only on continuous expressions. Then, there are the two corpora of nominal and adverbial MWEs from Laporte et al. (2008a; 2008b) which do not provide a typology. PolyCorp, from Tutin (2016) and Tutin and Esperança-Rodier (2019), provides a typology and uses a pre-identification of MWEs. Finally, the PARSEME project (Savary et al., 2015) uses FLAT - FoLiA Linguistic Annotation Tool, the PARSEME-customized online annotation platform[1], that on top of corpora in several languages, provides a typology as well as decision flow charts and many annotation and investigation tools.

On top of the PARSEME project, other research produces bilingual and multilingual Corpora, such as the bilingual English-Hungarian corpora with annotations solely on Light Verbal Expressions (Vincze, 2012) and AlphaMWE (Han et al., 2020), a multilingual corpus covering English, Chinese, Polish, and German where equivalent MWEs in all the languages are annotated to study MT behaviour. There are also many MWE annotations within treebanks in different languages, but we won't enumerate them in this paper.

Each of the above corpora relies on one or another typology, some corpora come along with an annotation tool or online MWE investigation

[1] https://github.com/proycon/flat

tools. Some are monolingual corpora and others are bilingual or multilingual (resulting in the translations of monolingual corpora). Some others use pre-identification of MWE, via lexicons or Deep Learning. Only continuous MWE annotations are addressed for most of those works.

Nevertheless, if one wants to build monolingual or bi/multilingual MWE annotated corpora to study how MT translates MWEs, one will have to navigate between the different projects, typologies, tools, and file formats. To our knowledge, no tool offers to annotate monolingual or aligned MT corpora in MWEs, nor to provide MWE annotation proposals on continuous and non-continuous MWE, to choose within several typologies, to allow annotators to collaborate to ensure intersubjectivity and to allow investigating the annotated corpora within the same framework.

This article describes the online collaborative platform ACCOLÉ primarily conceived to annotate translation errors but expanded to annotate continuous and non-continuous MWEs in monolingual corpora, bilingual, and multi-target aligned ones, by proposing MWEs.

After briefly presenting related works, we will present how ACCOLÉ is structured and describe how typologies can be integrated. We, then, explain how ACCOLÉ pre-identifies the MWE to be annotated, and we will show how the discussion feature allows achieving a high agreement among annotators if several. We will conclude with an example study on MWE translation using ACCOLÉ and further work.

2 Related works

As our work regards the automatic detection of MWE and their annotation, we have selected some tools that detect automatically MWE or only collocations and Named Entities mainly, as well as annotation tools.

MWEToolKit (Ramisch, 2015): “The Multiword Expressions toolkit is a tool that aids in the automatic identification of multiword units such as idiomatic expressions (kick the bucket) and phrasal verbs (take off, give up) in large text bases, independently of the language.” As mentioned on the website[2], it requires an installation and offers several tools to deal with MWEs. MWEToolKit now offers a library implementation in Python.

[2] http://mwetoolkit.sourceforge.net/PHITE.php?sitesig=MWE last accessed on 21/07/2022.

Sketch Engine[3]: offers the tool Word Sketch — collocations and word combinations. The Sketch grammar uses pre-defined rules that permit the detection of collocation by showing via the word sketch interface a list of word collocates and other words within a specified span.

spaCy: a python library dedicated to NLP that among other features offers to detect Named Entities.

Mind the Gap (Coavoux et al., 2019): is an unlexicalized transition-based parser for discontinuous constituency structures and identifies MWEs.

Flat — FoLiA Linguistic Annotation Tool, which is the PARSEME-customized online annotation platform that allows a lot of linguistic annotations, including MWE annotations.

PARSEME also offers a panel of tools that allows for visualizing annotated corpora, such as GrewMatch (Schmitt et al., 2019).

Once again, none of those tools offers, at once, the full treatment of MWEs, from their detection, and annotation, to their study. This is what we are offering while developing ACCOLÉ, to provide MWE annotated corpora for NLP and MT studies on MWEs.

3 ACCOLÉ

We propose an online platform that allows annotating continuous and non-continuous MWEs, of all types, on monolingual or multi-target corpora by validating the proposals of MWEs made by the platform or manually detecting MWEs and assigning the type of the chosen typology. Hence, any typology (section 3.1) can be integrated into ACCOLÉ and associated with a monolingual or multi-target corpus to create an annotation project. This allows loading the corpus only once. When creating a project, several annotators and supervisors can be assigned. The supervisor plays the role of a moderator and has also a decision-making role in discussions.

Within the annotation project, ACCOLÉ proposes potential MWEs to be validated by the annotator (Section 3.2) but the annotator can also annotate manually the MWEs (Section 3.3).

Being a collaborative tool, ACCOLÉ offers the Discussion feature (Section 3.4).

[3] https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/ last accessed 21/07/2022

All the pieces of information, i.e., annotation span, source, target selection, assigned type, and discussion opened on a specific annotation, are stored thanks to Structured String-Tree Correspondences (SSTC) (Boitet and Zaharin, 1988).

Once created, the corpus can be explored using the search feature, to look at the different MWEs, the way they are translated for multi-target corpus, and the patterns in which they appear.

To test the features, we have worked on the French-English ParaSHS Corpus – "Témoigner" (Kraif, 2018) a literary genre, and on its translations into Spanish.

Hereafter comes the description of the several features offered.

3.1 Typologies

ACCOLÉ offers to enter manually the typology one wants to use for annotations. Typologies are entered under a hierarchical scheme allowing 3 levels: Category level, Sub-Category Level, and Name. Hence, several typologies can be created and associated with any corpus via the project management system. As we worked on the typology created by Tutin (2016), this typology is already available on the platform.

3.2 MWE Detection

Lexicon based. ACCOLÉ detects MWEs in French monolingual corpora using a lexicon created within PolyCorp (Tutin, 2016; Tutin and Esperança-Rodier, 2019). However, this dictionary is not complete.

ACCOLÉ is a useful tool for both the automatic discovery of MWEs when annotating corpora through automatic detection of MWEs and human validation, as well as for the manual discovery of these structures. Automatic detection is performed by the morphosyntactic analysis tool «Mind The Gap[4]» (Coavoux et al., 2019), with a high success rate, particularly in terms of frozen expressions, and verb-particle constructions, often overlooked by other tools even in discontinuous MWE expressions. Examples of detected MWE include:

- "n'aura pas lieu" : (VN (ADV 22=n') (V 23=aura)) (ADV 24=pas) (NP-OBJ (NC 25=lieu)))
- "en guise de" : (PP-MOD (P+ (P 3=en) (NC 4=guise) (P 5=de))

4 https://github.com/mcoavoux/mtg

- “faisant autorité” : (VPpart (VN (VPR 5=faisant)) (NP-OBJ (NC 6=autorité))
- “se mettant lui-même en première ligne”: (VPpart (VN (CLR 8=se) (VPR 9=mettant)) (NP-MOD (PRO 10=lui-même)) (PP-P_OBJ (P 11=en) (NP (ADJ 12=première) (NC 13=ligne))))
- “j’ai toujours été frappé”: (VN (CLS-SUJ 0=J') (V 1=ai) (ADV 2=toujours) (VPP 3=été) (VPP 4=frappé))

Morphosyntactic analysis based. ACCOLÉ also allows the analysis of translation errors in aligned bilingual and multi-target corpora. Another feature of the platform is to automatically detect MWEs in the target and source language simultaneously. We use MWEToolKit (Ramisch, 2015) in combination with Treetagger (Schmid, 1994; 1995) to extract MWEs and obtain their POS tags.

The Python library spaCy is used to extract Named Entities in Spanish, French, and English. All tools were tested in the same corpora: ParaSHS-Témoigner. In all three languages, Named Entities (NE) were almost always identified.

- (‘Christa Wolf’, 1289, 1301) → {Christa: ‘PROPN’} {Wolf: ‘PROPN’}
- (‘The Trojan War Will Not Take Place’, 1873, 1907) → {The: ‘DET’} {Trojan: ‘PROPN’} {War: ‘PROPN’} {Will: ‘AUX’} {Not: ‘PART’} {Take ‘VERB’} {Place: ‘PROPN’}

As expected, erroneous MWE detection occurs in all three languages if part of the text is not translated by the MT system and left as it stands, in the source language, which can be the case of proper nouns in the corpus ParaSHS - Témoigner.

3.3 Annotation Scheme

ACCOLÉ displays monolingual corpora sentence per sentence. MWE proposals are underlined and must be validated by the annotator. The annotation is stored thanks to SSTC (Boitet & Zaharin, 1988).

When an MWE has not been detected by ACCOLÉ, the annotator proceeds to manual annotation. They select the MWE and associate it with a type from the typology.

For Multi-target corpora, the target sentences are presented next to the source one, and the MWE proposal validations and manual annotations must be done on the source as well as on the targets.

If an annotator has any doubt, they open a discussion and start a collaboration on the specific annotation with other annotators and with the project supervisor.

3.4 Collaboration

One of the distinctive features of ACCOLÉ is its collaborative aspect, the interaction between users. User interaction prevails when creating, for example, a new entry in the MWE lexicon. ACCOLÉ allows users to create discussions in which they can exchange opinions and reach a consensus to validate a new addition, ensuring the accuracy of new entries from a linguistic point of view, as opposed to a supervisor-oriented approach.

The annotated segments that opened discussions between the supervisor and annotators show a greater correspondence in terms of agreement. This can be attributed to the fact that the discussion generates an open, shared reflection among the participants, leading to validation in terms of typology and delimitation, which therefore implies a unique response among all the participants of the discussion.

4 Conclusion

We have presented ACCOLÉ: an online collaborative annotation platform for MWEs. This platform allows managing MWE annotation from the beginning to the end of the process, from the choice of the MWE typology, and the annotation process to the analysis of the annotations. It allows the creation of monolingual corpora annotated with MWEs that can help the NLP community as well as multi-target corpora also annotated with MWEs available for training, and testing in the MT community.

ACCOLÉ allows annotating continuous and non-continuous MWEs in several languages. The platform also offers to detect manually MWEs or by validating MWE automatic proposals.

Thanks to this platform, work on the comparative evaluation of MWE translation from French to Polish has already been performed by Es-

perança and Frankowski (2022), showing that DeepL and Google Translate were struggling mainly with the same type of MWEs, but DeepL achieved better translations than Google Translate for this language pair, even if it used English as a pivot.

Further work may include adding other part-of-speech taggers to investigate if MWEs appear in specific patterns at the morphosyntactic level and if those patterns induce translation errors. We also would like to integrate other MWE lexicons, and decision-making flow charts and to start working with Treebank corpora.

Acknowledgments

The work reported above has been granted by NeuroCoG/Pôle Grenoble Cognition funding as well as LIG/Emergence funding.

References

1. Abeillé, A., L. Clément, and F. Toussenel. "Building a treebank for French", in *Treebanks*, Kluwer, Dordrecht. (p.165-187) (2003)
2. Boitet, C. and Zaharin, Y. Representation trees and string- tree correspondences. In *Proceedings of International Conference on Computational Linguistics* COLING-88, pages 59-64 (1988).
3. Bouamor, D. Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables. Université Paris Sud - Paris XI, (2014).
4. Candito, M., Constant, M., Ramisch, C., Savary, A., Guillaume, B., et al. A French corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, Institute of Computer Science, Polish Academy of Sciences, Poland, 8 (2), pp.415-479. ⟨10.15398/jlm.v8i2.265⟩. ⟨hal-03016721⟩ (2020).
5. Coavoux, M., Crabbé, B., Cohen, S. H. Unlexicalized Transition-based Discontinuous Constituency Parsing. *Transactions of the Association for Computational Linguistics* 2019; 7 73–89 DOI: https://doi.org/10.1162/tacl_a_00255 (2019).
6. Constant, M., Tellier, I., Duchier, D., Dupont, Y., Sigogne, A., Billot, S. et al. Intégrer des connaissances linguistiques dans un crf : application à l'apprentissage d'un segmenteur- étiqueteur du français. In Actes de TALN, Montpellier, France (2011).
7. Esperança-Rodier, E., Brunet-Manquat, F., Eady, S. ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. Translating and the computer 41, Nov 2019, Londres, United Kingdom. ⟨hal-02363208⟩ (2019).
8. Esperança-Rodier, E. and Frankowski, D. DeepL vs Google Translate: who's the best at translating MWEs from French into Polish? A multidisciplinary approach to corpora creation and quality translation of MWEs. *Translating and the Computer - TC43*, Nov 2021, London, United Kingdom (2022).
9. Han, L., Jones, G. J. F. and Smeaton, A. AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations in the Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons Proceedings of the Workshop (MWE-LEX 2020). 2020. <hal-03013612> (2020).

10. Hausmann, F. J. La locution : entre langue et usages, 277-290 « Tout est idiomatique dans les langues », (1997).
11. Jackendoff, R. The architecture of the language faculty. MIT Press (1997).
12. Kraif, O. Constitution et traitement d'un corpus bilingue d'articles scientifiques : exemple de mise en œuvre automatique avec une architecture légère en Perl. In *Journées LTT 2018*, sept. 2018, Grenoble (2018).
13. Laporte, E., Nakamura, T., & Voyatzi, S. A French corpus annotated for multiword nouns. In Language Resources and Evaluation Conference (LREC). Workshop Towards a Shared Task on Multiword Expressions. pp. 27-30 (2008a).
14. Laporte, E., Nakamura, T., & Voyatzi, S. A French corpus annotated for multiword expressions with adverbial function. In Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop. pp. 48-51 (2008b).
15. Ramisch, C. Une plate-forme générique et ouverte pour le traitement des expressions polylexicales. Actes de *14e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2012*), Grenoble, France. ⟨hal-00959236⟩ (2012).
16. Ramisch, C. Multiword Expressions Acquisition: A Generic and Open Framework", Theory and Applications of Natural Language Processing series XIV, Springer, ISBN 978-3-319-09206-5, 230 p., (2015).
17. Rikters, M., and Bojar, O. Paying Attention to Multi-Word Expressions in Neural Machine Translation. Preprint (2017).
18. Sag, I. A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. Multiword Expressions: A Pain in the Neck for NLP. International Conference on Intelligent Text Processing and Computational Linguistics. 1-15. Springer, Berlin, Heidelberg (2002).
19. Savary, A., Piskorski, J. Language Resources for Named Entity Annotation in the National Corpus of Polish. Control and Cybernetics, Polish Academy of Sciences, 40 (2), pp.361-391 (2001).
20. Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rose´n, V., Przepio´rkowski, A., Krstev, C., Vincze, V., Wo´jtowicz, B., Losnegaard, G. S., Parra Escart´ın, C., Waszczuk, J., Constant, M., Osenova, P., and Sangati, F. PARSEME–PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)* (2015).
21. Schmid, H. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland. (1995).
22. Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK (1994).
23. Schmitt, M., Moreau, E., Constant, M., Savary, A. Démonstrateur en-ligne du projet ANR PARSEME-FR sur les expressions polylexicales. Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL), Jul 2019, Toulouse, France. pp.627-630 (2019).
24. Tutin, A. & Grossmann, F. Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif. Revue Française de Linguistique Appliquée, Lexique : recherches actuelles, vol VII:p 7–25 (2002).
25. Tutin, A. & Esperança-Rodier, E. The difficult identification of multiword expressions: from decision criteria to annotated corpora; European Society of Phraseology Conference (EUROPHRAS 2019), Sept. 2019, Malaga, Spain (2019).

26. Tutin, A., Esperança-Rodier, E., Iborra, M., and Reverdy, J. Annotation of multiword expressions in French, in *Proceedings of EUROPHRAS 2015*, pp. 60–67, Malaga, Spain (2016).
27. Vincze, V. Light verb constructions in the SzegedParalellFX English–Hungarian parallel corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12), pages 2381–2388, Istanbul, Turkey, May. European Language Resources Association (2012).

From Monolingual Multiword Expression Discovery to Multilingual Concept Enrichment: an Ontology-based approach

Gennaro Nolano[1], Maria Pia di Buono[1], and Johanna Monti[1]

[1]Unior NLP Research Group, University of Naples "L'Orientale"
{gnolano,mpdibuono,jmonti}@unior.it

Abstract. In this paper, we present a methodology for the semantic enrichment of cultural heritage (CH) data, based on the use of ontologies and Linked data. The proposed method aims at developing domain-specific resources enriched with multilingual conceptual information starting from monolingual RDF data. Particularly, our approach begins with a Multiword Expressions (MWEs) discovery process to select a starting list of domain-specific candidate mentions. Subsequently, we perform a concept discovery phase in order to link them to closely matching Dbpedia concepts through the use of two similarity measures. The semantic information related to these concepts is used to further filter the candidates and obtain representative mention-concept pairs by reweighting automatically computed scores making use of a graph representation.
We test our methodology on biographic information about authors extracted from the Europeana Data Collection. The final results are a resource of semantically enriched data, containing a list of domain-specific keywords and MWEs together with Dbpedia concepts they strongly match, and the multilingual labels representing these specific concepts.

Keywords: MWE discovery · Concept Discovery · Ontology

1 Introduction

In this paper, we present an approach to developing domain-specific multilingual resources starting from monolingual RDF data. Particularly, our approach focuses on Multiword Expressions (MWEs) discovery and exploits linking techniques through similarity measures to enrich the final linguistic resource (LR). The main rationale behind the proposed methodology is that RDF metadata contains fields for both descriptive texts (e.g., `dbo:abstract`[1]) and structured information from external Knowledge Bases (KBs) (e.g., `dbo:wikiPageExternalLink`[2]). These two sources of data could both be exploited in the creation of a domain-specific semantically enriched resource of entities and mentions (i.e., entities'

[1] `https://es.dbpedia.org/ontology/abstract`
[2] `https://dbpedia.org/ontology/wikiPageExternalLink`

surface forms) from descriptive texts are extracted and conceptually linked to structured information. A resource created in this way would be suitable for a series of Natural Language Processing (NLP) tasks, such as Terminology Extraction, Machine Translation and Entity Linking.
The paper is organized as follows: Section 2 describes some of the efforts made by researchers in the fields of MWE discovery and Semantic Enrichment; Section 3 explains in general terms the proposed methodology; Section 4 is devoted to the practical aspects related to the creation of the proposed LR; finally, Section 5 illustrates the final LR while also introducing possible future work.

2 Related Work

Multiword Expression Discovery There is a considerable body of works describing techniques to automatically detect MEWs. Generally, this is solved through statistical means by computing the correlation strengths between words forming the expression [14,6], with the most widely used association measure being pointwise mutual information [5]. Despite the effectiveness of these models, one of the main drawbacks is the need for a background corpus to compute statistical significance. This corpus might not exist, or might not be big enough when dealing with certain domains and certain languages.
Another option is to use syntactic patterns to generate MWE candidates. This, for instance, has been explored in works such as [11,1]. Since these patterns can be generated from heuristics, they can be applied to any kind of text, no matter their length. In this work, we propose the use of several syntactic patterns specifically tailored for the Italian language.

Semantic Enrichment Much effort has been made trying to fill the semantic gap between the "web of documents" and the "web of knowledge" [4], as shown in works such as [7,2,8,16].
Nevertheless, such models have generally focused on tasks related to Named Entities, such as Named Entity Recognition (NER), Named Entity Linking (NEL) and Named Entity Disambiguation (NED). While these tasks effectively integrate some sort of semantic knowledge into raw texts, they generally focus on just Named Entities which would directly leave out important domain-specific concepts, such as classes and topics (e.g., *classical music* and *Roman architecture*), which are rarely identified by proper names.
For this reason, connecting important spans of text to *concepts* rather than specific Named Entities can benefit many NLP tasks. Concept discovery [12] has been explored for domains such as news articles [10] and scientific knowledge [15].
One drawback of such models is that they generally focus on a single language, while connecting a raw text to specific concepts present in a knowledge base as Wikidata or Dbpedia can help provide multilingual access to data, thus improving the reusability of LR.
In this work, we integrate multilingual data in the final resource by exploiting labels used to describe Dbpedia concepts.

3 Methodology

Our methodology relies on the use of monolingual RDF data and makes use of the unstructured text presented by descriptive metadata and structured information in the form of external links. Basically, we extract MWEs from the unstructured texts and connect them to the concepts (in the form of links) which are conceptually related to the specific RDF item that is being described. More specifically, we perform three main steps:

1. Monolingual MWE Discovery
2. Concept Discovery
3. Ontology-based filtering

The first two steps of the methodology are shown in graphical form in Figure 1, while an example of a graph used for ontology-based filtering is shown in Figure 2.

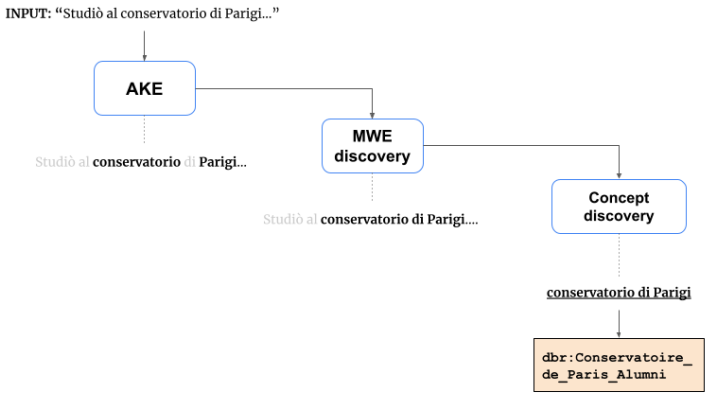


Fig. 1. Graphical representation of the process of monolingual mwe discovery: in the first step, Keywords are automatically extracted using the pke Python library. The automatically extracted keywords are then expanded based on specific patterns. The extracted MWEs are connected to the link they most likely refer to, through similarity measure with the italian label of said link.

Monolingual MWE Discovery In order to perform this step, we first rely on off-the-shelf libraries to extract keywords from the texts at hand. While some of these keywords might be represented by Named Entities, we do not want to put any restriction on the type of semantic information we want to extract from the

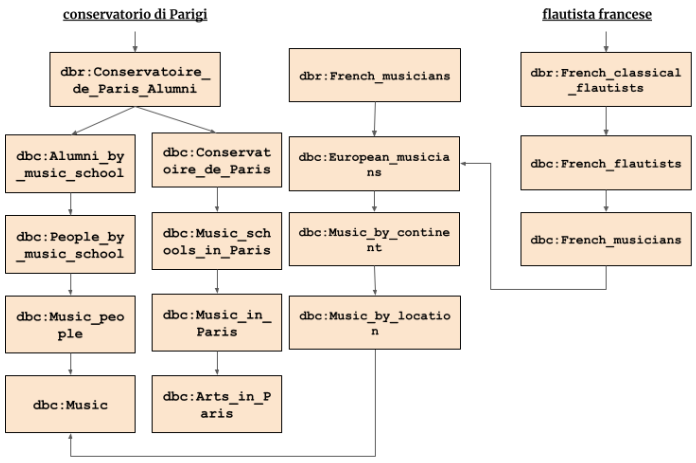


Fig. 2. Graphical representation of the graph implemented for ontology-based filtering. The graph is constructed traversing the `skos:broader` property 4 times from the links available for the entity. The nodes representing concepts related to the entity will generally share several edges, thus giving more weights to the MWEs linked to them.

texts. Thus, we do not rely on NER tools (which usually limit their information to People, Locations and Organizations), but rather automatic keyword extraction tools, which rely on purely statistical information. The extracted results can thus be any sort of concept, without restrictions.
These results represent a first list of scored keyword candidates which are then used as inputs for an MWE discovery phase.
To achieve that, after the automatic keyphrase extraction phase, we check the extracted candidates in context, that is in the source texts, and assume that a candidate w_x is part of an MWE when it is close to another candidate w_y and they are separated by specific elements, according to hand-defined linguistic patterns based on heuristics (Table 1).
Then, we assign a relevance score to the extracted MWEs and classify them using the conceptual category already associated with the main item.

Concept Discovery In order to improve the results from the previous step and semantically enrich them, we inject external knowledge from DBpedia[3]. Since most RDF data are linked to other external KBs, it is possible to include it as a source of additional information, which can usually be accessed through specific endpoints. In particular, the data used in this work is connected to DBpedia entries, and we make use of the DBpedia SPARQL endpoint[4] and the `dbo:wikiPageWikiLink` property, which connects a specific DBpedia entry to

[3] `https://www.dbpedia.org/`
[4] `https://dbpedia.org/sparql`

Distance	Pattern	Example
Zero	w_1 w_2	Adelaide Festival
1 Element	w_1 *PREP* w_2	nuova generazione *di* musicisti folk
	w_1 *ADJ* w_2	Aleksandra Aleksandrovna *nata* Grigorovič
2 Elements	w_1 *PREP DET* w_2	premio Nobel *per la* letteratura
	w_1 *ADJ PREP* w_2	direttore *principale della* Philharmonia Orchestra

Table 1. MWE candidate patterns.

The translation for the presented examples are, respectively: *Adelaide Festival*, *new generation of folk musicians*, *Aleksandra Aleksadrovna nee Grigorovič*, *Nobel prize for Literature*, *main directory of the Philharmonic Orchestra*.

every other entry present as a link in its Wikipedia article. This means that concepts that are related to a certain topic or domain are likely to share similar links.
Once these entities are extracted, we calculate similarities between them and extracted MWEs. Thus, it is possible to create a network of connections between raw spans of text and entities in DBpedia, consequently obtaining a certain level of semantic enrichment, while also enabling the inclusion of multilingual data in the form of the labels representing the various concepts.

Ontology-based Filtering The information collected during the concept discovery phase is then used to filter the extracted candidates. All the links extracted in the previous steps are also used to recreate a hierarchical graph by traversing the `skos:broader` relation. In particular, for each link, we extract its parent nodes until we get to the 4rd highest node in the hierarchy.
This way, we end up with a graph-based representation of the connected concepts, in particular the links that were connected to MWEs during concept discovery. We make use of this graph to filter and reweight domain-specific keywords and MWEs, while also using it as a source of additional semantic information to enrich the data at hand.

4 Experiment and Results

Data Collection In order to collect domain-specific texts, we refer to the Europeana Entity API[5], which allows for the search and retrieval of RDF data about entities from the Europeana Entity Collection.
These entities represent a collection of Named Entities harvested from and linked to several online data catalogues, such as Geonames, DBpedia and Wikidata. In particular, for the purpose of this work, we extract biographical information about entities of type `agents`, which represent artists from different cultural heritage sub-domains such as music and fashion.

[5] `https://pro.europeana.eu/page/entity`

Using the SPARQL API for the Europeana Data Collection[6], we recollect the following information for 500 agent entities:

- their English label,
- the DBpedia entry they are linked to, and
- the Italian text for their biographical information.

Monolingual MWE discovery We first extract keyphrases from each text using the `pke` Python library[7], which returns a list of weighted results representing extracted keyphrases and their relevance according to a specific model. The library provides several models for AKE, among which we opt for the MultiPartite Ranking [3]. The resulting relevance score for each keywords ranges from 0 to 1. As already stated, from this list of automatically extracted keywords, we aim at discovering MWEs within the text. To do so, we check whether two keyphrases are close enough[8], and whether the sequence of words between them is acceptable according to pre-defined patterns of co-occurring elements (as shown in Table 1). We check each candidate occurring either in the w_1 or w_2 position.
To assign a score to the newly extracted MWEs, we calculate the average MultiPartite Ranking value for each of the keyphrases involved in the MWE by summing the values of each keyphrase and then dividing the result by the number of keyphrases belonging to the discovered MWE. Keyphrases which cannot be used to build any new MWE are kept as they are.
In total, we extract 4770 keywords and MWEs. In Table 2 we show the different effectiveness of each linguistic pattern in discovering new MWEs, together with the number of linked concepts for each of these patterns, as described in the next paragraph.

Pattern	**Occurrences**	**Linked to Concepts**
w_n	3795	1620
w_1 w_2	12	6
w_1 PREP w_2	878	555
w_1 ADJ w_2	3	1
w_1 PREP DET w_2	66	38
w_1 ADJ PREP w_2	16	12
Total	**4770**	**2232**

Table 2. Number of MWEs and connected concepts

Concept Discovery For each agent entity, we exploit its entry in DBpedia to extract all the hyperlinks present in the entity's corresponding Wikipedia page by accessing the `dbo:wikipageWikiLink` property using the DBpedia SPARQL

[6] `http://sparql.europeana.eu/`
[7] `https://github.com/boudinfl/pke`
[8] In this work we set the maximum distance window at 2 tokens.

endpoint.
One of the main issues of such hyperlinks is that in some cases they only present labels for the English language. This is the case for most of the category-defining entries such as `dbp:Victorian_poets` and `dbc:19th-century_English_poets`. Since these links generally refer to domain-specific knowledge classification, we want to access them even in absence of an Italian label. In order to do so, in case an Italian label is unavailable for a specific link, we automatically translate it from the English label using the Argos Translate Python library[9]. In order to make full use of the linked information available on DBpedia, we connect each keyword and MWE to the concept it most closely matches, by applying similarity measures over the links present in each specific page.
In particular, we use pre-trained fastText word vectors for Italian[10] to represent both MWEs and the Italian labels in vector space. Once we obtain these distributional representations, for each agent entity we compute the similarity scores between each MWE and each page link's Italian label. The similarity scores are calculated as the raw product of two different measures: cosine similarity between the embeddings and the overlap coefficient (i.e., the Szymkiewicz–Simpson coefficient [13]) between the surface forms. This way, we take into account semantic similarity between vectors, while also accounting for cases in which a specific MWE and a label share similar surface form despite their vectors being distant. From this list of computed similarity scores, we discard any MWE-link pair with a score lower than 0.4. Then, for each remaining MWE, we keep the link with the highest similarity score as the closest match. In case a MWE is not linked to any concepts, we leave it as it is for the following steps.
Regarding the 2232 keyphrases linked to Dbpedia concepts, in Table 3 we report the number of translated labels for each available languages in the final resource.

Language	**#Concepts**	**Language**	**#Concepts**
ar	933	ja	1028
ca	1.005	ko	893
cs	948	nl	1.085
de	1.228	pt	1.075
el	746	pl	1.075
en	2.232	ru	1.131
eu	815	sv	1.054
fr	1.266	uk	1.007
ga	514	zh	948
in	823		
Total			**19.806**

Table 3. Number of translated concepts

[9] `https://pypi.org/project/argostranslate/`
[10] `https://fasttext.cc/docs/en/crawl-vectors.html`

Ontology-based filtering Starting from the links collected from the entity, we recreate a hierarchical graph by traversing the `skos:broader` property 4 times. In this graph, concepts that are related to each other will generally share edges connecting to common nodes. In particular, we are interested in the links connected to specific MWEs.
We can then use the graph we obtain to re-rank the candidate keywords on the basis of their correlation to topics and concept: for each node linked to a specific MWE we calculate its betweenness centrality [9], which is then integrated together with the score calculated in the previous step.
For each MWE, the final score is computed as the sum $v_final = v_{mwe} + bc_{node}$, where $node_{mwe}$ is the value of the specific MWE, and bc_{node} the betweenness centrality of a the node connected to it. In case a MWE is not connected to any link, its score will be left as it was originally.
Finally, the extracted MWEs, ranked according to their reweighted scores, are enriched with related concepts and their multilingual labels (when present) from DBpedia.

5 Conclusion and Future Work

In this paper we described the process of semantic enrichment employed in the creation of a domain-specific multilingual resource.
The development makes use of statistical information to extract keyphrases, linguistic patterns to discover new MWE on the basis of automatically extracted keyphrases, similarity measures to link those to close concepts in Dbpedia, and finally graph-based representations to combine all these information together. The final proposed resource[11] is a collection of the following data extracted from the original biographic texts for 500 agent entities:

- MWE discovered through the combination of AKE and linguistic patterns,
- Dbpedia correlated entities linked to similar MWEs,
- multilingual labels for the Dbpedia entities for all available languages,
- broader concepts for each of the Dbpedia entity linked to a specific MWE.

In future work, we aim at improving the current results by refining the current process. For instance, a domain-specific BERT-like embedding model might help improving the concept discovery stage. The MWE discovery stage: for instance would benefit from a more flexibile and lexically (rather than just sintactically) grounded set of linguistic patterns might help improving current results while also reducing the noise currently present in the resource.

Acknowledgements

Maria Pia di Buono has been supported by Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 - Fondo Sociale Europeo, Azione I.2 “Attrazione

[11] `https://github.com/Nolanogenn/multilingual_ontology_based_ch`

e Mobilità Internazionale dei Ricercatori" Avviso D.D. n 407 del 27/02/2018. Authorship Attribution is as follows: Gennaro Nolano is author of Section 2 and Section 4, Maria Pia di Buono is author of Section 1 and Section 3, and Johanna Monti is author of Section 5 and supervised the project.

References

1. Baldwin, T.: Deep lexical acquisition of verb–particle constructions. Computer Speech Language **19**(4), 398–414 (2005). https://doi.org/https://doi.org/10.1016/j.csl.2005.02.004, https://www.sciencedirect.com/science/article/pii/S0885230805000070, special issue on Multiword Expression
2. Batchelor, C.R., Corbett, P.T.: Semantic enrichment of journal articles using chemical named entity recognition. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. pp. 45–48. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), https://aclanthology.org/P07-2012
3. Boudin, F.: Unsupervised keyphrase extraction with multipartite graphs (2018). https://doi.org/10.48550/ARXIV.1803.08721, https://arxiv.org/abs/1803.08721
4. Buitelaar, P., Cimiano, P.: Bridging the gap between text and knowledge **167**, v–ix (01 2008)
5. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics **16**(1), 22–29 (1990), https://aclanthology.org/J90-1003
6. Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Survey: Multiword expression processing: A Survey. Computational Linguistics **43**(4), 837–892 (Dec 2017). https://doi.org/10.1162/COLI_a_00302, https://aclanthology.org/J17-4005
7. Desmontils, E., Jacquin, C., Simon, L.: Ontology enrichment and indexing process (06 2003)
8. Dojchinovski, M., Sasaki, F., Gornostaja, T., Hellmann, S., Mannens, E., Salliau, F., Osella, M., Ritchie, P., Stoitsis, G., Koidl, K., Ackermann, M., Chakraborty, N.: FREME: Multilingual semantic enrichment with linked data and language technologies. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 4180–4183. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1660
9. Freeman, L.: A set of measures of centrality based on betweenness. Sociometry **40**, 35–41 (03 1977). https://doi.org/10.2307/3033543
10. Hassanzadeh, O., Trewin, S., Gliozzo, A.: Semantic Concept Discovery over Event Databases, pp. 288–303 (06 2018). https://doi.org/10.1007/978-3-319-93417-4_19
11. Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering **1**(1), 9–27 (1995). https://doi.org/10.1017/S1351324900000048
12. Lin, D., Pantel, P.: Concept discovery from text. In: COLING 2002: The 19th International Conference on Computational Linguistics (2002), https://aclanthology.org/C02-1144
13. M.K, V., Kavitha, K.: A survey on similarity measures in text mining (2016)

14. Pecina, P., Schlesinger, P.: Combining association measures for collocation extraction. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. pp. 651–658. Association for Computational Linguistics, Sydney, Australia (Jul 2006), https://aclanthology.org/P06-2084
15. Shen, Z., Wu, C.H., Ma, L., Chen, C.P., Wang, K.: SciConceptMiner: A system for large-scale scientific concept discovery. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. pp. 48–54. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-demo.6, https://aclanthology.org/2021.acl-demo.6
16. Smrz, P., Otrusina, L.: Semantic enrichment across language: A case study of Czech bibliographic databases. In: Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017). pp. 523–532. NLP Association of India, Kolkata, India (Dec 2017), https://aclanthology.org/W17-7563