Edited by

Gloria Corpas Pastor, Miriam Buendía Castro and Rut Gutiérrez Florido

Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives

Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües

2015.

ORGANISING COMMITTEE/COMITÉ ORGANIZADOR

Chair/Presidencia

- Gloria Corpas Pastor

Organisers/Organizadores

- Rosario Bautista Zambrana
- Cristina Castillo Rodríguez
- Hernani Costa
- Isabel Durán Muñoz
- Jorge Leiva Rojo
- Gema Lobillo Mora
- Pablo Pérez Pérez
- Míriam Seghiri Domínguez
- M.ª Cristina Toledo Báez
- Míriam Urbano Mendaña
- Anna Zaretskaya

Secretary/Secretaría

- Míriam Buendía Castro
- Rut Gutiérrez Florido

COMITÉ CIENTÍFICO/ PROGRAMME COMMITTEE

- Margarita Alonso Ramos (Universidad da Coruña, Spain)
- Ignacio Bosque (Universidad Complutense de Madrid, Spain)
- Jenny Brumme (Universitat Pompeu Fabra, Spain)
- František Čermák (Univerzita Karlova v Praze, Czech Republic)
- Jean-Pierre Colson (Université Catholique de Louvain, Belgium)
- Dmitrij Dobrovol'skij (Russian Academy of Sciences, Russia)
- Peter Ďurčo (Univerzita sv. Cyrila a Metoda v Trnave, Slovakia)
- Xesús Ferro Ruibal (Centro Ramón Piñeiro para a Investigación en Humanidades, Spain)
- Sabine Fiedler (Universität Leipzig, Germany)
- Natalia Filatkina (Universität Trier, Germany)
- Thierry Fontenelle (Translation Centre for the bodies of the European Union (CdT), Belgium)
- Miguel Á. García Peinado (Universidad de Córdoba, Spain)
- Jose Enrique Gargallo Gil (Universitat de Barcelona, Spain)
- Maribel González Rey (Universidade de Santiago de Compostela, Spain)
- Annelies Häcki Buhofer (Universität Basel, Switzerland)
- Patrick Hanks (University of Wolverhampton, United Kingdom)
- Ulrich Heid (Universität Hildesheim, Germany)
- Adam Kilgarriff (University of Brighton, United Kingdom)
- Ramesh Krishnamurthy (Aston University, United Kingdom)
- Elvira Manero Richard (Universidad de Murcia, Spain)
- Josep Marco (Universitat Jaume I, Spain)
- Manuel Martí Sánchez (Universidad de Alcalá, Spain)
- Carmen Mellado Blanco (Universidade de Santiago de Compostela, Spain)
- Salah Mejri (Université Paris 13, France)
- Ruslan Mitkov (University of Wolverhampton, United Kingdom)

- Pedro Mogorrón (Universitat d'Alacant, Spain)
- Johanna Monti (Università degli Studi di Sassari, Italy)
- Esteban T. Montoro del Arco (Universidad de Granada, Spain)
- Rosamund Moon (University of Birmingham, United Kingdom)
- Carmen Navarro (Universita' degli studi di Verona, Italy)
- Michael Oakes (University of Wolverhampton, United Kingdom)
- Inés Olza (Universidad de Navarra, Spain)
- Antonio Pamies (Universidad de Granada, Spain)
- Inmaculada Penadés Martínez (Universidad de Alcalá, Spain)
- Rosa Piñel (Universidad Complutense de Madrid, Spain)
- Carlos Ramisch (Aix Marseille Université, France)
- Leonor Ruiz Gurillo (Universitat d'Alacant, Spain)
- Agata Savary (Université Francois Rabelais, France)
- Elmar Schafroth (Universität Düsseldorf, Germany)
- Violeta Seretan (Université de Genève, Switzerland)
- Julia Sevilla Muñoz (Universidad Complutense de Madrid, Spain)
- Inès Sfar (Université Paris 13, France)
- Kathrin Steyer (Institut für Deutsche Sprache, Germany)
- Joanna Szerszunowicz (Uniwersytet w Białymstoku, Poland)
- Aina Torrent (Fachhochschule Köln, Germany)
- Eric Wehrli (Université de Genève, Switzerland)
- Gerd Wotjak (Universität Leipzig, Germany)
- Stefanie Wulff (University of Florida, United States)
- Pablo Zamora (Universidad de Murcia, Spain)

INVITED SPEAKERS/CONFERENCIANTES PLENARIOS

Jean-Pierre Colson

Professor of Translation Studies and Linguistics at the Université Catholique de Louvain, and President of the Louvain School of Translation and Interpreting (LSTI).

"The contribution of corpus-based phraseology to translation studies: from experiments to theory"

29 June 2015/29 de junio de 2015

The notion of phraseology is now used across a wide range of linguistic disciplines: Phraseology (proper), Corpus Linguistics, Discourse Analysis, Pragmatics, Cognitive Linguistics, Computational Linguistics. It is, however, conspicuously absent from most studies in the area of Translation Studies (e.g. Delisle 2003, Baker & Saldanha 2011). The paradox is that many practical difficulties encountered by translators and interpreters are directly related to phraseology in the broad sense (Colson 2008, 2013), and this can most clearly be seen in the failure of SMT-models (statistical machine translation) to deal efficiently with the translation of set phrases (used here as a generic term for all categories of phraseological constructions, from collocations to proverbs).

Although corpus-based and computational phraseology still need to be clearly delineated from other concurrent disciplines, a possible way of narrowing the gap between phraseology and translation studies is proposed here: the recourse to experiments involving on the one hand set phrases and, on the other, evidence from parallel translation corpora or SMT-machines such as Google Translate. We will argue that both phraseology and translation studies have much to gain from this cross fertilisation, because both disciplines are regularly criticised for their lack of coherent terminological description and for the insufficient number of reproducible experiments they involve. The aim of this paper is not to draw up an exhaustive list of the possible experiments showing

the interweaving of phraseology and translation studies, but to propose directions for future research involving a number of key issues that are posed by phraseology and are illustrated by translation practice.

A first series of experiments relating to this subject matter concerns the problems posed by phraseology to human translation. Decoding phraseology in the source text is far from easy for translators and interpreters, all the more so as they are usually not native speakers of the source language. Also, finding a natural formulation in the target language and avoiding translationese requires an excellent mastery of the phraseology of the target language. I will argue that experiments with translation corpora may precisely shed some light on some crucial notions of phraseology and of translation studies. Experiments have shown that translation errors due to phraseology are legion in many translation corpora, even in the official translations of the European Union. A contribution of corpus-based phraseology would therefore consist in making human translators aware of the pitfalls of phraseology in the source text. Even experienced professionals sometimes fail to detect the fixed or semi-fixed character of a source text construction. Experiments along these lines should therefore also include the creation of large, multilingual phraseological databases, which brings us back to two serious shortcomings of computational phraseology:

1. There is no universally accepted algorithm for the automatic extraction of phraseology, especially not for ngrams larger than bigrams.
2. There is no consensus as to the proportion of set phrases in relation with the rest of the vocabulary: according to Jackendoff (1995), there are about as many fixed expressions as there are single words in the dictionary, but others (such as Mel'čuk 1995) hold the view that fixed expressions far outnumber single words.

I will argue in that respect that algorithms derived from text mining and information retrieval techniques (Baeza-Yates, R. & B. Ribeiro-Neto 1999) can be efficient and (computationally) cost-effective in order to build up unfiltered collections of recurrent fixed or semi-fixed phrases, from which translators could gain information about the number of set phrases in the source text. Such an algorithm has been proposed in Colson (2014), and a provisional database of about 700,000 English set phrases (tokens) has been assembled, which seems

to confirm that Jackendoff's view about the total number of fixed expressions was not correct.

A second series of experiments that would turn out to be profitable to a better theoretical understanding of both phraseology and translation studies, has to do with the specific problems posed by phraseology to automatic translation. Phraseology has only recently been identified as one of the main sources of errors in automatic translation systems, including the most recent SMT-systems (Monti, Mitkov, Corpas Pastor & Seretan 2013). I will however point out that the theoretical underpinnings of phraseology are at stake in order to provide a coherent explanation for the serious shortcomings in the automatic translation of sentences containing phraseology. The crux of the matter seems to be the complex interplay between association and frequency in fixed expressions. Recent evidence shows that, contrary to what is assumed by most statistical scores, there should be no relationship between the statistical association of the grams constituting a set phrase, and its frequency in a huge corpus. The countless examples of wrong translations of phraseologically rich sentences by Google Translate, for instance, all point to the fundamentally wrong way in which ngrams were traced down, namely by giving the highest priority to frequency.

Further experimentation should also shed some light on the overall statistical distribution of set phrases in large corpora. The well-know zipfian distribution of words in a corpus poses theoretical problems as far as phraseology is concerned. Corpus-based studies (Baroni 2008) indicate that the distribution of ngrams themselves may display a Zipf-Mandelbrot curve. This is an important theoretical challenge to the theory of phraseology and also to semantics, having therefore consequences on the way meaning may be expressed in different languages and be adequately translated from one language into another. I will point out that a general theory of phraseology, as outlined by Mejri (2006), may offer a new insight into the statistical underpinnings of both morpheme associations (in words) and of word association (in set phrases).

References

BAEZA-YATES, R. & B. RIBEIRO-NETO (1999). *Modern Information Retrieval*. New York: ACM Press, Addison Wesley.

BAKER, M. & G. SALDANHA (EDS.) (2011). *Routledge Encyclopedia of Translation Studies.* New York: Routledge.

BARONI, M. (2008). Distributions in text. In: A. Lüdeling & M. Kytö, (eds.), *Corpus linguistics. An international handbook*. Berlin, New York: Walter de Gruyter, p. 803-821.

BARONI, M., BERNARDINI, S., FERRARESI, A. & E. ZANCHETTA. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation*, 43, p. 209-226.

COLSON, J.-P. (2008). Cross-linguistic phraseological studies: An overview. In: Granger, S. & F. Meunier (eds.), *Phraseology. An interdisciplinary perspective*. John Benjamins, Amsterdam / Philadelphia, p. 191-206.

COLSON, J.-P. (2010a). The Contribution of Web-based Corpus Linguistics to a Global Theory of Phraseology. In: Ptashnyk, S., Hallsteindóttir, E. & N. Bubenhofer (eds.), *Corpora, Web and Databases. Computer-Based Methods in Modern Phraselogy and Lexicography*. Hohengehren, Schneider Verlag, p. 23-35.

COLSON, J.-P. (2010b). Automatic extraction of collocations: a new Web-based method. In: S. Bolasco, S., Chiari, I. & L. Giuliano, *Proceedings of JADT 2010,Statistical Analysis of Textual Data*, Sapienza University of Rome, 9-11 June 2010. Milan, LED Edizioni, p. 397-408.

COLSON, J.-P. (2013). Pratique traduisante et idiomaticité : l'importance des structures semi-figées. In: Mogorrón Huerta, P., Gallego Hernández, D., Masseau, P. & Tolosa Igualada, M. (eds.), *Fraseología, Opacidad y Traduccíon. Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation* (Herausgegeben von Gerd Wotjak). Frankfurt am Main, Peter Lang, p. 207-218.

COLSON, J.-P. (2014). Set phrases around *globalization* : an experiment in corpus-based computational phraseology. Paper presented at *CILC 2014, 6th International Conference on Corpus Linguistics*. University of Las Palmas de Gran Canaria, 22-24 May 2014.

CORPAS PASTOR, G. (2013). Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In Olza, I. & R. Elvira Manero (eds.) *Fraseopragmática*. Berlin: Frank & Timme, p. 335-373.

DELISLE, J. (2003). La traduction raisonnée. Ottawa: Presses de l'Université d'Ottawa.

JACKENDOFF, R. (1995). The boundaries of the lexicon. In M. Everaert, E.-J. van der Linden, A. Schenk & R. Schroeder (eds.), *Idioms: Structural and psychological perspectives*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, p. 133-165.

MEJRI, S. (2006). Polylexicalité, monolexicalité et double articulation. *Cahiers de Lexicologie*, 2 :209-221.

MEL'CUK, I. 1995. Phrasemes in language and phraseology in linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk & R. Schroeder (eds.), *Idioms: Structural and psychological perspectives*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, p. 167-232.

MONTI, J., MITKOV, R., CORPAS PASTOR, G. & V. SERETAN (EDS) (2013). Workshop Proceedings: *Multi-word units in machine translation and translation technologies*, Nice 14th Machine Translation Summit.

Ulrich Heid

Professor of computational linguistics and language technology at University of Hildesheim

"Extracting linguistic knowledge about collocations from corpora"

29 June 2015/29 de junio de 2015

We start from the assumption that (lexical) collocations and verbal idioms are a type of multiword expressions which deserve a detailed linguistic and lexicographic description (Gouws/Heid 2006: treatment units of their own in dictionaries). If this is so, then there is a need for corpus-based tools which allow us to find out about the contextual (= syntagmatic) and paradigmatic properties of collocations. We intend to show how much of these data can be identified with acceptable quality in large enough corpora.

Syntagmatic properties have to do with the distributional behavior of collocations: preferences in number (have high hopes, pl.), determination (article use) or modifiability are well-known examples; some collocations and many verbal idioms have their own syntactic valency constructions (cf. be in a position [+to+INF]) or they co-occur preferentially with certain lexical items, e.g. as modifiers (cf. DE Kritik üben ("criticize") which prefers adjectives that typically collocate with Kritik: harsche, scharfe Kritik üben ("criticize severely"), cf. Häcki-Buhofer et al. 2014).

Examples of paradigmatic properties include the exchangeability of lexical elements of the collocation against synonyms, or the availability of nominalizations (submit a proposal – submission of a proposal) or of compounds in Germanic languages (DE Antrag einreichen ("submit a proposal") – Einreichung eines Antrags – Antragseinreichung). Another example are pragmatic marks and preferences with respect to domain-specific languages.

We will give examples of such data from German, English, French and Italian and we will assess to what extent such linguistic knowledge may be needed in translation and in (mother tongue or foreign language) text production. Thereafter, we intend to show which types of data of the above kind can be

extracted with acceptable quality from corpus texts, and with which language processing techniques; we claim that state of the art dependency parsing provides a fair amount of such data thus facilitating the description work of terminologists and lexicographers.

References

GOUWS, RUFUS H. AND HEID, ULRICH. 2006. “A model for a multifunctional dictionary of collocations”, in Corino, Elisa et al. (Eds.): Proceedings of the XIIth EURALEX International Congress, (Alessandria: Edizioni dell’Orso), 979 – 988.

HÄCKI-BUHOFER, ANNELIES ET AL. 2014. Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag, (Tübingen: Francke).

Patrick Hanks

Professor in Lexicography at the Research Institute of Information and Language Processing in the University of Wolverhampton.

"Meaning and Phraseology: a Corpus-Driven Approach"

30 June 2015/30 de junio de 2015

Interesting aspects of the meaning are revealed by corpus-driven lexical analysis. The typical function of nouns is to create referring expressions—terms that either refer to objects in the world or denote abstract concepts. The typical function of verbs, on the other hand, is to create propositions, in which noun phrases play roles that are mediated by a verb. According to the Theory of Norms and Exploitations (Hanks, 1994, 2004, 2013), a verb has only meaning potential (not meaning as such) until it is put in context. There is no 'semantic invariable' that is common to all normal uses of a verb. Consider the verb blow. 'A gale was blowing', 'They blew up the bridge', 'He blew his nose', and 'She blew the whistle on government malpractice' have little in common semantically, but all four sentences represent realizations of conventional lexico-syntactic patterns of English. The meanings lie in collocation and phraseology, not in the words themselves.

Different questions must be asked about nouns and verbs, and different apparatuses are required for corpus analysis of these two categories. When shower is used as a noun, we can ask how many different kinds of shower there are—rain showers, snow showers, spring showers, etc., as opposed to bathroom showers and power-driven showers. What distinctive properties or common features does each category have? On the other hand, if shower is used as a verb, relevant question are prompted by the collocates and syntagmatics: for example, 'Is it normal to say in English, "It showered yesterday"?' Patterns with prepositions such as with and on prompt questions such as 'Who showers what on whom?' 'Who showers whom with what?' 'What is the relationship between such patterns?' In this way, we can start to compile an inventory of patterns of word use that seem to be already available to the unconscious minds of users of a language.

Ruslan Mitkov

Director of the Research Institute of Information and Language Processing (RIILP) and Professor of Computational Linguistics and Language Processing at the University of Wolverhampton

"Knowledge- and resource-poor identification and translation of multiword expressions across language pairs"

30 June 2015/30 de junio de 2015

The correct identification, interpretation and translation of multiword expressions in both general and specialised languages, is vital for the successful operation of most Natural Language Processing applications and computer-aided tools supporting various users including phraseologists, translators, interpreters, terminologists, language learners and teachers. As parallel corpora are scarce and in order to benefit from the wider availability of comparable corpora (which can be also compiled with a specific task in mind), there is a pressing need to develop approaches which can extract and translate multiword terms from comparable corpora. In this presentation the speaker will propose a novel methodology based on computing semantic similarity to extract and translate multiword units from comparable corpora termed as 'knowledge- and resource-poor' as it is not dependent on the use of any dictionaries or linguistic knowledge. While this particular study focuses on English and Spanish, the methodology is not restricted to any particular pair of languages.

TABLE OF CONTENTS/ÍNDICE DE CONTENIDOS

CORPUS-BASED PHRASEOLOGY
FRASEOLOGÍA BASADA EN CORPUS

NLP AND/OR CORPUS-BASED IDENTIFICATION AND CLASSIFICATION OF PHRASEOLOGICAL UNITS

IDENTIFICACIÓN Y CLASIFICACIÓN DE UNIDADES FRASEOLÓGICAS BASADA EN CORPUS O MEDIANTE TÉCNICAS DE PLN

COMPUTER-AIDED AND/OR CORPUS-BASED ANALYSIS OF PHRASEOLOGICAL UNITS

ANÁLISIS DE UNIDADES FRASEOLÓGICAS BASADO EN CORPUS O ASISTIDO POR ORDENADOR

PHRASEOLOGY IN E-LEXICOGRAPHY AND E-TERMINOLOGY

LA INFORMACIÓN FRASEOLÓGICA EN LA LEXICOGRAFÍA Y LA TERMINOLOGÍA ELECTRÓNICAS

Corpus-based phraseology Fraseología basada en corpus

COLOCACIONES EN UN CORPUS DE LECTURAS GRADUADAS

Margarita Alonso Ramos
Universidade da Coruña
lxalonso@udc.es

Marcos García Salido
Universidade da Coruña
marcos.garcias@udc.es

Ana Orol González
Universidade da Coruña
ana.orol.gonzalez@udc.es

Orsolya Vincze
Universidade da Coruña
ovincze@udc.es

Entre las unidades fraseológicas, las colocaciones son las que cuentan con menos recursos lingüísticos en español. Solo disponemos de dos diccionarios (Bosque 2004 y Alonso Ramos 2004) y los materiales didácticos orientados al mundo del Español como Lengua Extranjera (ELE) que traten las colocaciones son escasos (Prada et al. 2012). Una de las principales carencias es que no existe todavía una clasificación de las colocaciones por los niveles del *Marco común europeo de referencia* (MCER). Dicha clasificación sería especialmente necesaria para seleccionar las colocaciones que deben aparecer en diferentes manuales correspondientes a diferentes niveles. Con el objetivo de examinar cómo están niveladas las colocaciones en materiales didácticos ya nivelados, nos propusimos examinar un corpus de lecturas graduadas de ELE, en particular, los textos que aparecen en el siguiente enlace del Centro virtual Cervantes (http://cvc.cervantes.es/aula/lecturas/).

El objetivo de este trabajo es presentar la metodología utilizada para la anotación de colocaciones en ese corpus, así como el perfil de las colocaciones ahí encontradas. Nuestro trabajo está enmarcado en la Lexicología explicativa y combinatoria (Mel'čuk 2012) en donde el concepto de colocación no se basa en frecuencias sino en la coocurrencia léxica restringida entre los dos elementos componentes de la colocación, la base y el colocativo (como *miedo cerval*, *odio mortal*, pero no **miedo mortal* u **odio cerval*). Por esta razón, el trabajo de anotación fue semiautomático. Para la anotación adaptamos el programa *Knowtator* (Ogren 2006) que permite anotar las bases y los colocativos individualmente y como constituyentes de la colocación. Una vez identificadas las colocaciones en los textos correspondientes a tres niveles de ELE (inicial, intermedio y avanzado), buscamos algún rasgo que pudiera explicar su distribución. Dado que la frecuencia es uno de los elementos más utilizados para nivelar el vocabulario, optamos por buscar las frecuencias de las colocaciones en un corpus de español utilizado como referencia (*esTenTen*, Kilgarriff et al. 2014). Los resultados muestran que las colocaciones de frecuencia muy baja se usan menos en el nivel inicial que en el intermedio y en este menos que en el avanzado, por lo que parece confirmarse la tendencia también presente en el léxico general de asignar niveles más avanzados a las palabras menos frecuentes. Queda por confirmar si esta tendencia también existe en otros corpus de lecturas graduadas.

En la presentación por extenso mostraremos en detalle los datos cuantitativos así como los posibles criterios de nivelación de colocaciones que se pueden derivar a partir de nuestros resultados.

References

ALONSO RAMOS, M. (2004). *Diccionario de colocaciones del español*, [online] Disponible en: <http://www.dicesp.com> [Accessed 15 de marzo 2015]

BOSQUE, I. (2004). *Diccionario combinatorio práctico del español*. Madrid: Ediciones SM.

KILGARRIFF, A. RYCHLY, P. SMRZ, P. AND TUGWELL, D. (2004). The Sketch Engine. In: G. Williams and S. Vessier, eds. (*Proceedings of the Eleventh EURALEX International Congress, Euralex 2004*. Lorient: Université de Bretagne-Sud. pp. 105–116.

MEL'ČUK, I. (2012). Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology*, *3*, 31–56.

OGREN, P. V. (2006). Knowtator: A Protégé plug-in for annotated corpus construction. In: *Proceedings of the Human Language Technology*

Conference of the NAACL. Companion Volume. New York: Association of Computational Linguistics. pp. 273–275.

PRADA, M. DE; SALAZAR, S. Y MOLERO, C.M. (2012). *Uso interactivo del vocabulario. Y sus combinaciones más frecuentes*. Madrid: Edelsa.

LOCUCIONES IDIOMÁTICAS EN EUSKERA: NECESIDAD Y PAUTAS PARA SU RECOPILACIÓN Y ORDENACIÓN A PARTIR DE LOS CORPUS TEXTUALES EXISTENTES

Xabier Altzibar Aretxabaleta

Universidad del País Vasco

josejavier.alcibar@ehu.eus

Xabier Bilbao López

Universidad del País Vasco

xabier.bilbao@ehu.eus

Hoy día existen valiosos corpus textuales en euskera que facilitan la búsqueda de unidades fraseológicas a partir de palabras clave (EPG, EPD, LBC, WCA, XXMECE), si bien las herramientas de búsqueda de dichos corpus no están específicamente orientadas a la extracción de fraseologismos (tan solo se pueden buscar pares frecuentes de palabras en algunos corpus). Existen también repertorios fraseológicos consultables online (Mokoroa 1990), así como un diccionario de autoridades que contiene entradas de locuciones (DGV-OEH).

Sin embargo, a pesar de contar con estas colecciones, hechas en diferentes épocas y con diferentes criterios, la lengua vasca carece de una recopilación de locuciones confeccionada y organizada con un punto de vista actual, de manera que sirva de utilidad para la miríada de traductores, enseñantes, periodistas y demás profesionales. Sería especialmente valiosa una recopilación, ordenación y explicación de las locuciones léxicas figurativas o

idiomáticas (de diversos grados), dado el alto valor expresivo de las mismas. Denominamos locuciones a un determinado tipo de unidades fraseológicas, propias del sistema de la lengua, que tienen fijación interna, unidad de significado y fijación externa pasemática (Corpas 1996).

Pensamos que hay razones urgentes para la recopilación que proponemos. Constatamos la pérdida constante de locuciones, sobre todo idiomáticas, en el uso real del idioma hablado y escrito, además de un creciente recurso a calcos del español, y, por otra parte, creemos que los agentes de la normativización no han prestado a los fraseologismos la atención que merecen ni han conseguido fomentar el uso de los mismos al primar el euskera formal de tipo académico sobre el euskera popular y hablado. Ciertamente existen dificultades para la normalización y/o normativización de las locuciones, debido a la fragmentación dialectal y a la gran variedad sintáctica y léxica. Por ello, un trabajo actual de recopilación y ordenación de locuciones idiomáticas podría contribuir a impulsar un mayor uso de las mismas y orientar en la normativización.

En un trabajo como el que proponemos, el lexicógrafo o fraseólogo debería, a nuestro juicio, seleccionar prioritariamente las locuciones figurativas o idiomáticas realmente usadas. Asimismo, debería establecer la forma de las locuciones, dar información de las variantes y de la categoría (y, a veces, del registro o nivel de uso, actitud del hablante, ámbito, frecuencia y extensión geográfica o dialectal), definir o explicar el significado o las diversas acepciones de la locución, y dar al menos un ejemplo actual en su contexto o uno por cada estructura morfosintáctica diferente.

Así pues, presentaremos los puntos de partida, beneficios y procedimiento que tendría el trabajo de recopilación que proponemos, así como una muestra preliminar del mismo. Por otra parte, esperamos aprovechar las aportaciones de los congresistas para obtener nuevas ideas para una utilización eficaz de los corpus y las herramientas de búsqueda de que disponemos.

References

CORPAS PASTOR, G. (1996). *Manual de fraseología española*. (Biblioteca Románica Hispánica. Manuales, 76). Madrid: Gredos.

DGV-OEH: MITXELENA, L., SARASOLA, I. (1987-2005). [3a. edic. (2013]. *Diccionario General Vasco–Orotariko Euskal Hiztegia*. 16 vls. Bilbao:

Academia de la Lengua Vasca – Euskaltzaindia. (Corpus no etiquetado de 6 mill. de palabras). URL: <http://www.euskaltzaindia.eus/oeh>.

EPD: INSTITUTO DE EUSKERA DE LA UNIVERSIDAD DEL PAÍS VASCO, EPD: *Ereduzko Prosa Dinamikoa* [Corpus dinámico de prosa de referencia] (Corpus de textos periodísticos y literarios 2009-2013). URL: <http://ehu.eus/ehg/epd/>.

EPG: INSTITUTO DE EUSKERA DE LA UNIVERSIDAD DEL PAÍS VASCO, EPG: *Ereduzko Prosa Gaur* [Prosa de referencia actual] (Corpus de textos periodísticos y literarios 2001-2008; 25 mill. de palabras). URL: <http://ehu.eus/euskara-orria/euskara/ereduzkoa/> .

LBC: EUSKALTZAINDIA – ACADEMIA DE LA LENGUA VASCA. *Lexikoaren behatokiaren corpusa*. [Corpus del observatorio del léxico] (Corpus de textos periodísticos del s. XXI; 11 mill. de palabras)] URL: <http://lexikoarenbehatokia.euskaltzaindia.eus>.

MOKOROA, J. M. (1990). *Ortik eta emendik. Repertorio de locuciones del habla popular vasca*. Labayru-Eusko Jaurlaritza-Etor eds. URL: <http://www.hiru.com/hirupedia>.

XXMECE: EUSKALTZAINDIA – ACADEMIA DE LA LENGUA VASCA. *XX. mendeko Euskararen Corpusa*. [Corpus vasco del s. XX] (muestra variada de textos; 4,6 mill. de palabras). URL: <http://www.euskaracorpusa.net/Xxmendea>.

WCA: ELHUYAR. *Web-corpusen Ataria*. [Portal de corpuses de la web] (Corpus de textos extraídos de la web; 125 mill. de palabras, aprox.). URL: <http://webcorpusak.elhuyar.org>.

LA MODALIDAD EN LOS ENUNCIADOS FRASEOLÓGICOS: DELIMITACIÓN Y AUTONOMÍA EN LAS FÓRMULAS RUTINARIAS

María Belén Alvarado Ortega

Universidad de Alicante

belen.alvarado@ua.es

Este trabajo estudia la modalidad en las fórmulas rutinarias para delimitar su grado de independencia y autonomía en los enunciados fraseológicos. Tal y como afirmamos en Alvarado (2010), las fórmulas rutinarias son formas lingüísticas que pueden codificar la modalidad del enunciado, que se centra en mostrar la actitud que tiene el hablante con respecto al mensaje. Las fórmulas rutinarias se encuentran dentro de la Esfera III de las UFs (Corpas 1996), los llamados enunciados fraseológicos, que tienen carácter de enunciado y están fijados en el habla. Esta denominación se basa, en gran medida, en los presupuestos que ya había dado Casares (1950) y en los trabajos de Zuluaga (1980, 1992). Así, las fórmulas rutinarias se encuentran en esta esfera y, como UFs que son, deben poseer las características comunes a todas ellas, la fijación y, en ocasiones, la idiomaticidad, pero además, pueden presentar algún tipo de independencia como enunciados fraseológicos que son. Consideramos que toda fórmula rutinaria posee fijación formal, entendida como perdurabilidad de los componentes que la constituyen, y fijación psico-lingüística, referida a la convencionalización en la comunidad lingüística, es decir, a la estabilidad en su

reproducción y a su frecuencia de uso. Sin embargo, para las fórmulas rutinarias estos rasgos definitorios se pueden dar de manera gradual. Observaremos si estos datos se dan de igual manera en todos los enunciados fraseológicos que componen la Esfera III.

En este contexto aplicaremos el sistema de Briz y el Grupo Val.Es.Co. (2003, 2014) para segmentar y delimitar las fórmulas rutinarias en la conversación y establecer su autonomía. Este sistema, que delimita las unidades en la conversación, distingue en el orden estructural del plano monológico entre intervención, acto y subacto. La característica principal del acto es la aislabilidad que presenta con respecto al resto de unidades. Por tanto, esta aislabilidad se corresponde a su vez con la independencia en el habla que poseen las fórmulas. Así, toda fórmula rutinaria debería ser un acto; sin embargo, comprobaremos que no siempre se cumple esta afirmación. Este sistema de unidades será aplicado de igual modo a las UFs que componen la Esfera III, para constatar si funcionan como enunciados independientes y tienen características similares a las fórmulas rutinarias.

La metodología empleada se corresponde con el enfoque fraseológico y pragmático, y los ejemplos se han extraído del Corpus de Conversaciones Coloquiales de Briz y el Grupo Val.Es.Co. (2002).

Así, a partir de la extracción de datos estudiaremos los contextos de uso de las fórmulas rutinarias y comprobaremos si son o no enunciados independientes. Además, este estudio va a hacer que reestructuremos la Esfera III enunciada por Corpas (1996) para delimitar las unidades fraseológicas que la componen, ya que veremos que muchas de ellas no cumplen esas características comunes.

References

ALVARADO ORTEGA, M.B. (2010). *Las fórmulas rutinarias: teoría y aplicaciones*. Frankfurt. Peter Lang.

BRIZ, A. y GRUPO VAL.ES.CO. (2002). *Corpus de conversaciones coloquiales*. Madrid, Arco Libros.

BRIZ, A. y GRUPO VAL.ES.CO. (2003). «Un sistema de unidades para el estudio del lenguaje coloquial*», Oralia*, 6, págs. 7-61.

CASARES, J. ([1950] 1969). Introducción a la lexicografía moderna. Madrid, Revista de Filología Española, Anejo LII.

CORPAS, G. (1996). *Manual de Fraseología Española*, Madrid, Gredos.

ZULUAGA, A. (1980*). Introducción al estudio de las expresiones fijas*. Tübingen, Max Hueber, Verlag.

FRASEOLOGÍA ¿SEXISTA?: LO QUE EL CORPUS ESCONDE

Gloria Corpas Pastor

Universidad de Málaga

gcorpas@uma.es

En este trabajo abordamos los usos sexistas del lenguaje, con especial referencia a la fraseología como uno de los recursos más efectivos, pero menos estudiados desde la perspectiva de género. El lenguaje es un reflejo del androcentrismo típico de la sociedad. Por androcentrismo entenderemos el estudio, análisis o investigación realizado desde una perspectiva eminentemente masculina, que se presenta como central a la experiencia humana. Desde esta visión del mundo, el varón es la referencia por defecto, y la mujer, "lo otro". El sexismo lingüístico es una de las consecuencias del androcentrismo, como también lo es la discriminación y la (in)visibilidad de "lo otro". Los estereotipos, los roles preestablecidos, la publicidad o los medios de comunicación, entre otros, contribuyen sin duda a alguna a perpetuar actitudes discriminatorias y de exclusión social. Los movimientos feministas han llamado la atención desde hace varias décadas sobre estos aspectos y su reflejo en el lenguaje.

En cierto sentido, las guías de lenguaje no sexista han servido para denunciar la invisibilidad social de la mujer, y reivindicar, al mismo tiempo, su derecho fundamental a la igualdad y a alcanzar el protagonismo que se le ha venido negado sistemáticamente. No obstante, dichas guías se han ocupado casi exclusivamente del masculino genérico, proponiendo otras formas y posibilidades que ofrece la propia lengua para expresar el plural. Otros

aspectos estudiados han sido los duales aparentes, las formas de tratamiento y las denominaciones de profesiones tradicionalmente reservadas a los varones. La fraseología sexista ha sido objeto de estudio con respecto a los valores e ideas contenidos en refranes y algunas locuciones que denigran la imagen de la mujer, expresan misoginia, o fomentan actitudes misóginas o de dominación hacia la mujer. Algunos ejemplos son La mujer en casa y con la pata quebrada, Palabra de mujer no vale un alfiler, Más tiran dos tetas que dos carretas y similares. Pero frente a los ejemplos típicos de sexismo en el lenguaje, de los cuales forma parte también la fraseología explícitamente sexista, existen otras unidades aparentemente inocuas, pero de contenido extraordinariamente machista. De ellas nos ocuparemos en este trabajo, en el cual vamos a realizar un estudio pormenorizado de diversas unidades fraseológicas con metodología de corpus. Estos sesgos sutiles del lenguaje no pueden ser desmontados con el seguimiento de simples guías de lenguaje no sexista, por cuanto conllevan implicaturas y valoraciones sancionadas por la comunidad hablante. Tales matices no son fácilmente detectables y, por tanto, requieren ser descubiertos mediante el análisis de corpus. Los aspectos de género desde el punto de vista de la lingüística de corpus, van mucho más allá del uso del masculino o femenino, o de la explicitación sistemática de la relación entre género y sexo.

References

AMORÓS, C. (DIR). (1995). *10 palabras clave sobre mujer*. Estella: Verbo Divino.

CALERO FERNÁNDEZ, M.ª A.; FORGAS BERDET, E.; LLEDÓ CUNILL, E. (2004). *De mujeres y diccionarios: evolución de lo femenino en la 22. a edición del DRAE.* Madrid: Ministerio de Trabajo y Asuntos Sociales, Instituto de la Mujer.

COLSON, J.-P. (2010). "The Contribution of Web-Based Corpus Linguistics to a Global Theory of Phraseology". En Ptashnyk, S., Hallsteindóttir, E., Bubenhofer, N. 2010. (Eds.). *Corpora, Web and Databases. Computer-Based Methods in Modern Phraseology and Lexicography*. Hohengehren: Schneider Verlag. 23-35.

CORPAS PASTOR, G. (1996). *Manual de fraseología española*. Madrid: Gredos.

CORPAS PASTOR, G. (2013). "Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas". En: I. Olza y E. Manero (eds.): Fraseopragmática. Berlín: Frank & Timme, 335-373.

GILMAN, C. P. 1911. *The Man-made World, or Our Androcentric Culture*. Nueva York: Charlton Company.

LAKOFF, G. (1987). *Women, Fire and Dangerous Things*. Chicago: University of Chicago Press.

MONTI, J.; MITKOV, R.; CORPAS PASTOR, G.; SERETAN, V. (EDS.). (2013). *Workshop Proceedings for: Multi-word units in Machine Translation and Translation Technologies*. Allschwil (Suiza): European Association for Machine Translation (EAMT). <http://www.mtsummit2013.info/files/proceedings/WkSh4_proceedings.pdf>

SYSTEMATIC VS. NON-SYSTEMATIC COLLOCATIONAL PATTERNS IN LSP: PARADIGMATIC VARIATION IN THE TECHNICAL DOMAIN

Laura Giacomini

University of Heidelberg

lgiacomini@yahoo.com

This paper presents detailed empirical observations on the topic of collocational variation in LSP texts of the technical domain and on its relevance for text production and translation (Kerremans 2010, Schmitt 2006). The low level of linguistic standardization of this specialised field allows for a high degree of variability in the way in which languages encode domain-related concepts. Despite the great heterogeneity of variation examples, however, encoding often takes place along systematic and reproducible combinatory models, which this study aims to capture with the help of small-sized comparable corpora of English, German and Italian texts. These corpora include a restricted sample of domain-specific textual genres (e.g. handbooks, technical data sheets, product catalogues, specialised magazines, support pages offered by companies) with varying level of specialization and intended for a broad public, ranging from experts to non-experts.

The study deals with the technical subdomain of the automotive sector and concentrates on paradigmatic, i.e. lexical variation in collocation, a subtype of formal variation in which a set of two or more collocations refer to the same conceptual item. Different from syntagmatic variation, variation on the

paradigmatic level does not depend on the syntactic properties of a word combination, but on the interchangeability of its lexical constituents. Given a cluster of collocations with the same syntactic structure, this phenomenon may involve any of their constituents; for instance, the Italian collocations meaning “fog light” and displaying the structure A + N allow for variation on the level of the noun/base (*faro fendinebbia, proiettore fendinebbia, luce fendinebbia*) as well as on the level of the adjective/collocate (*faro fendinebbia, faro antinebbia*).

Context, in which semantic proximity (contextual synonymy) of collocation constituents is embedded, is the result of several functional factors, such as communicative situation, specific topic and textual genre. The availability of LSP corpora that include a representative and balanced selection of typical textual genres can be of great advantage in identifying and explain relations between a collocational variant and genre-related features (cf. Roelcke 2010) and not infrequently discloses a direct link between terminology, corporate identity and geographical areas (Sofer 2006). Collocations are extensively understood as phraseological n-grams including multiword expressions and compounds (cf. Roth 2014), a comprehensive view that enables better cross-linguistic evaluation of paradigmatic variation (e.g. the nominal compounds En. *fog light/fog lamp* = It. *antinebbia/fendinebbia* = De. *Nebelscheinwerfer/Nebelleuchte*). Moreover, the paper analyses from a contrastive perspective the phraseological status of different kinds of borrowings (Giacomini 2012), for instance synonymic calques (*brake drum*, *Bremstrommel*, *tamburo del freno*).

The findings from the study on paradigmatic variation of collocation in domain-specific contexts have been summarised in a variational description model, which brings to light the intrinsic concept-oriented structure of terminology and the frequent lack of a one-to-one correspondence between concepts and designations. At the same time, the study highlights the phraseological relevance of the variational phenomenon for text-productive and translation purposes (Quirion 2014, Scarpa 2008, Delpech 2011), hinting at a possible application of the description model to other specialised domains.

References

DELPECH, E. (2011). Un protocole d’évaluation applicative des terminologies

bilingues destinées à la traduction spécialisée. *RNTI - Revue des Nouvelles Technologies de l'information* 2011, pp. 23–48.

GIACOMINI, L. (2012). Lexical borrowings in German and Italian IT terminology: At the crossroads between language interference and translation procedures. In: *Proceedings of the BDÜ Conference "Übersetzen in die Zukunft 2012"*, Berlin 28-30.09.12.

KERREMANS, K. (2010). A Comparative Study of Terminological Variation in Specialised Translation. In: C. Heine / J. Engberg, eds. *Reconceptualizing LSP*. *Online proceedings of the XVII European LSP Symposium 2009*, Aarhus

QUIRION, J. (2014). La mesure de la variation terminologique comme indice de l'évolution des connaissances dans un environnement bilingue. In: R. Temmerman/M. Van Campenhoudt, eds. *Dynamics and Terminology. An interdisciplinary perspective on monolingual and multilingual culture-bound communication*. Amsterdam/Philadelphia: John Benjamins. pp. 281-302.

ROELCKE, T. (2010). *Fachsprachen*, Berlin: Erich Schmidt.

ROTH, T. (2014). *Wortverbindungen und Verbindungen von Wörtern. Lexikografische und distributionelle Aspekte kombinatorischer Begriffsbildung zwischen Syntax und Morphologie*. Tübingen: Francke Verlag.

SCARPA, F. (2008). *La traduzione specializzata*, Milano: Hoepli.

SCHMITT, P. A. (2006). *Translation und Technik*. Tübingen: Stauffenburg.

SOFER, M. (2006). *The Translator's Handbook*. 6th edition. Rockville USA: Schreiber Publishing.

COMPUTERISATION OF PHRASE-TO-PHRASE MATCHING FROM A STANDARD MARINE COMMUNICATION PHRASES CORPUS: A PRELIMINARY EMPIRICAL STUDY

María Araceli Losey

Universidad de Cádiz

araceli.losey@uca.es

Standard Marine Communication Phrases (IMO, 2002), formerly Standard Marine Navigational Vocabulary (IMO, 1985), are a collection of phrases conceived by the International Maritime Organization as a restricted language in the specialized setting of maritime communications to enhance maritime safety by avoiding language misunderstandings at sea. These standardized phrases were created for use over VHF radio in bridge external communications and land-based station exchanges or face to face on board communications. Its prominent characteristics include the lexical and grammatical restrictions (controlled vocabulary, modality limitations) and the controlled discourse use (sender and receiver's identification, pattern repetition, message markers, distress, urgency and safety procedural patterns, broadcast entries) they are submitted to. The standardized language restrictions in the vocabulary and phrase structure selection were designed taking into account predictable areas of language confusion and error. Availability of training material suited to the

peculiar pedagogy that a controlled language involves is of utmost importance. In this context, several studies have been made (Strevens & Johnson, 1983; Losey, 2000; Pritchard & Kalogjera, 2000; Cole, Pritchard & Trenkner, 2007; CAPTAINS, 2012). However, there is still a lack of computerised corpus-based material oriented towards the specific gradual practice of the SMCP through its phraseology in semantic scenarios that may train and prepare the learner (or user) for building up predictable questions-to-answers about situations that may emerge at sea.

Within the NLP framework, the present corpus-based study attempts to develop an application for automatic generation of SMCP message phrase replies based on the Question Answering (QG) system for restricted domains (RDQA) (Mollá, Vicedo, 2007) . The integration of domain-specific information is especially useful in situations in which a user needs to know a very specific piece of information and does not have the time. For the research methodology purposes, an inventory of lexical collocations extracted from the written SMCP corpus was developed and matching rules for the association of direct questions, statements and instructions with the appropriate replies were created taking into account semantically defined contextual frameworks. On the other hand, the system architecture shall address the Question Generation (QG) system (Skalban, Ha, Specia, Mitkov, 2012; Liu, Calvo & Rus, 2014) to encompass instances in which a given reply may be extended into a question. Findings revealed during this preliminary empirical study shall be presented and examined. Finally, it is expected that this approach, which to this author's knowledge has not been explored so far, may contribute to SMCP training by e-learning, blended, face-to-face courses or Intelligent Tutoring Systems (ITS).

References

CAPTAINS (2012). *The Captains English Learning Tool. Standalone Course.* [online] Available at: <http://www.captains.pro> [Accessed 10 December 2012].

COLE, C., PRITCHARD, B. AND TRENKNER, P. (2007). Maritime English Instruction –ensuring instructors' competence. *Ibérica*, 14. pp. 123-148.

CORPAS PASTOR, G. (2013). Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In: I. Olza, E. Manero, eds. 2013. *Fraseopragmática.* Berlin: Frank & Timme GmbH. pp. 335-373.

IMO (2002). *IMO Standard Marine Communication Phrases.* London: IMO.

LIU, M., CALVO, R. A., RUS, V. (2014). Automatic generation and ranking of questions for critical review. *Educational Technology & Society* 17(2). pp. 333-346. [online] Available at: <http://www.ifets.info/journals/17_2/27> [Accessed 15 January 2015].

LOSEY LEÓN, M. A. (2000). Facing new changes in the Maritime English curriculum: Tasks' design towards the acquisition of the Standard Marine Communication Phrases. In: Piniella, F. et al. eds. 2000. *Proceedings of the 2nd. International Congress on Maritime Technological Innovations and Research.* Cádiz: Servicio de publicaciones de la Universidad de Cádiz. pp. 1289-1299. *The ESP Journal.* 2(2). pp. 123-129.

MOLLÁ, D., VICEDO, J.L. (2007). Question answering in restricted domains: An overview. *Computational Linguistics.* 33(1). pp. 41-61.

PRITCHARD, B., KALOGJERA, D. (2000). On some features of conversation in maritime VHF communication. In: M. Coulthard, J. Cotterill, F. Rock, eds. 2000. *Dialogue Analysis VII: Working with Dialogue*. Tübingen: Max Niemeyer Verlag. pp. 185-196

SKALBAN, Y., HA, L.A., SPECIA, L. MITKOV, R. (2012). Automatic question generation in multimedia-based learning. *Proceedings of COLING 2012: Posters.* pp. 1151-1160. [online] Available at: <http://aclweb.org/anthology/C12-2112> [Accessed 19 December 2014].

STREVENS, P., JOHNSON, E. (1983). Seaspeak: A project in applied linguistics, language engineering, and eventually ESP for sailors. *The ESP Journal.* 2(2). pp. 123-129.

VARIACIÓN DIALECTAL EN LA FRASEOLOGÍA DEL ESPAÑOL DE MÉXICO: ANÁLISIS CONTRASTIVO DE CORPUS ORALES

Niktelol Palacios

El Colegio de México

niktelolpalaciosc@gmail.com

El *Diccionario del español de México* (DEM), es una obra de lexicografía integral cuyo corpus —*Corpus del español mexicano contemporáneo*, CEMC— está constituido por textos orales y escritos que suman dos millones de ocurrencias, distribuidas en trece géneros (literatura, periodismo, jergas, conversaciones de habla culta y popular, etc.) y abarca la sincronía 1921-1974. En la actualidad el mismo equipo lexicográfico trabaja en el DEM2, para lo cual está construyendo un nuevo corpus (que reproduce los criterios del primero pero abarca una sincronía posterior) y se planea además la redacción del *Diccionario fraseológico del español de México*.

A diferencia del diccionario de lengua general, en el *Fraseológico* se pretende que cada artículo lexicográfico contenga además de definición y ejemplo de uso, el tipo de UF (locución, colocación), la categoría gramatical de cada locución, la materia, la región y el nivel de lengua.

El objetivo de esta ponencia es identificar y analizar unidades fraseológicas que aparecen en tres corpus orales del español de México: *Corpus sociolingüístico de la ciudad de México-PRESEEA*, *Corpus sociolingüístico de*

la ciudad de Puebla y *Corpus del habla de Monterrey-PRESEEA*. Partimos del supuesto de que, por tratarse de materiales orales, podremos registrar sobre todo colocaciones y locuciones coloquiales y por ser corpus locales recolectados por hablantes de la misma comunidad, se favorecerá la documentación de variantes dialectales. Sirvan de ejemplo las siguientes unidades:

1) Corpus de Puebla:

Chido, baril y coqueto loc. adv (*Pue*) (*Pop*) 1 De manera agradable: "Espero te lata la idea y juntos la pasemos *chido, baril y coqueto*"

2) Corpus de la ciudad de México:

Así y asado loc. adv. (*Col*) De esa manera, de alguna forma: "qué le voy a hacer a mi terreno, pues voy a hacer mi casa así y asado"

3) Corpus del habla de Monterrey:

Tirar carrilla loc. verb. (*N* y *Occ*) (*Col*) 1 Hacer burla a alguien: "en mi escuela todos se *tiran carrilla*"

Los corpus que sirven de base a este análisis se encuentran preestratificados con tres variables sociales (sexo, edad e instrucción educativa). Cada entrevista dura como mínimo 45 minutos. Para este análisis se estudian 3 entrevistas por variable (162 en total), lo cual nos da la posibilidad de registrar la riqueza y variedad del español de México y, con ello, evitar estereotipar las hablas coloquiales y populares.

La elección de posibles unidades fraseológicas será obtenida de acuerdo a los siguientes criterios: 1) análisis cuantitativo a partir de un programa automatizado de generación de concordancias; 2) análisis cualitativo (idiomaticidad, fijación y unidad de significado) y 3) contraste de las unidades que aparecen sólo en un corpus con las definidas en el DEM.

Las preguntas que pretendemos responder son: ¿cuáles son los criterios cuantitativos y cualitativos que permiten el reconocimiento de las unidades fraseológicas en un corpus de habla? y ¿cuáles de estas unidades deben llevar una marca de uso regional?

References

LARA, L. F. (2010). *Diccionario del español de México*. México: El Colegio de México.

MARTÍN, P. (2011). *Corpus sociolingüístico de la ciudad de México: materiales de PRESEEA-México*. México: El Colegio de México.

PEJOVIC, A. (2009). De la idea a la locución: hacia la propuesta de un nuevo diccionario fraseológico basado en los ejemplos procedentes de corpus textuales. In: P. Cantos, ed. 1970. *Panorama de investigaciones basadas en corpus*. Murcia: Universidad de Murcia, AELINCO. pp. 430-442.

PENADÉS I. (2008). Proyecto para la redacción de un diccionario de locuciones del español. In: E. Bernal, ed. 2008. *Proceedings of the XIII Euralex International Congress* (Barcelona, 17-19 de julio de 2008). Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada. pp. 1379-1384.

EL TRATAMIENTO DE LAS UNIDADES FRASEOLÓGICAS EN LOS EXÁMENES DELE DEL INSTITUTO CERVANTES

M. ª Ángeles Recio Ariza
Universidad de Salamanca
recio@usal.es

Maddalena Ghezzi
Universidad de Salamanca
maddy.ghezzi@usal.es

En esta comunicación se presentan los resultados del análisis de las unidades fraseológicas (UF) que aparecen en una muestra seleccionada de exámenes pertenecientes a los Diplomas de Español como Lengua Extranjera (DELE) del Instituto Cervantes de nivel superior, C1 y C2.

El mismo carácter definitorio de las UF (estructura sintagmática, idiomaticidad, fijación, etc.) y sus peculiaridades hacen que, tradicionalmente, la fraseología se haya considerado un aspecto complicado de enseñar en una clase de Español como Lengua Extranjera (ELE) y, por ello, se haya incluido principalmente en las programaciones didácticas de niveles avanzados y superiores.

Por esta razón, y con el fin de disponer de una muestra amplia y variada de UF, hemos decidido centrar nuestra atención en los exámenes oficiales de dominio del español de nivel C, que definen el "usuario competente" según el *Marco Común Europeo de Referencia* (MCER).

Así, nuestro corpus está constituido por las UF encontradas tras el análisis de cinco modelos del DSE (Diploma Superior de Español, en vigor hasta la convocatoria de agosto de 2011) y otros diez modelos de los actuales DELE de nivel C1 y C2 (en vigor desde la convocatoria de noviembre de 2011).

Tras una breve introducción sobre la didáctica y evaluación de las UF en los niveles superiores de ELE, explicaremos las características generales de los DELE, exámenes diseñados según las directrices internacionales del Consejo de Europa y del MCER.

La base de datos creada *ad hoc* para la organización del corpus del que disponemos nos permitirá analizar los fraseologismos hallados según distintos parámetros.

Por un lado, clasificaremos las UF dependiendo de su contexto de aparición (en ítems de evaluación de uso de la lengua, en textos de comprensión lectora o comprensión auditiva, etc.), dedicando especial atención a los casos en los que las mismas son objeto directo de pregunta de examen, con el fin de estudiar la tipología de ítem elegida. De este modo, evidenciaremos también las diferencias existentes entre el DSE y los nuevos DELE C1 y C2 en lo que se refiere a este aspecto.

En un segundo momento, procederemos a presentar la clasificación tipológica de las UF, siguiendo el modelo marcado por G. Corpas (1996), para determinar qué fraseologismos suelen ser los predominantes en estas pruebas de evaluación y, por lo tanto, cuáles son las UF que se suelen considerar más propias de los niveles superiores de ELE.

Concluiremos este breve estudio introductorio con las posibles vías de investigación futuras en este campo y una posible mejora de los exámenes de dominio en lo que se refiere a la fraseología.

References

BACHMAN, L. Y A. S. PALMER. (2010). *Language Assessment in practice*. Oxford: Oxford University Press.

ORDÓN, T. (2006). *La evaluación en el marco de E/L2*. Madrid: Arco/Libros.

BUITRAGO JIMÉNEZ, A. (1995). *Diccionario de dichos y frases hechas*. Madrid: Espasa Calpe.

CONSEJO DE EUROPA. (2001). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación.* [en línea] Disponible en: < http://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/> [Último acceso: 30/01/2015].

CONSEJO DE EUROPA. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A Manual*. [en línea]. Disponible en: http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf [Último acceso: 30/01/2015].

CORPAS PASTOR, G. (1996). *Manual de fraseología española*. Madrid: Gredos.

GARCÍA-PAGE, M. (2008). *Introducción a la fraseología española: estudio de las locuciones*. Barcelona: Anthropos.

INSTITUTO CERVANTES. (2007). *Plan Curricular del Instituto Cervantes. Niveles de referencia para el español.* Madrid: Edelsa, Biblioteca Nueva. 3 vols.

CERVANTES. (2015). *Página oficial de los DELE del Instituto Cervantes*. [en línea] Disponible en: <http://diplomas.cervantes.es/> [Último acceso: 30/01/2015].

LÓPEZ-MEZQUITA MOLINA, Mª. T. (2007). *La evaluación de la competencia léxica: tests de vocabulario, su fiabilidad y validez*. Madrid: MEC.

MARTÍNEZ BAZTÁN, A. (2011). *La evaluación de las lenguas: garantías y limitaciones*. Granada: Octaedro.

PRATI, S. (2007). *La evaluación en español lengua extranjera. Elaboración de exámenes*. Buenos Aires: Libros de la Araucaria.

PENADÉS MARTÍNEZ, I. (1999). *La enseñanza de las unidades fraseológicas*. Madrid: Arco/Libros.

RUIZ GURILLO, L. (1997). *Aspectos de fraseología teórica española*. Anejo XXIV de Cuadernos de Filología, Valencia: Universidad de Valencia.

RUIZ GURILLO, L. (2001). *Las locuciones en el español actual*. Madrid: Arco/Libros.

SÁNCHEZ LOBATO, J. E I. SANTOS GARGALLO (2004). *Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2)/lengua extranjera (LE).* Madrid: SGEL.

SECO, M., O. ANDRÉS Y G. RAMOS (2004). *Diccionario fraseológico documentado del español actual: locuciones y modismos españoles*. Madrid: Aguilar Lexicografía.

ZULUAGA OSPINA, A. (1980). *Introducción al estudio de las expresiones fijas*. Frankfurt: Peter D. Lang.

FRASEOLOGÍA E IDENTIDAD FEMENINA EN LOS MONÓLOGOS HUMORÍSTICOS DE EVA HACHE

Leonor Ruiz Gurillo

Universidad de Alicante

Leonor.Ruiz@ua.es

El objetivo de esta comunicación es mostrar el papel que desempeña la fraseología en la construcción del discurso femenino de la humorista Eva Hache. El corpus está compuesto por un total de 96 monólogos audiovisuales, extraídos de las últimas 3 ediciones del programa *El Club de la Comedia*, emitido por el canal LaSexta en 2011 y 2012, así como con los guiones publicados en ECC (2011). La humorista Eva Hache es la primera mujer que presenta el programa y, en ocasiones, la única mujer que sube al escenario esa noche. En sus monólogos desarrolla a menudo aspectos sobre la conceptualización del género, como diversas ideas preconcebidas sobre hombres y mujeres. Por otra parte, construye su *talante* y, en consecuencia, su autoridad cómica (Greenbaum, 1999) por medio de un discurso femenino que la identifica como una de las mejores conductoras que ha tenido el programa y que la convierte con bastante frecuencia en la mejor cómica de la noche. A menudo dramatiza monólogos donde se sitúa en el punto de vista de la mujer, por lo que su *feminolecto* (Litosseliti y Suderland, 2002) se enfrenta al discurso masculino.

En la construcción de esta identidad femenina interviene uno de los *indicadores* del humor más rentables, la fraseología. Así, emplea compuestos

sintagmáticos como *ropa interior*, colocaciones como *cometer un delito*, locuciones nominales como *plato de alta cocina,* locuciones verbales como *ponerse cañón,* locuciones clausales como *subírsele* [a alg.] *a la cabeza* o *hacérsele* [a alg.] *la boca agua*, o enunciados fraseológicos como *el tiempo todo lo cura.*

Observaremos cómo estas elecciones fraseológicas no constituyen procedimientos al azar, sino recursos metapragmáticos al servicio del humor que cumplen con los requisitos de variabilidad, negociabilidad y adaptabilidad (Verschueren, 2002). De este modo, muchas de las unidades fraseológicas empleadas cumplen un papel destacado en los *ganchos* o bromas humorísticas (Attardo, 2008). Además, algunos de estos indicadores fraseológicos se reinterpretan contextualmente tanto de forma fraseológica como literal. En consecuencia, la fraseología se convierte en un elemento fundamental en la construcción del discurso humorístico de Eva Hache.

References

ATTARDO, S. (2008). A primer for the linguistics of humor. In Raskin, V. (ed.). *The Primer of Humor Research*. Berlin: Mouton de Gruyter, pp.101-155.

ECC 2011= GLOBO MEDIA/SOGECABLE. 2011. *El Club de la Comedia*. (Presenta: *Qué mal repartido está el mundo... y el universo, ni te cuento*). Madrid: Aguilar.

GREENBAUM, A. (1999). *Stand-up comedy* as rhetorical argument: An investigation of comic culture. *Humor*, 12-1, pp. 33-46.

LITOSSELITI, L. AND SUDERLAND, J. eds. (2012). *Gender identity and discourse analysis.* Amsterdam: John Benjamins.

RUIZ GURILLO, L. (2012). *La lingüística del humor en español*. Madrid: Arco/Libros.

VERSCHUEREN, J. (2002). *Para comprender la pragmática*. Madrid, Gredos.

ANNOTATION OF MULTIWORD EXPRESSIONS IN FRENCH

Agnès Tutin

Université Grenoble-Alpes

agnes.tutin@u-grenoble3.fr

Emmanuelle Esperança-Rodier

Université Grenoble-Alpes

emmanuelle.esperanca-rodier@imag.fr

Manolo Iborra

Université Grenoble-Alpes

manolo.iborra@e.u-grenoble3.fr

Justine Reverdy

Université Grenoble-Alpes

justine.reverdy@e.u-grenoble3.fr

This study presents an experiment of multiword expression annotation on the French part of a French-English bilingual corpus. Our aim is to achieve:

a) a corpus-based and robust typology of MWEs;
b) a basis for linguistic studies on MWEs, especially in relation to diverse textual genres.
c) a corpus of evaluation for MT tasks, and especially SMT tasks.

To our knowledge, very few corpora with such annotations are currently available for French, except specialized annotation of MWE nouns and MWE adverbs (Laporte *et al.* 2008a; Laporte *et al.* 2008b), but these corpora do not include any typology of expressions. The French Treebank (Abeillé *et al.* 2003) includes several kinds of MWEs including verbs, but only on contiguous MWEs. In English too, there are not many resources. One of the most interesting ones is undoubtedly Schneider *et al.* 2014's social web corpus with MWE

annotations, which distinguishes between strong and weak MWEs, but does not include any fine-grained typology.

Building such a corpus is a challenging task for several reasons. First, deciding what belongs or not to the class of MWEs is not trivial. While function words such as *as long as*, obviously fall into the class of MWES, boundaries are less clear with collocations such as *to take a walk* or some nominal expressions like *address book*. Second, delimiting clearly the boundaries of MWEs is also a complicated task. For example, do determiners belong to the class of MWEs when they are highly variable? Third, labelling every expression with a MWE tag is far more complicated than presenting a typology with clear and prototypical examples.

Our first experiment is based on the annotation of two kinds of French texts: a scientific report of the BAF corpus and an extract of a literary text, *Thérèse Raquin* (22,000 words). Our annotation process is based on a semi-automatic annotation process using a finite-state transducer tool (NooJ, Silberztein 2008) and lexicons extracted from the French Treebank (Abeillé 2003) and the *Dictionnaire Électronique des Mots* (Dubois & Dubois-Charlier 2010). Our typology is inspired by several models (Heid 2008; Mel'čuk 2013; Tutin 2010) and includes a large set of MWEs: collocations (e.g. *pay attention*), full phrasemes (*to kick the bucket, dead end*), functional words (e.g. *as long as, in front of*), pragmatemes (e.g. *see you later*), complex terms (e.g. *multi-word expression*), routine formulae (e.g. *as expected*), proverbs, named entities (e.g. *Thérèse Raquin*).

An interannotator agreement experiment on an excerpt of 3,500 words of this corpus has been performed in order to identify the most difficult decisions. The results show a very good agreement on the scientific report (80,5 % of agreements between two annotators; Cohen's Kappa: 0,743) and good agreement on the literary extract (78,8 % of agreements; Cohen's Kappa: 0,683). The best agreements concern functional words and named entities, and the worst ones are related to uncertainty between collocations and full phrasemes, which will lead us to provide more precise criteria for these two categories.

References

ABEILLE, A., CLEMENT, L., & TOUSSENEL, F. (2003). Building a treebank for French. In *Treebanks.* Springer Netherlands. pp. 165-187.

DUBOIS, J., & DUBOIS-CHARLIER, F. (2010). La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration. *Langages*. 179(3). pp 31-56.

HEID, U. (2008). Computational phraseology. An overview. *Phraseology. An interdisciplinary perspective.* Amsterdam: Benjamins. pp 337-360.

LAPORTE, E., NAKAMURA, T., & VOYATZI, S. (2008a). A French corpus annotated for multiword nouns. In *Language Resources and Evaluation Conference. Workshop Towards a Shared Task on Multiword Expressions.* pp. 27-30.

LAPORTE, E., NAKAMURA, T., & VOYATZI, S. (2008b). A french corpus annotated for multiword expressions with adverbial function. In *Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop.* pp. 48-51.

MEL'CUK, I. (2013). Tout ce que nous voulions savoir sur les phrasèmes, mais... *Cahiers de lexicologie*. Revue internationale de lexicologie et de lexicographie*.* 102. pp. 129-149.

SCHNEIDER, N., ONUFFER, S., KAZOUR, N., DANCHIK, E., MORDOWANEC, M. T., CONRAD, H., & SMITH, N. A. (2014). Comprehensive annotation of multiword expressions in a social web corpus. *Proc. of LREC. Reykjavík, Iceland*.

SILBERZTEIN, M. (2008). Complex annotations with NooJ. In *Proceedings of the 2007 International NooJ Conference* Cambridge Scholars Publishing. pp-214.

TUTIN, A. (2010). Sens et combinatoire lexicale: de la langue au discours. *Unpublished Dossier en vue de l'habilitation à diriger des recherches). Grenoble: Université de Stendhal*.

RESTRICTED COLLOCABILITY AND ITS USE IN ARABIC CORPUS LINGUISTICS

Petr Zemánek
Charles University, Prague
petr.zemanek@ff.cuni.cz

Jiří Milička
Charles University, Prague
milicka@centrum.cz

It is generally accepted that in every language, some words have restricted collocability and some of them extremely (such as “pulmonary embolism”: for the whole expression, BNC reports 15 occurrences, whereas for “embolism” 21, i.e. 71% of the instances of the word "glottal" occurs in the phrase "glottal stop"). These restrictions are manifested in the ratios of the occurrences of individual items with the frequency of the whole string. The occurrence of such collocations can be considerably frequent, such as “look forward to”, where the intersection of the whole and “look forward” represents almost 92% (BNC search: 1095/1005), thus exhibiting a very strong relation between the two parts. Although the description of these multi-word units can be found in lexicographical textbooks (e.g. (Granger and Meunier 2008), the usage so far reported in literature is rather limited to purposes of studying lexicalization (e.g. (Barkema 1996 in connection with terminology), they are part of studies on language teaching (e.g. Howarth 1998) or are used in enriching existing databases with phrases (for WordNet, Bentivogli – Pianta 2003). They are, among other fields, an issue in translatology (e.g. (Hansen et al. 2001). All of these perceive these collocations as semantic units (semantic cohesion criterion). A pioneering contribution that slightly deviates from this mainstream is Čermák 2006, where the phenomenon is used for mapping prepositional

valency or Bentivogli – Pianta 2004 who suggest to enrich the lexical databases with syntactic information based on restricted collocations.

The method can be viewed as comparing sets of occurrences of individual words and their overlap (intersection). We would like to show that such phenomenon can be used for the study of many linguistic purposes or even go further across the border of linguistics, and sometimes it is fruitful to look aside from a purely lexicographic approach which prefers to find a special meaning of such collocations. We argue that the method brings interesting linguistic data that offer many possibilities of analysis, based both on a presupposed semantic cohesion criterium as well as without considering solely the semantic part.

Our dataset is based on a historical corpus of Arabic containing ca 440 million words, reflecting the use of the language in medieval times. Based on our data we will try to show some of the possible applications of our method that go beyond the attempts to establish a meaning of a collocation. These principles will be demonstrated on several case studies that will cover lexically oriented focus, usage in purely linguistic analyses and possible employment in culturally focused treatises. These case studies will include:

-study of prepositional valency (a follow-up on Čermák 2006);

-possible application in syntax (study of subsystems of clauses – conditional clauses, circumstantial qualifiers);

-lexical profiling of a given concept (certain expressions connected exclusively or predominantly with a given lexical item or their group, resulting in a prototypical characteristics of these concepts – e.g. the difference between angels and demons).

References

BARKEMA, H. (1996). 'Idiomaticity and Terminology: A Multi-Dimensional Model.' *Studia Linguistica* 50(2), 125-160.

BENTIVOGLI, L. AND PIANTA, E. (2003). 'Beyond lexical units: Enriching wordnets with phrasets.' *Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL03)*, Vol. 2, 67-70.

BENTIVOGLI, L. AND PIANTA, E. (2004). 'Extending wordnet with syntagmatic information.' In *Proceedings of Second Global WordNet Conference*, 47-53. Brno (Czech Republic).

ČERMÁK, F. (2006). 'Collocations, Collocability and Dictionary.' In: In E. Corino, C. Marello, and L. Onsti (eds.), *Proceedings of XII EURALEX International Congress,* Torino: Edizioni dell' Orso, Vol. II, 929–937.

GRANGER, S. AND MEUNIER, F. eds. (2008). *Phraseology: An interdisciplinary perspective.* Amsterdam: John Benjamins.

HANSEN, G. MALMKJÆR, K. AND GILE, D. eds. (2004). *Claims, Changes and Challenges in Translation Studies. Selected Contributions from the EST Congress.* Copenhagen 2001. Amsterdam/Philadelphia: John Benjamins.

HOWARTH, P. (1998). 'Phraseology and second language proficiency.' *Applied Linguistics*, 19, 24-44.

GERMAN-INTO-BASQUE/SPANISH TRANSLATION ANALYSIS OF BINOMIALS IN A PARALLEL AND MULTILINGUAL CORPUS

Zuriñe Sanz Villar
Universidad del País Vasco
zurine.sanz@ehu.es

As can be read in Piirainen (2012: 43) binomials are "sequences of two or more constituents that belong to the same grammatical category, have some semantic relationship and are joined by a conjunction like *and* or *or*". This type of PU has been analysed from several perspectives and across different languages (Malkiel 1959; Gustaffson 1984; Müller 2009; Čermák 2010; Toury 2012; Pontrandolfo 2011), and this is the first attempt to carry out a corpus-based translation analysis of binomials in the language combination German-Basque.

It is important to emphasise that a multilingual (and not a bilingual) corpus has been created. Since a considerable number of translations have been labeled as indirect — that is, as translations that have not been made directly from the German source text, in this case — at the Aleuska catalogue — a catalogue consisting of all literary works that have been translated from German into Basque over history —, intermediary texts have been added to the corpus in the case of indirect translations. By doing so, we were able to compare the data obtained from direct and indirect translations.

Regarding binomials, it is said they represent a widespread PU type across languages. As far as my corpus is concerned — the mentioned AleuskaPhraseo

parallel and multilingual corpus, which consist of literary texts translated from German (or other intermediary version(s)) into Basque — all in all, and in terms of tokens, 1,456 German binomials and 2,230 Basque binomials were extracted. The mentioned PU-lists of German and Basque binomials have been created with AntConc, a multiplatform freeware, which, among others, enables the user to automatically scan “the entire corpus for 'N' (e.g. 1 word, 2 words) length clusters” (Laurence 2014), in order to find common expressions in the uploaded corpus.

According to different criteria, a representative sample of German and Basque binomials has been selected. For the purpose of obtaining the counterparts of the extracted binomials, the AleuskaPhraseo parallel corpus, which is linked to a search-engine, has been consulted. Then, from a descriptive approach, the translation of the extracted binomials has been analysed.

The aim of the present paper will be to present the AleuskaPhraseo corpus, a “small-scale topic-specific PTC [Parallel Translational Corpus]” (Ji 2010: 6) consisting of around 3.5 million words from where the binomials under analysis have been extracted, to describe the actual process of binomials’ extraction using AntConc, and drawing on the theoretical and methodological framework proposed by Toury (2012), to show the results obtained from the translation analysis.

References

ČERMÁK, F. (2010). Binomials: Their nature in Czech and in general. In: J. Korhonen, W. Mieder, E. Piirainen, and R. Piñel, eds. 2010. *Phraseologie global-areal-regional.* Tübingen: Narr. p.309-315.

GUSTAFFSON, M. (1984). The syntactic features of binomial expressions in legal English. *Text*, 4, p.123–141.

JI, M. (2010). *Phraseology in Corpus-Based Translation Studies*. Frankfurt am Main: Peter Lang.

LAURENCE, A. (2014). AntConc. [online] Available at: <http://www.laurenceanthony.net/software/antconc/releases/AntConc343/help.pdf> [Accessed 27 January 2015].

MALKIEL, Y. (1959). Studies in irreversible binomials. *Lingua*, 8, p.113–160.

MÜLLER, H. G. (2009). *Adleraug und Luchsenohr: deutsche Zwillingsformeln und ihr Gebrauch*. Frankfurt am Main: Peter Lang.

PIIRAINEN, E. (2012). *Widespread idioms in Europe and beyond. Towards a Lexicon of Common Figurative Units*. New York: Peter Lang.

PONTRANDOLFO, G. (2011). *La fraseología en las sentencias penales: un estudio contrastivo español, italiano, inglés basado en corpus*. PhD thesis, Università degli studi di Trieste.
TOURY, G. (2012). *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

NLP and/or corpus-based identification and classification of phraseological units

Identificación y clasificación de unidades fraseológicas basada en corpus o mediante técnicas de PLN

TRADUIRE DES EXPRESSIONS FIGEES FRANÇAISES EN LANGUE ETRANGERE (ITALIEN, ALLEMAND, ESPAGNOL): TRAITEMENT COGNITIF, STRATEGIES D'INTERPRETATION ET ELABORATION

Mariangela Albano
Université de Bourgogne et Université
Sorbonne Nouvelle – Paris 3
albanomariangela@gmail.com

Dans le cadre de l'apprentissage des langues étrangères, le figement a retenu l'attention de plusieurs chercheurs. Puisque les allophones n'ont pas accès à la dimension figurée qui caractérise les expressions figées d'une langue, la maîtrise de celles-ci est très difficile et n'est atteinte, généralement, qu'à un stade avancé de l'acquisition. Cependant le domaine des expressions figées se prête à accueillir une quantité incalculable d'éléments, ce qui se produit au fur et à mesure que les locuteurs d'une langue s'efforcent d'obtenir un maximum d'économie en répétant des expressions qu'ils ont déjà entendues, au lieu d'en créer de nouvelles. En ce sens, la décodification d'un message est plus facile pour les locuteurs natifs qui activent un processus de mémorisation ayant des implications psycholinguistiques. Dans ce contexte, le but de notre étude est de fournir une contribution à l'analyse du traitement

sémantique qui se déroule au niveau cognitif dans l'interprétation des unités phraséologiques.

Notre communication porte sur le traitement d'expressions figées françaises par des apprenants italophones, germanophones et hispanophones adultes en contexte universitaire. Le corpus est représenté par un nombre suffisant d'expressions figées tirées de dictionnaires, pas directement traduisibles en italien/allemand/espagnol et données en contexte (brefs textes de presse). La singularité des expressions figées choisies est de mettre en place des mécanismes cognitifs différents car il s'agit des métaphores, métonymies, similitudes, proverbes et personnifications. On abordera ces processus à partir de deux perspectives d'analyse : l'une relevant de la linguistique acquisitionnelle, ayant pour but de s'interroger sur les démarches et sur les stratégies interprétatives mises en place par des apprenants étrangers lorsqu'ils sont confrontés à une expression idiomatique, et l'autre relevant de la linguistique cognitive et ayant pour but l'analyse des opérations cognitives qui créent des réseaux sémantiques de nature analogique pendant le traitement d'une expression figée. Notre attention sera focalisée sur les stratégies interprétatives mises en place pour accomplir la tâche. Les processus interprétatifs élaborés par les étudiants étrangers commencent par une approche analytique, qui leur permet tout d'abord de repérer l'idiome et de l'isoler à l'intérieur de la phrase. La non-compositionnalité des idiomes et leur nature métaphorique pousse les apprenants à se servir de stratégies différentes, comme l'ancrage au contexte et le recours aux analogies avec les idiomes de leur langue maternelle

References

BENSON, M. (1985). Collocations and idioms. In R. Ilson, éd., *Dictionaries, Lexicography and Language Learning* VIII. Oxford : Pergamon, pp. 61-68.

BRINTON, L. J. & TRAUGOTT E. C. (2005). *Lexicalization and Language Change*. Cambridge : CUP.

BURGER, H., éd. (2007). *Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung*. Berlin : Walter de Gruyter.

DOBROVOL'SKIJ, D. (2004). Idiome aus kognitiver Sicht. In K. Steyer, éd., *Wortverbindungen - mehr oder weniger fest*. Berlin : Walter de Gruyter, pp. 117-143.

GIBBS, R. W. JR. (1986). Skating on thin ice : Literal meaning and understanding idioms in conversation. *Discourse Processes, 7*, pp. 17-30.

GONZALEZ, R., éd. (2007). *Les expressions figées en didactique des langues étrangères*. Fernelmont : E.M.E.
GROSS, G. (1996). *Les expressions figées en français; noms composés et autres locutions*. Paris : Éditions Ophrys.
HUDSON, J. (1998). *Perspectives on fixedness: applied and theoretical.* Lund : Lund University Press.
KLEIBER, G. (1999). Les proverbes: des dénominations d'un type 'très très spécial'. *Langue française,* 123, pp. 52-69.
KÖVECSES, Z. & SZABO P. (1996). Idioms: a view from cognitive linguistics. *Applied Linguistics*, 17, pp. 326-355.
MARTIN, R. (1997). Sur les facteurs du figement lexical. In M. Martins-Baltar, éd., *La locution entre langue et usages*. Fontenay Saint Cloud : ENS Éditions, pp. 291-305.
MEL'CUK, I. (1993). La phraséologie et son rôle dans l'enseignement/apprentissage d'une langue étrangère. *Étude de Linguistique Appliquée,* 92, pp. 82-113.
NUNBERG, G.; SAG, I. A. & WASOW, T., éd. (1994). Idioms. *Language,* 70, 3. Washington DC: Linguistic Society of America, pp. 491-538.
SHAPIRA, CH. (1999). *Les stéréotypes en français : proverbes et autres formules*. Paris : Éditions Ophrys.
WOOD, D. (2006). Uses and Functions of Formulaic Sequences in Second Language Speech: An Exploration of the Foundations of Fluency. *The Canadian Modern Language Review*, 63, 1, pp. 13-33.
WRAY, A. (2002). *Formulaic Language and the Lexicon*. Cambridge : Cambridge University Press.

METAPHORICAL UNIVERSALS AND CULTURAL VARIATIONS IN BODY IDIOMS: TWO ROMANCE LANGUAGES IN CONTRAST

Marilei Amadeu Sabino

São Paulo State University,
Câmpus de São José do Rio Preto

amadeusm@ibilce.unesp.br

Ariane Lodi

São Paulo State University,
Câmpus de São José do Rio Preto

ariri_nena@hotmail.com

Because human body is common to all human beings, as well as the physical and mental actions we perform through it, and the feelings and emotions we experience, it is believed that body idioms may belong to the group of idiomatic expressions that mostly shares similar metaphors in different languages. Kövecses (2010) analyzed linguistic expressions related to the emotion "happiness" and found that several of them have similar metaphors in different cultures. Therefore, this author wonders how so different languages like English, Chinese and Hungarian, for example, can conceptualize happiness in a so similar way. To answer this question, Kövecses (op. cit.) raises three possible causes that could justify his findings: (1) this coincidence would have happened accidentally; (2) one or more languages would have borrowed their metaphors from another; or (3) there would be some universal motivations for the metaphors that emerge in these different cultures. Although Kövecses recognizes that none of these factors can be completely disregarded, he bets on the third alternative. Thus, this study will analyze some somatic idioms of

Italian in contrast with (Brazilian) Portuguese language, in order to examine closely some issues relating to metaphorical and metonymic similarities and differences underlying these different languages and cultures. To carry out this investigation, we will rely on Lakoff and Johnson's conceptual metaphor studies (1980) as well as on phraseologism researches carried out by Zuluaga (1980), Corpas Pastor (1996), and others. Kövecses' arguments (2000, 2005, 2010) seem to be logical and well supported by research. One of our methodological resources is the use of the world wide web as a corpus for phraseological research (Fletcher, 2004). Colson (2007, p. 1071) states that its use presents some advantages in identifying, extracting and describing phraseological units and "is a fairly recent development" for linguistic purposes. According to our data, we may conclude that a great amount of the studied idioms are metaphorically similar in both languages. However, there are visibly noticeable metaphorical variations in both too. In this paper we will discuss some of these similarities and variations that shape human cognition, based on these considerations.

References

COLSON, JEAN-PIERRE (2007). The World Wide Web as a Corpus for Set Phrases. In: Burger, H. (Dobrovol'skij, D. (Kuhn, P. (Norrick, N. ed. *Phraseologie / Phraseology, Handbooks of Linguistics and Communication Science.* Berlin, New York, Mouton de Gruyter. pp. 1071-1077.

CORPAS PASTOR, G. (1996). *Manual de fraseología Española. Madrid*. Editorial Gredos.

GIBBS, R. W. JR; STEEN, G. Eds. (1999). *Metaphors in cognitive Linguistics*. Amsterdam: John Benjamins.

GRADY, J. (1997). 'Theories are Buildings' revisited. *Cognitive Linguistics 8,* pp. 267-290.

IBARRETXE-ANTUÑANO, I. (2008). Vision Metaphors for the Intellect: Are They Really Cross-Linguistic? In: *Atlantis - Journal of the Spanish Association of Anglo-American Studies*. 30 (1), pp. 15–33.

JOHNSON, M. (1987). *The body in the mind: the bodily basis of meaning, imagination and reason*. Chicago: University of Chicago Press.

KOVECSES, Z. (2002). *Metaphor: a practical introduction*. Oxford: OUP.

KOVECSES, Z. (2005). *Metaphor in Culture: universality and variation*. Cambridge: CUP.

LAKOFF, G.; JOHNSON, M. (1980). *Metaphors we live by*. Chicago: The University of Chicago Press.

LAKOFF, G.; JOHNSON, M. (2002). *Metáforas da Vida Cotidiana*. (Coordenação da Tradução Mara Sophia Zanotto). Campinas, SP: Mercados de Letras.

ROBERTS, R. P. (1996). O tratamento das colocações e das expressões idiomáticas nos dicionários bilíngües. In THOIRON, Philipe; BEJOINT, Henri. *Les dictionaires bilíngües*. Louvain-La-Neuve: Duculot.

SABINO, M. A. (2010a). Provérbios e Expressões Idiomáticas: desfazendo confusões teóricas e práticas. In: XATARA, C. M.. *Estudos em Lexicologia e Lexicografia Contrastiva*. Curitiba: Honoris Causa. pp. 129-152.

SABINO, M. A. (2010b). Expressões Idiomáticas, Provérbios e Expressões Idiomáticas Proverbiais: iguais, semelhantes ou diferentes? In: ISQUERDO, A. N.; BARROS, L. A. (Ed. *O léxico em foco: múltiplos olhares*. São Paulo: Cultura Acadêmica/Fundação Editora da Unesp (FEU). pp. 331-347.

TAGNIN, S. O. (1989). *Expressões idiomáticas e convencionais*. São Paulo: Ática.

TONFONI, G; TURBINATI, L. (1995). Visualizzazione dei processi di traduzione: i proverbi e le espressioni idiomatiche. In: AA.VV. *La Traduzione: Saggi e Documenti II*. Roma: Divisione Editoria. pp. 239-252.

XATARA, C. M. (1998). A tradução para o português das expressões idiomáticas em francês. Tese. Araraquara: FCL da UNESP.

YU, N. (2008). The relationship between metaphor, body and culture. In FRANK, R.; DIRVE, R.; ZIEMKE, T.; BERNÁRDEZ, E. eds. *Cognitive Linguistics Research: Body, Language and Mind*. V. 2. Berlim: Walter de Gruyter. pp. 387-408.

ZULUAGA, A. (1980). *Introducción al estudio de las expressiones fijas*. Frankfurt: Peter D. Lang.

[Financial Support: FAPESP - São Paulo Research Foundation (process n° 2015/06484-5)].

OPINEXPRESS: UN LEXIQUE D'OPINION EN FORME DE GRAMMAIRES LOCALES

Oto Araujo Vale

Université Catholique de Louvain

Universidade Federal de São Carlos

oto.araujovale@uclouvain.be

Dans cette communication, nous présentons la construction d'un lexique d'opinions composé d'expressions multimots du portugais. Cette ressource est produite à partir de l'étude des occurrences d'expressions figées dans des textes réels. Ce travail s'inscrit dans le domaine de l'analyse de sentiments, un domaine de recherche nouveau qui consiste, approximativement, à déterminer de façon automatique l'opinion de l'auteur d'un texte à propos d'une entité ou d'un produit. La littérature du domaine établit qu'il est possible identifier l'opinion à partir de trois niveaux non-exclusifs : l'analyse au niveau du texte entier, l'analyse au niveau de la phrase ou encore l'analyse de l'entité. Cependant, pour que l'analyse soit bien établie dans chacun de ces trois niveaux, il est nécessaire d'utiliser une ressource fondamentale : un lexique d'opinion bien construit.

Notre hypothèse part de l'observation selon laquelle les expressions figées sont dans une large mesure porteuses de l'opinion de leur énonciateur. Par exemple dans une phrase comme:

Ana pagou os olhos da cara por essa mesa

(litt: Ana a payé les yeux de la tête pour cette table),

le locuteur exprime une opinion négative sur le prix trop élevé payé par le sujet de la phrase. En effet, il est intéressant de constater que la plupart des expressions relatives à l'argent dénotent l'argent perdu, beaucoup plus rarement l'argent gagné.

Ainsi, l'identification de ces expressions pourrait constituer un outil important pour la fouille d'opinion. Or, les travaux sur les lexiques d'opinion comme Wiebe et al (2005) ou encore Taboada et al (2011) concernent surtout des listes de mots simples comportant un petit nombre d'expressions multimots. Pour le portugais, des travaux comme Freitas (2013) ou Silva et al (2012) décrivent des expressions multimots, mais dans un nombre limité par rapport au reste du lexique présenté. Dans notre travail, nous proposons de construire ce lexique d'expressions du portugais au moyen de de grammaires locales. Notre approche rejoint les observations de Hunston et Sinclair (2003) sur l'application des grammaires locales proposées par Gross (1993 ; 1995 ; 1997) pour l'identification des sentiments. Les grammaires locales sont un outil très flexible qui peut être utilisé aussi bien pour la description des faits de langue comme les expressions multimots, que pour des tâches comme celles de la fouille d'opinion. Cette approche permet de décrire les différentes variations de chaque expression et a aussi pour avantage de permettre de construire les grammaires locales à partir des tables du lexique-grammaire. Cette caractéristique rend possible l'introduction de la description de certaines propriétés, comme la polarité de l'expression, le rôle joué par les actants et aussi les traces de l'énonciateur.

References

FREITAS, C. (2013). Sobre a construção de um léxico da afetividade para o processamento computacional do português. *Rev. bras. linguist. apl.* [online]. vol.13, n.4, pp. 1031-1059. <http://dx.doi.org/10.1590/S1984-63982013005000024> [Accessed 22 March 2015].

GROSS, M. (1993). Local Grammars and their Representation by Finite Automata. In: Hoey, M. ed 1993 *Data, Description, Discourse. Papers on the English Language in honour of John McH Sinclair*. Londres: Harper Collins, 1993, p. 26-38 <http://halshs.archives-ouvertes.fr/docs/00/27/83/08/PDF/sinc.pdf> [Accessed 08 December 2013].

GROSS, M. (1995). Une grammaire locale de l'expression des sentiments. Langue Française. 105 Paris: Larousse, 1995, p. 70-87. < http://persee.fr/web/revues/home/prescript/issue/lfr_0023-8368_1995_num_105_1> [Accessed 08 December 2013].

GROSS, M. (1997). The construction of Local Grammars. In: Roche, E; Schabes, Y. *Finite-States Language Processing.* Cambridge (EUA): MIT Press, pp. 329-354.

HUNSTON, S.; SINCLAIR, J. (2003) A Local Grammar or Evaluation. In: Hunston, S.; Thompson, G. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: OUP, pp. 74-101.

SILVA, M.; CARVALHO, P.; SARMENTO, L. (2012). Building a Sentiment Lexicon for Social Judgement Mining. In: Caseli et al eds. 2012 *Computational Processing of the Portuguese Language. 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012*. Berlin: Springer Berlin Heidelberg, pp. 218-228.

TABOADA, M.; BROOKE, J.; TOFILOSKI, M.;VOLL, K.; STEDE, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics* Vol. 37:2. <http://cgi.sfu.ca/~mtaboada/docs/Taboada_etal_SO-CAL.pdf> [Accessed 22 November 2013].

WIEBE, J.; WILSON, T.; CARDIE, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation,* 39(2-3), pp. 165–210. <http://people.cs.pitt.edu/~wiebe/pubs/papers/lre05.pdf> [Accessed 22 November 2013].

IMPLEMENTING EUROPEAN PORTUGUESE VERBAL IDIOMS IN A NATURAL LANGUAGE PROCESSING SYSTEM

Jorge Baptista **Francisco Dias** **Maria da Graça Fernandes**

Rui Talhadas **Nuno Mamede**

Verbal idioms (e.g. *kill two birds with one stone*) can be defined as frozen sentences where the verb and at least one of its arguments are frozen together and their overall meaning cannot be derived from the mere composition of the meanings of their individual elements, when used separately (undisclosed ref1). They constitute a large set of the lexicon-grammar of many languages, in the order of several thousands, though their frequency in texts is often very low. In fact, their occurrence it is highly dependent on text genre/type, being more common in oral than in written texts. The integration of verbal idioms in natural language processing (NLP) systems is relevant for an accurate semantic parsing. However, this integration is a challenge to NLP systems as these idioms cannot be dealt with like other idioms, as frozen strings of words. In spite of their being semantically non-compositional, they do have syntactic structure, allowing inflection, insertions, several transformations and creative pragmatic reuse ().

In this paper, we present an solution to the integration of verbal idioms in a fully-fledged NLP system (undisclosed reference). This is based on an extant

lexicon-grammar of European Portuguese verbal idioms, containing about 2,400 expressions, e.g. *deitar mãos à obra* (lit: throw hand to work) 'start working' (undisclosed ref., undisclosed ref.). Conceived in a tabular format, and organized in 10 main classes, according to the formal structure of the idioms (Gross, 1996), these tables present the verb and the frozen arguments of each idiom, along with the encoding of distributional constraints on free argument slots and the structural changes (or transformations) the sentence can undergo (passive, pronominalization, etc.). Each idiom is also illustrated by an example.

In order to integrate the lexicon-grammar of verbal idioms in the rule-based parsing module of the NLP system (undisclosed ref), the following strategy was adopted: firstly, the general parsing rules are applied, so the frozen sentence is given a structure like any ordinary sentence; then, another set of rules extracts a (semantic) dependency (called FIXED), based on the previous parse, and groups together the frozen elements of the idiom, while keeping intact the syntactic structure of the sentence; finally, the FIXED dependency is then used to further calculate the sentence's semantics: for example, the semantic roles of the verb's free argument are to be extracted not from the information attached to the simple verb (undisclosed ref.) but to those of the verbal idiom; the part-whole semantic relations extraction is blocked (undisclosed ref.); the verb sense statistical disambiguation module (undisclosed ref., undisclosed ref.) is prevented from acting; and so on. A script automatically reads the tabular format and converts it into the syntax of the rule-based parser for the extraction of the FIXED dependency. The assess the conversion process, the set of rules was applied to the examples provided in the lexicon-grammar. A small percentage of errors was detected and some rules were manually adjusted. Most errors, though, were due to incorrect POS tagging, from previous processing stages.

To evaluate the system's new module, sentences including all the key-elements of each idiom were extracted, and a representative sample, taken from a freely available corpus, was manually annotated by a team of linguists to build a golden standard. Then the sentences were parsed are results were automatically compared to the golden standard. A detailed error analysis is also briefly presented.

References

COWIE A. (1998). *Phraseology. Theory, analysis, and applications*. Clarendon press. Oxford.

GROSS, M. (1996). Lexicon-Grammar. *Concise Encyclopedia of Syntactic Theories*. Cambridge. Pergamon. pp. 244-258.

SAG, I.A., BALDWIN, T., BOND, F., COPESTAKE, A., AND FLICKINGER, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.) *Proceedings of the Third International Conference, CICLing - Computational Linguistics and Intelligent Text Processing*, Mexico City, Mexico, February, 2002, pp. 1-15.

POS-PATTERNS OR SYNTAX? COMPARING METHODS FOR EXTRACTING WORD COMBINATIONS

Sara Castagnoli
University of Bologna
s.castagnoli@unibo.it

Gianluca E. Lebani
University of Pisa
gianluca.lebani@for.unipi.it

Alessandro Lenci
University of Pisa
alessandro.lenci@ling.unipi.it

Francesca Masini
University of Bologna
rancesca.masini@unibo.it

Malvina Nissim
University of Groningen
m.nissim@rug.nl

Lucia C. Passaro
University of Pisa
lucia.passaro@for.unipi.it

It is widely acknowledged that lexicographers' introspection alone cannot provide comprehensive information about word meaning and usage, and that investigation of language in use is fundamental for any reliable lexicographic work (Atkins and Rundell 2008). This is even truer for dictionaries that record the combinatorial behaviour of words, where the lexicographic task is to detect the typical combinations a word participates in. In fact, it was much harder to study lexical combinatorics empirically before the advent of large corpora and the definition of statistical techniques for the analysis of word associations (Hanks 2012).

This paper reports on work carried out in the framework of an ongoing project aimed at building a lexicographic resource for Italian Word Combinations. We use the term Word Combinations (WoCs) to encompass both Multiword Expressions (MWEs) – namely a variety of WoCs characterised by different degrees of fixedness and idiomaticity that act as a single unit at some level of linguistic analysis (Calzolari et al. 2002, Sag et al. 2002, Gries 2008) – and the

distributional properties of a word at a more abstract level (argument structure, subcategorization frames, selectional preferences), along the lines of Benson et al. (2010).

Currently, apart from purely statistical approaches, the most common methods for the extraction of WoCs involve searching a corpus via sets of shallow morphosyntactic patterns and then ranking the extracted candidates according to various association measures (Villavicencio et al. 2007, Ramisch et al. 2010). Whereas most studies have so far focused on the use of POS-patterns, which yields satisfactory results for relatively fixed, short and adjacent WoCs, others have suggested that syntactic dependencies might also be exploited, especially to capture discontinuous and syntactically flexible WoCs (Seretan 2011).

We aim to test and compare the performance of the two methods with respect to the task of extracting Italian WoCs for inclusion in a combinatory resource. To this purpose we select a sample of Italian target lemmas (TLs) – including verbs, nouns and adjectives – by combining frequency information derived from the *la Repubblica* corpus (Baroni et al. 2004) and inclusion in the largest existing Italian combinatory dictionary (DiCI, Lo Cascio 2013), which we use for evaluation. For each TL, we extract from *la Repubblica*:

- all its occurrences in a set of pre-defined POS sequences deemed representative of Italian Word Combinations, using the EXTra tool (Passaro and Lenci, submitted);
- all its occurrences in syntactic frames together with the lexical fillers of the relevant syntactic slots, using the LexIt tool (Lenci et al. 2012).

For evaluation, we calculate the recall of the two methods using as benchmark the list of combinations recorded in DiCI, which is manually compiled. This sheds light on the independent performance of the two methods overall, and with respect to the extraction of different types of WoCs. In addition, manual inspection of the top candidates in both datasets is used to assess the proportion of *valid* WoCs that are extracted from the corpus but unattested in DiCI. This also provides information towards improving dictionary coverage.

References

ATKINS, B.T.S. AND RUNDELL, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

BARONI, M., BERNARDINI, S., COMASTRI, F., PICCIONI, L., VOLPI, A., ASTON, G. AND MAZZOLENI, M. (2004). Introducing the "La Repubblica" Corpus: A Large, Annotated, TEI(XML)- Compliant Corpus of Newspaper Italian. *Proceedings of LREC 2004*, pp.1771–1774.

BENSON, M., BENSON, E. AND ILSON, R. (2010). *The BBI Combinatory Dictionary of English*. 3rd revised edition. Amsterdam/Philadelphia: John Benjamins.

CALZOLARI, N., FILLMORE, C.J., GRISHMAN, R., IDE, N., LENCI, A., MACLEOD, C. AND ZAMPOLLI, A. (2002). Towards best practice for multiword expressions in computational lexicons. *Proceedings of LREC 2002*, pp.1934–1940.

GRIES, S. TH. (2008). Phraseology and linguistic theory: a brief survey. In: S. Granger and F. Meunier, eds. *Phraseology: an interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins. pp. 3–25.

HANKS, P. (2012). Corpus evidence and Electronic Lexicography. In: S. Granger and M. Paquot eds. *Electronic Lexicography*. Oxford: Oxford University Press. pp.57-82.

LENCI, A., LAPESA, G. AND BONANSINGA, G. (2012). LexIt: A Computational Resource on Italian Argument Structure. *Proceedings of LREC 2012*, pp.3712–3718.

LO CASCIO, V. ed. (2013). *Dizionario combinatorio italiano*. Amsterdam/Philadelphia: John Benjamins.

PASSARO, L.C. AND LENCI, A. (2015). Extracting Terms with EXTra. Submitted.

RAMISCH, C. VILLAVICENCIO, A. AND BOITET, C. (2010). mwetoolkit: a framework for multiword expression identification. *Proceedings of LREC 2010*, pp.662–669.

SAG, I.A., BALDWIN, T., BOND, F., COPESTAKE, A. AND FLICKINGER, D. (2002). Multiword expressions: A pain in the neck for NLP. *Proceedings of CICLing 2002*, pp.1–15.

SERETAN, V. (2011). *Syntax-based collocation extraction*. Dordrecht: Springer.

VILLAVICENCIO, A., KORDONI, V., ZHANG, Y., IDIART, M. AND RAMISCH, C. (2007). Validation and evaluation of automatically acquired multiword expressions for grammar engineering. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp.1034–1043.

SEMANTIC STRUCTURING OF VERBAL IDIOMS FROM THE CONCEPTUAL DOMAIN “DEATH”

Svitlana Chornoba
Crimean Federal University
svetoniy@gmail.com

Jorge Baptista
University of Algarve
jbaptist@ualg.pt

The paper deals with comparative scrutiny into the semantics of Portuguese, English and Russian verbal idioms constituting the conceptual domain {DEATH}. Languages comparison gives the opportunity to demonstrate “which various ways a human created a language and which part of the world of thoughts he managed to transfer into it” (Humboldt, 1985). The concept {DEATH} is a basic concept, at the core of conceptual sphere of any linguistic culture (Karakevich, 2009; Grabarova, 2005; Kytayhorods’ka, 2008; Hnapovs’ka, 2008). The paper is set in the context of an on-going research program to systematically describe – and compare – the verbal idioms of Russian and Portuguese (undisclosed reference).

A verbal idiom can be defined (M. Gross, 1996) as a syntactic-semantic lexical-grammatical unit, composed of a verb and at least one main constituent (usually a complement) that is distributionally frozen with the verb; the overall meaning of the idiom cannot be calculated by composition of the meaning of the individual elements of the expressions, as they are interpreted when used separately in other contexts. Hence, for example, in the idiom, *отбросить коньки* (‘throw the skates’; cp. English idioms *pop one’s clogs, kick the bucket*) the meaning is not derived from the meaning of its lexical constituents.

The goals of this paper are twofold: 1) to propose a strategy to the semantic structuring of verbal idioms from the conceptual domain {DEATH},

based on connotative processes (Lakoff and Johnson, 1980) involved in the idioms decoding mechanisms; 2) to assess the usefulness of this strategy by systematically comparing data gathered from three distantly related languages, Russian, English, and Portuguese. By this time, the collected data from this conceptual domain consists of 52, 76, 47 verbal idioms, respectively.

The main idea supporting this paper proposes that it is possible to structure a conceptual domain of idioms, in this case {DEATH}, and to obtain coherent subsets, by relying on the tropological processes involved in the idiomatic interpretation of the expressions. This approach has several advantages over a simple, flat, classification that would only rely on the generic concept, or even one that would distinguish grammatical features associated with the main concept, such as {CAUSE} (e.g. *send smb. into the bottomless pit*), {AGENT-VOLITION} (*Smb. put the horse, etc. to sleep*), or even aspectual features {UNACCOMPLISHED}, as in ***глядеть в гроб*** ('look into the coffin').

The adoption of this strategy has to do with dual readings of many idioms: while their meaning is non-compositional, their insertion in discourse still holds a link with the non-idiomatic, literal meaning of the components, which must be accounted for in many situations.

For the structuring of the idioms associated with the conceptual domain of {DEATH}, the following subsets were found: {DIE–GO}, {DIE–SLEEP}, {DIE–STRETCH–LEGS}, {DIE–SURRENDER–SPIRIT}, {DIE–BREATHE–AWAY}, {DIE-SACRIFICE} and some isolated metaphors.

Semantic classification of verbal idioms can be approached from different perspectives. The first step is to try and build general, semantically course classes, such as the one here discussed, {DEATH}. We posit that the metaphoric processes underlying the idiomaticity these expressions convey can be a useful strategy to semantically organize a conceptual domain, especially in a contrastive setting. Detailed tropological analysis is required, but this paper shows that it can be done. Extension to other semantic constructs, such as {BIRTH}, {MARRIAGE}, {DIVORCE}, *etc*., should provide a better perspective of this approach – aspects that will be tackled in future work.

References

GRABAROVA, E.V. (2005). Koncept “savoir vivre” – “umenie zhit” (Concept “savoir vivre” – “ability to live”). In: Karasik, I.V., Prokhvacheva, O.G., Grabarova, E.V. and Zubkova, L.V., 2005. *Inaya mental’nost’* (*Other Mentality*). Moscow: Gnozis. pp.257-333.

GROSS, M. (1996). Lexicon-Grammar, in: Brown, K. and Miller, J. (eds.). *Concise Encyclopedia of Syntactic Categories*. Cambridge: Pergamon. pp. 244–259.

HNAPOV’KA, L. (2008). Lingvokognityvni oznaky anglomovnoi metaforychnoi reprezentacii konceptu “smert’” (Linguocognitive signs of English-speaking metaphorical representation of the concept “death”). *Naukovi zapysky* (*Scientific notes*) 75 (4), p.55.

HUMBOLT, W.Von (1985). *Yazyk i filosofia kultury*. (*Language and Philosophy of Culture*) Translated from German by I.I.Levina *et al*. Moscow: Progress.

KARAKEVYCH, R.O. (2009). Konceptopolya “LEBEN”/ “ZHYTTYA” v nimeckij ta ukrains’kij movnyh kartynah svitu: asymetria frazeologychnyh lingvokulturem. *Materialy X Mizhnarodnoi naukovo-praktychnoi konferencii “Semantyka movy i tekstu”* (Conceptual fields “LEBEN”/ “LIFE” in German and Ukrainian language pictures of the world: asymmetry of phraseological culturemes. *Materials of X International scientific practical conference “Semantics of Language and Text”*, 21-23 September 2009 in Ivano-Frankivs’k,). p.131.

KYTAYHORODS’KA, K. (2008). Koncepty “zhyttya” i “smert’” u frazeologychnyh kartynah svitu anglijs’koi ta ukrains’koi mov. *Naukovi zapysky 75 (4),* p.205. (Concepts “life” – “death” in phraseological pictures of the world of the English and Ukrainian languages. *Scientific notes 75 (4)*, p. 205).

LAKOF, G., JOHNSON, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

LEXICON-GRAMMAR OF RUSSIAN VERBAL IDIOMS

Tetyana Fukova
Universidade do Algarve
tatyanafukova@gmail.com

Svitlana Chornobay
Crimean Federal University
svetoniy@gmail.com

Jorge Baptista
Universidade do Algarve
jorge.manuel.baptista@gmail.com

This paper describes an on-going project to build a lexicon-grammar of Russian verbal idioms for natural language processing. The aim is to produce a language resource that can be used to automatically identify these idioms in naturally occurring texts. Such resource can also be useful to several fields of research, such as language acquisition, or language learning and teaching (Cowie, 1998), among others.

Though many definitions, and also many conceptual and terminological disputes, (*idioms*, *collocations*, *phrasemes*, etc.) can be found in the literature around the term *idiom*, for this paper we adopt the definition of *verbal idioms* given by author1, who called it *frozen sentences*): “Frozen sentences are elementary sentences where the main verb and at least one of its argument are distributionally constraint, and usually the global meaning of the expression cannot be calculated from the individual meaning of its component elements when they are used independently.” For example, in Russian, an idiom like держать язык за зубами (deržat’ jazyk za zubami), literally ‘hold one’s tongue behind one’s teeth’, ‘keep one's tongue between one's teeth’, has nothing to do with the relative position of the tongue and the teeth, but rather ‘to be prudent when speaking, not saying things that should not be said’. The body part nouns are frozen with the verb and the locative preposition. Furthermore, the nouns can establish a part-whole relation with a free human noun (in the form of a

possessive), as the translation illustrates, and this have referential value, so that there is some structure and structural variation that must be adequately captured, and not just take the sequence as a whole. The idiom, however, can be parsed like an ordinary (semantically compositional) sentence. Ambiguous idioms like this need to be specifically marked, so that the adequate reading is found from context, if possible.

The automatic identification of the meaning units in texts involves the correct delimitation and tagging of idioms in texts. Using available linguistic resources, such as phraseological dictionaries (Molotkov, 1986; Fedosov and Lapisky, 2003) and the linguistic development platform Unitex (Paumier 2003, 2014)[1], along with the machine readable dictionary distributed with this software, we intend to determine the relevant linguistic information required to process this type of expressions, and to formalize it into a database of idioms.

To this date, we collected over 1,000 Russian verbal idioms from phraseological dictionaries and other sources, classified them using the lexicon-grammar framework (Gross, 1996) of and formalized them into a tabular format for computational processing and automatic identification in texts. This consist of a fine-grained description of the syntactic structure of those idioms, the lexical content of frozen elements, the distributional constraints on their free syntactic slots and their transformational properties, that is, the alternative, paraphrastically equivalent, forms (or alternations) they can yield. For each idiom, a word-by-word English translation and the relevant morphosyntactic information is provided, along with a free translation (gloss) or the English equivalent and an illustrative example. Using finite-state tools provided by the Unitex platform, some experiments on corpora are presented and preliminary results are reported.

References

COWIE A. (1998). *Phraseology. Theory, analysis, and applications*. Clarendon press. Oxford.

GROSS, M. (1996). Lexicon-Grammar. *Concise Encyclopedia of Syntactic Theories*. Cambridge. Pergamon. pp.244-258.

[1] http://www-igm.univ-mlv.fr/~unitex/

FEDOSOV, I. and LAPITSKY, A. (2003). *Phraseological dictionary of the Russian language*. (Федосов, И. и Лапицкий, А. *Фразеологический словарь русского языка*). Moscow: Unves.

MOLOTKOV, A. (1986). *Phraseological dictionary of the Russian language* (Молотков, А. *Фразеологический словарь русского языка*). Moscow: АСТ.

PAUMIER, S. (2003). De la reconnaissance des formes linguistiques à l'analyse syntaxique. PhD thesis, Université de Marne-la-Vallée, 2003.

PAUMIER, S. (2014). *Unitex 3.0 - User's Manual*. Paris: Université Paris-Est Marne-la-Vallée.

COLLOCATION DICTIONARY FOR SLOVENE: CHALLENGE FOR AUTOMATIC EXTRACTION OF DATA AND CROWDSOURCING

Polona Gantar

University of Ljubljana

apolonija.gantar@ff.uni-lj.si

Simon Krek

Jožef Stefan Institute, Center for language resources

simon.krek@guest.arnes.si

Iztok Kosem

Trojina, Institut for Applied Slovene Studies

iztok.kosem@trojina.si

Vojko Gorjanc

University of Ljubljana

vojko.gorjanc@ff.uni-lj.si

The contribution describes the compilation of a collocation dictionary for Slovene which is based on the data already registered in the Slovene Lexical Database (LBS) (Gantar and Krek 2011), and explores the procedures of automatic extraction of collocations and matching corpus examples from the reference corpus of Slovene.

The rationale behind the project is the non-existence of any collocation dictionary for Slovene and the lack of collocation data available for Natural Language Processing (NLP) tasks, as collocations are only sparsely available in general dictionary examples, they are not available in machine-readable formats and due to obsolescence do not reflect modern Slovene. The compilation of the collocation dictionary is devised as a pilot project with several goals: collocation data will be later integrated in the online dictionary of modern

Slovene, and it serves as a testbed for the introduction of automatic extraction procedures (from corpora) and crowdsourcing into lexicographic workflow.

The process of automatic extraction of collocation data from the corpus was tested during the compilation of Slovene Lexical Database (Kosem et al. 2013). However, subsequent work showed that the use of the Sketch Engine (Word Sketch and GDEX modules: Krek 2012; Kosem et al. 2012) could be upgraded by applying more rigorous extraction criteria and by automatic post-processing of the extracted data, with the aim to prepare better quality data for further lexicographic treatment. In addition, in the pilot project we wanted to test the intuition of dictionary users in relation to the optimal sense division in polysemous headwords. For this purpose a crowdsourcing platform was prepared for a semantic annotation task assigning extracted sentences with headwords and collocates to pre-prepared sense menus or signposts (Lew and Ptasznik 2014).

The contribution will describe the process of assigning automatically extracted collocation data (together with corpus examples) to individual senses in polysemous entries from the existing Slovene Lexical Database. The process involves the use of crowdsourcing for several tasks. First, we want to establish if users recognise the adequacy of sense descriptions in the existing sense menus, using newly extracted sentences which contain the headword and the collocate. This process should show the consistency and comprehensibility of existing sense menus. The resulting crowdsourced data will be evaluated and used for finalising sense division in polysemous entries. We expect that this process will also show to what extent users recognise expressions which are strucutrally more solid and semantically less transparent than collocations, such as fixed expressions and (non-compositional) phraseological units.

References

GANTAR, P., KREK, S. (2011). Slovene lexical database. In: D. Majchraková, R. Garabík, ed. 2011. Proceedings of the Sixth International Conference. Modra, Slovakia, 20-21 October 2011, *Natural language processing, multilinguality*. Slovenská akadémia vied, Jazykovedný ústav L'udovíta Štúra: Tribun EU. pp.72-80.

KOSEM, I., GANTAR, P., KREK, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowdsourcing. In: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik ed. 2013. Proceedings of the eLex 2013 conference, 17–19 October 2013,

Electronic lexicography in the 21st century: thinking outside the paper. Tallinn, Estonia. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut. pp. 32-48.

KOSEM, I. HUSÁK, M., MCCARTHY, D. (2011). GDEX for Slovene. In: I. Kosem, K. Kosem ed. 2011. Proceedings of eLex conference, Bled, 10-12 November 2011, *Electronic Lexicography in the 21st Century: New applications for new users.* Ljubljana: Trojina, zavod za uporabno slovenistiko.

KREK, S. (2012). New Slovene sketch grammar for automatic extraction of lexical data. *SKEW3*, Brno, 21.-22. march 2012. [online] Available at: <https://trac.sketchengine.co.uk/attachment/wiki/SKEW-3/Program/Krek_SKEW-3.pdf?format=raw>

PTASZNIK, B., LEW, R. (2014). Do menus provide added value to signposts in print monolingual dictionary entries? An application of Linear Mixed-Effects Modelling in dictionary user research. *International Journal of Lexicography* , 27(3), p. 241-258.

SELECTING COLLOCATIONS AMONG MULTIWORD UNITS FROM A SPECIALIZED CORPUS

Larisa Grčić Simeunovic
University of Zadar
lgrcic@unizd.hr

Paula de Santiago Gonzales
Freelance translator and interpreter
desantiagopaula@gmail.com

Corpus research on multiword units has led to the recognition of the importance of phraseology for specialized language description. Corpus linguistics is actually the most frequently used method for the study of phraseology. It allows us to find out not only quantitative data but to analyze the grammar and lexis/semantics interface.

Among the influential works based on the identification and analysis of collocations, this study find s great support on Hunston and Francis' Pattern grammar (2000) and Martin's semantic frames (2008). Both consider the phraseological tendency of language as it motivates sense and syntax to be associated. Hunston and Francis (2000) suggest that frequent combinations of words can reveal patterns of words that display different senses. Martin (2008) establishes a frame-based approach to predict the combinatory potential of words. He understands collocations as a combination of two concepts or frames which are in dependency relation.

A crucial issue in specialized phraseology research is to know how to recognize multiword units as phraseologisms in contrast to complex terms. We believe that a clear notion of the object of study should help. According to Gries' (2008) defining criteria of phraseology, our study deals with lexical collocations made of two to three lexical items, that is, an adjectival collocate premodifying a

single or complex nominal term (Adj + N/Adj + NP). The selected syntagmatic combinations have a minimum frequency of 3 occurrences in the corpus being found in at least 2 different texts. Regarding the permissible distance between constituents, we will adopt an open perspective considering continuous and discontinuous collocations. Thus, flexible patterns will be taken into account allowing lexical units maintain a syntactic relationship in which the function of collocates may change from attributive to predicative position. Finally, it should be noted that the overall meaning of collocations in our study results from the sum of meanings of the constituents.

On the basis of an English specialized corpus in the field of karstology, collocations will be distinguished from complex terms on the basis of a morphological and a semantic analysis of collocates as well as an analysis of the syntactic relation between the constituents of the combinatory structure.

Morphologically, modifiers of base nouns will be classified (relational adjectives, deverbal adjectives, adjectival nouns, and qualifying adjectives) so as to prove which type tend to trigger collocations. Semantically, modifiers will be grouped in semantic frames (genesis, constitutive material, size, shape, etc.) to discover which of them tend to promote collocations. Syntactically, the relation of the constituents will be analyzed on the basis of fixed or non-fixed order within examples of use extracted from the corpus.

Secondly, once the previous analyses provide us with a list of collocations, a further analysis based on the frame approach will be performed. Frames will not only help in grouping words in knowledge sets, they will also clarify the combinatorial behavior. Therefore generalizations on the formation of collocations in karstology will be drawn according to the combinatorial possibilities of lexical units. Our phraseological model aims to illustrate terms in use and more specifically the different lexicosyntactic patterns in which each term in this specialized field regulary participates.

References

De SANTIAGO, P. and GRČIČ, L. (2015). The polymorphic behaviour of adjectives in terminology. *Meta*. In print

GRIES, S. (2008). Phraseology and linguistic theory. In: S. Granger and F. Meunier, eds. *Phraseology. An Interdisciplinary Perspective*. Amsterdam/Philadelphia: John Benjamins. pp 3-21.

KAEGURA, K. and UMINO, B. (1996). Methods of automatic term recognition: a

review. *Terminology,* 3(2), p. 259-289.
MARTIN, W. (2008). A unified approach to semantic frames and collocational patterns. In: S. Granger and F. Meunier, eds. 2008. *Phraseology. An Interdisciplinary Perspective*. Amsterdam/Philadelphia: John Benjamins. pp 51-65.
OSTER, U. (2006). Classifying domain-specific intraterm relations: A schema-based approach. *Terminology*, 12(1), p. 1-17.
PAVEL, S. (1994). *Archived guide on phraseology research in languages for special purposes* [online]. *Available at:* <http://www.bt-tb.tpsgc-pwgsc.gc.ca/btb-pavel.php?page=guidephras&lang=eng&contlang=eng> [Accessed 5 March 2015]
ROBERTS, R. (1998). Phraseology and translation. In: P. Fernández and J.M. Bravo, eds. 1998. *La traducción: orientaciones lingüísticas y culturales*. Valladolid: Universidad de Valladolid. pp 61-77.
RÖMER, U. (2009). The inseparability of lexis and grammar. *Annual Review of Cognitive Linguistics*, 7, p. 141-163.
SINCLAIR, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
SINCLAIR, J. (2004). *Trust the text. Language, corpus and discourse*. London: Routledge.
STANISLAW, G.R. (2012). Discovering patterns and meanings: corpus perspectives on phraseology in legal discourse. *Roczniki Humanistycne*, LX(8), p. 47-70.
ZELINSKY-WIBBELT, C. (2012). Identifying term candidates through adjective-noun construction in English. *Terminology,* 18(2), p. 226-242.

AUTOMATIC ACQUISITION OF MULTI-WORD TERMS IN FRENCH

Xiaoqin Hu

Université Paris 13

franhuxqin@hotmail.fr

Pierre-André Buvet

Université Paris 13

pierreandre.buvet@gmail.com

Multi-word terms extraction is always a challenge for natural language processing because of their flexible structures and variable forms. Nowadays, many of the methods for multi-word terms extraction are based on either the linguistic descriptions (such as the TERMINO solution (Lauriston, 1994), the LEXTER system (Bourigault, 1996, etc.) or the statistical calculation, such as n-gram classification method used in the system ANA (Enguehard, 1993), statistical parsing of Spence Green (2013), etc. There are also many other methods which combine the linguistic models and statistical models, such as the semi-automatic system ACABIT (Daille, 1994) in which the multi-word term candidates are extracted with the morphosyntactic patterns and are filtered by a statistical calculation. This article describes a hybrid method for automatic acquisition of multi-word terms in French. In this method, a morphosyntactic analysis of the multi-word terms is done for identifying the multi-word term candidates and the term candidates are filtered by integrating a semantic distributional method which is based on the appropriate relation between the appropriate predicates and their class of arguments. We've studied two specific vocabularies in French: artifact nouns and occupational nouns for the method experimentation. For artifact nouns, a set of morphosyntactic patterns is established on the basis of inner structure analysis about the multi-word artifact nouns. The semantic relation between the constituents of the multi-word nouns

is also taken into account. The morphosyntactic patterns are divided into two groups: simple patterns and complex patterns which are formed from the simple patterns. Next, a syntactic-semantic analysis about the appropriate predicates of artifact nouns is carried out for locating the predicate-argument structures associated with artifact nouns and a set of appropriate predicate candidates are obtained. An intersection technic is used for filtering the appropriate predicate candidates and a probabilistic calculation is integrated for a selection of syntactic patterns for the appropriate predicates. With the selected appropriate predicates and their syntactic patterns, the nominal distribution of artifact nouns can be located and a semantic filtering can be executed to the multi-word artifact nouns candidates. The same method is applied to the occupational nouns, but some peculiarities about the occupational nouns' morphosyntactic features and their appropriate predicates are considered. In comparison to other morphosyntactic description-based methods, the proposed method takes into account the natural language transformation for establishing the morphosyntactic patterns and appeals to the appropriate relation between the appropriate predicates and their semantic class of arguments for a semantic filtering to the multi-word term candidates obtained by morphosyntactic patterns.

References

ALECU, B. P. (THOMAS, Izabella, RENAHY, Julie. (2012). La "multi-extraction" comme stratégie d'acquisition optimisée de ressources terminologiques et non terminologiques. In : *Actes de la conférence conjointe JEP-TALN-RECITAL, Grenoble, 2012*. Vol(2) : TALN, pp.511-518.

BISKRI, Ismaïl, MEUNIER, Jean-Guy AND JOYAL, Sylvain. (2004). L'extraction des termes complexes : une approche modulaire semi-automatique. *7es Journées internationales d'Analyse statistique des Données Textuelle*, *Louvain-la-Neuve (Belgique), 2004*. pp.192-201.

BOURIGAULT, D. (1996). LEXTER: a natural language tool for terminology extraction. *Proceedings of the 7th EURALEX International Congress*, *Göteborg, 1996*. pp.771-779.

BUVET, P.-A. (2009). Des mots aux emplois : la représentation lexicographique des prédicats. *Le Français Moderne,* 77(1), pp.83-96.

CONSTANT, Matthieu, SIGOGNE, Anthony AND WATRIN Patrick (2012). La reconnaissance des mots composés à l'épreuve de l'analyse syntaxique et vice-versa : évaluation de deux stratégies discriminantes. In : *Actes de la conférence conjointe JEP-TALN-RECITAL, Grenoble, 2012*. Vol(2) : TALN, pp.57-70.

DAILLE, B. (1994). Study and implementation of combined techniques for automatic extraction of terminology. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language, Proceedings of the "Workshop of the 32nd Annual Meeting of the ACL", Las Cruces, New Mexico, USA, 1994*. pp.29-36.

ENGUEHARD, Chantal (1993). Acquisition de Terminologie à partir de gros corpus. *Proceedings of Informatique & Langue Naturelle (ILN'93), Nantes*. pp.373-384.

GREEN, Spence, DE MARNEFFE, Marie-Catherine AND MANNING, Christopher D. (2013). Parsing Models for Identifying Multiword Expressions. *Computational Linguistics*, Vol(39), pp.195-227.

GROSS, M. (1986). *Grammaire transformationnelle du français : Syntaxe du verbe ; Syntaxe du nom.* Malakoff : Cantilène.

HARRIS, Z. S. (1976). *Notes du cours de syntaxe*. Paris : Le seuil.

HARRIS, Z. S. (1979). *Mathematical structures of language*. New York : R. E. Krieger.

HARRIS, Z. S. (1988). *Language and information*. New York : Columbia University Press.

HEARST, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of the 14th conference on Computational linguistics, Stroudsburg, PA, USA, 1992*. pp.539-545.

JIANG Jing (2012). Information Extraction from Text. In: Charu C. Aggarwal et al. (ed.) 2012. *Mining Text Data*. US: Springer. pp. 11-41.

IBEKWE-SANJUAN, Fidelia. (2007). *Fouille de textes*. Paris: Lavoisier.

LAURISTON, A. (1994. Automatic recognition of complex terms: problems and the TERMINO solution. *Terminology*, 1(1), pp.147-70.

MAUREL, D. (1993). Reconnaissance automatique d'un groupe nominal prépositionnel. Exemple des adverbes de date. *Lexique,* 11, pp.147-161.

Mejri, S. (2009). Le mot, problématique théorique. *Le Français Moderne*, 77(1), pp.68-82. PLANANS, Emmanuel, 2012. BiTermEx : un prototype d'extraction de mots composés à partir de documents comparables via la méthode compositionnelle. In : *Actes de la conférence conjointe JEP-TALN-RECITAL, Grenoble, 2012*. Vol(2) : TALN, pp.415-422.

QUINIOU, Solen, CELLIER, Peggy, CHARNOIS, Thierry AND LEGALLOIS, Dominique (2012). What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics?. *Computational Linguistics and Intelligent Text Processing*, *Heidelberg, 2012*. pp.166-177.

SEEKER, Wolfgang AND KUHN Jonas (2013). Morphological and Syntactic Case in Statistical Dependency Parsing. *Computational Linguistics*, Vol (3), pp.23-55.

SNOW, Rion, JURAFSKY, Daniel AND NG Andrew Y (2004). Learning Syntactic Patterns for Automatic Hypernym Discovery. *Advances in Neural Information Processing Systems*, *British Columbia, 2004*. pp.1297-1305.

PAREMIOLOGÍA BASADA EN CORPUS WACKY: ENFOQUE (INTRA- E INTER-) LINGÜÍSTICO Y CONCEPTUAL

Vincenzo Lambertini
Universidad de Bolonia
vincenzo.lambertini2@unibo.it

El estudio está encaminado a analizar los refranes bajo la vertiente conceptual, para categorizarlos según categorías conceptuales. De tener refranes en diferentes lenguas categorizados según los mismos criterios, podríamos obtener equivalentes funcionales de refranes cuando éstos compartan unas categorías. En este trabajo, los refranes serán en italiano y francés, pero se prestará mucha atención a lo que ocurre en español, para tener otro idioma de comparación.

En primer lugar, cabe colocar los refranes en el panorama más amplio de los fraseologismos. Para algunos investigadores los refranes no tienen mucho que ver con fraseología y expresiones idiomáticas, por su carácter composicional y su falta de fijación. Sin embargo, a través de la observación de los datos que nos brinda la lingüística de corpus, se nota que sí existen diferencias entre expresiones idiomáticas y refranes, pero también que tienen características comunes muy importantes y que algunos argumentos utilizados para demostrar su diversidad no son irrefutables.

A estas alturas, la lingüística de corpus se convierte en una herramienta para analizar los aspectos lingüísticos de los refranes así que buscar una metodología de categorización conceptual de los mismos, para agrupar estos elementos según rasgos conceptuales comunes.

Este estudio se inserta en una tesis doctoral llevada a cabo en el Departamento de Interpretación y Traducción (DIT) de la Universidad de Bolonia (Italia). El enfoque adoptado no es el "clásico" que se sigue para el análisis de paremias, es decir de carácter folklórico, histórico o simplemente filológico/lexicográfico. Esta investigación se propone satisfacer una necesidad práctica que es propia de los traductores e intérpretes: la traducción de refranes y, en general, la búsqueda de paremias equivalentes que se empleen realmente en la lengua de llegada en contextos parecidos.

Lamentablemente, escasean estudios pragmáticos sobre los refranes, por lo menos en italiano, nuestra lengua de partida. La primera dificultad ha sido encontrar los refranes más utilizados en italiano en el lenguaje común por falta de listas de frecuencias de refranes. Por lo tanto, la fase de recogida de los datos ocupa una parte relevante de este estudio.

Los datos y su análisis han sido útiles incluso para producir consideraciones lingüísticas sobre los refranes. Por ejemplo, la variabilidad de las paremias es algo muy frecuente pero es también lo que indica qué tiene que ser invariable para que el refrán no cese de ser reconocido (de la literatura paremiológica no se desprende claramente hasta qué punto se puede modificar un refrán) y cuáles son los conceptos más generales y abstractos que permanecen incluso en el refrán variado.

References

ANSCOMBRE, J.-C. (1997). Reflexiones críticas sobre la naturaleza y el funcionamiento de las paremias. *Paremia*, 6, pp.43-54.

ANSCOMBRE, J.-C. (2005). Les proverbes : un figement du deuxième type ? *Linx*, 53, [online] Available at: <http://linx.revues.org/255> [Accessed 11 October 2012].

BARBADILLO DE LA FUENTE, M. T. ET AL. (1997). *Refranero Multilingüe*. [online] Available at: <http://cvc.cervantes.es/lengua/refranero/Default.aspx> [Accessed 20 March 2015].

BARONI, M., BERNARDINI, S., FERRARESI, A. AND ZANCHETTA, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3), pp. 209-226.

BARSANTI VIGO, M. J. (2006). Problemática en torno al refrán y otras categorías paremiales. In: M. Alonso Ramos, ed. 2006. *Diccionarios y fraseología. Anexos de Revista de Lexicografía*, 3, pp.197-206.

BIDAUD, F. (2002). *Structures figées de la conversation. Analyse contrastive français-italien*. Bern et al.: Lang.

BOGGIONE, V. AND MASSOBRIO, L. (2007). *Dizionario dei proverbi*. Torino: UTET.

CABRÉ, M. T. (2000). Sur la représentation mentale des concepts. In: H. Bejoint and P. Thoiron, eds. 2000. *Le sens en terminologie*. Lyon: Presses universitaires de Lyon. pp.20-39.

CAMPOS, J. G. AND BARELLA, A. (1995). *Diccionario de refranes*. Madrid: Espasa Calpe.

CANELLADA, M. J. AND PALLARES, B. (2001). *Refranero español. Refranes, clasificación, significación y uso*. Madrid: Castalia.

CASADEI, F. (1996). *Metafore ed espressioni idiomatiche: uno studio semantico sull'italiano*. Roma: Bulzoni.

CORREAS, G. (1992). *Vocabulario de refranes y frases proverbiales y otras fórmulas comunes de la lengua castellana en que van todos los impresos antes y otra gran copa que juntó el Maestro Gonzalo Correas*. Madrid: Visor.

CRISTILLI, C. (1989). Il proverbio come esempio di testualità popolare. In: C. Vallini, ed. 1989. *La pratica e la grammatica. Viaggio nella linguistica del proverbio.* Napoli: Istituto Universitario Orientale, Dipartimento di studi letterari e linguistici dell'Occidente. pp.177-206.

DEIGNAN, A. (2009). Searching for Metaphorical Patterns in Corpora. In: P. Baker, ed. 2009. *Contemporay Corpus Linguistics*. London, New York: Continuum. pp.9-31.

DEPECKER, L. (2011). Comment aborder le concept d'un point de vue linguistique ? In: J.-J. Briu, 2011. *Terminologie (I): analyser des termes et des concepts.* Bern: Lang. pp.17-32.

DUNETON, C. AND CLAVAL, S. (1990). *Le bouquet des expressions imagées : encyclopédie thématique des locutions figurées de la langue française*. Paris: Seuil.

GROSS, G. (1996). *Les expressions figées en français.* Gap: Ophrys.

GUAZZOTTI, P. AND ODDERA, M. F., 2010. *Il grande dizionario dei proverbi italiani.* Bologna: Zanichelli.

HALLIDAY, M. A. K. (1992). Language Theory and Translation Practice. *Rivista internazionale di tecnica della traduzione*, 0, pp.15-25.

HERBST, T., FAULHABER, S. AND UHRIG, P. eds. (2011). *The Phraseological View of Language. A Tribute to John Sinclair.* Berlin: De Gruyter Mouton.

HONECK, R. P. (1997). *A proverb in mind: the cognitive science of proverbial wit and wisdom*. Mahwah (N. J.): Lawrence Erlbaum associates.

JUNCEDA, L. (1991). *Del dicho al hecho*. Barcelona: Obelisco.

KLEIBER, G. (1990). *La sémantique du prototype. Catégories et sens lexical*. Paris: PUF.

KLEIBER, G. (1999). Les proverbes: des dénominations d'un type «très très spécial». *Langue française*, 123, pp.52-69.

KLEIBER, G. (2000). Sur le sens des proverbes. *Langages*, 139, pp.39-58.

LAKOFF, G. AND JOHNSON, M. (1980). *Metaphors We Live By.* Chicago: University of Chicago.

LAPUCCI, C. (1993). *Dizionario dei modi di dire della lingua italiana*. Milano: Garzanti.

LONGHI, J. (2008). *Objet discursif et doxa. Essai de sémantique discursive*. Paris: L'Harmattan.

MALOUX, M. (1990). *Dictionnaire des proverbes, sentences et maximes*. Paris: Larousse.

MONTREYNAUD, F. ET AL. (1989). *Dictionnaire de proverbes et dictons*. Paris: Dictionnaires Le Robert.
NORRICK, N. R. (1985). *How proverbs mean: semantic studies in English proverbs*. Berlin: Mouton.
PERRIN, L. (2000). Remarques sur la dimension générique et sur la dimension dénominative des proverbes. *Langages*, 139, pp.69-80.
PHILIP, G. (2011). *Colouring Meaning. Collocation and connotation in figurative language.* Amsterdam: John Benjamins.
PITTANO, G. (1996). *Frase fatta capo ha: dizionario dei modi di dire, proverbi e locuzioni*. Bologna: Zanichelli.
POCCETTI, P. (1989). Aspetti della teoria e della prassi del proverbio nel mondo classico. In: C. Vallini, 1989. *La pratica e la grammatica. Viaggio nella linguistica del proverbio*. Napoli: Istituto Universitario Orientale, Dipartimento di studi letterari e linguistici dell'Occidente. pp.61-85.
QUARTU, B. M. (2001). *Dizionario dei modi di dire della lingua italiana: 10.000 modi di dire ed estensioni figurate in ordine alfabetico per lemmi portanti e campi di significato*. Milano: Rizzoli.
QUIROGA, P. (2006). *Fraseología italo-española. Aspectos de lingüística aplicada y contrastiva*. Granada: Granada lingüística.
REY, A. (1989). *Dictionnaire de proverbes et dictons*. Paris: Dictionnaires Le Robert.
SHAPIRA, C. (2000). Proverbe, proverbialisation et déproverbialisation. *Langages*, 139, pp.81-97.
SILVESTRI, D. (1989). Osservazioni in margine ai proverbi sumerici: strutture linguistiche e architetture testuali. In: C. Vallini, 1989. *La pratica e la grammatica. Viaggio nella linguistica del proverbio*. Napoli: Istituto Universitario Orientale, Dipartimento di studi letterari e linguistici dell'Occidente. pp.11-29.
SIMPSON, J. (1992). *The Concise Oxford Dictionary of Proverbs*. Oxford: Oxford University Press.
SINCLAIR, J. (2003). *Reading Concordances. An Introduction*. London: Pearson Longman.
TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work.* Amsterdam: John Benjamin.
TURRINI, G. ET AL. (1999). *Capire l'antifona*. Bologna: Zanichelli.
VISETTI, Y. M. AND CADIOT, P., 2006. *Motifs et proverbes. Essai de sémantique proverbiale.* Paris: Presses Universitaires de France.
XATARA, C. M. ET AL. (2009). *Dictionnaire d'expressions idiomatiques: Français – Portugais –Français.* [online] Available at: <http://www.cnrtl.fr/dictionnaires/expressions_ idiomatiques/> [Accessed 10 March 2015].

IDENTIFICACIÓN AUTOMÁTICA DE UNIDADES FRASEOLÓGICAS Y VALIDACIÓN LINGÜÍSTICA EN NOTAS PERIODÍSTICAS

Luis Meneses Lerín
Université d'Artois
jluis.meneseslerin@univ-artois.fr

En los últimos años, los medios de comunicación se han visto influenciados por los avances tecnológicos. Desde el punto de vista de los *usuarios*, es importante acceder a la información de manera rápida y a bajo costo. Desde el punto de vista del Procesamiento Natural del Lenguaje (PNL), los periódicos en línea permiten construir corpus homogéneos y ordenados por secciones (política, internacional, nacional, economía, deportes, etc.). Finalmente y desde el punto de vista de la lingüística de corpus, los artículos de periódicos en línea permiten realizar tareas que implican la búsqueda de información, la representación de conocimientos y el análisis lingüístico.

Uno de los grandes retos para procesar automáticamente los artículos de periódicos que circulan en Internet es el tipo de lenguaje empleado. Se trata de un lenguaje formal que utiliza el argot y las frases hechas o expresiones idiomáticas del tema en cuestión (sección). Este lenguaje se diferencia claramente del usado en obras literarias o en artículos científicos. El enfoque tradicional de la minería de textos utiliza la frecuencia de las palabras para discriminar/agrupar entre las categorías de una colección de documentos (Sebastiani, 2002). Sin embargo, este enfoque no da muy buenos resultados

cuando se trata de reconocer, identificar o detectar unidades fraseológicas, conocidas en el ámbito del PNL como secuencias multipalabras.

El interés de proponer subconjuntos de corpus tiene como finalidad reducir el área nocional del léxico para así tratar la ambigüedad de las palabras y los bloques de palabras que se repiten. Estos bloques son en su gran mayoría unidades fraseológicas que poseen la característica de ser “escurridizos” cuando se aplican enfoques estadísticos. Por lo cual, su identificación y validación resultan de suma importancia para así proponer un análisis estadístico que no sólo tome en cuenta las unidades lexicales sino también las unidades fraseológicas. De esta forma, las frases hechas o expresiones idiomáticas impiden contar con una modelización adaptada a las exigencias de los sistemas informáticos (cf. Laporte 1988; Danlos 1988) y, de ahí, surge la necesidad de contar con recursos lingüísticos, validados por lingüistas, suficientemente vastos y claramente estructurados con el objetivo de ser automatizados (cf. Habert 1991) y así poder identificar las *unidades fraseológicas*.

En esta presentación mostraremos que la identificación automática de las unidades fraseológicas que utilizan bases de datos requiere de una validación lingüística. La validación lingüística permite identificar: i) las unidades fraseológicas no-composicionales semánticamente, ii) las unidades fraseológicas composicionales semánticamente, es decir, las secuencias de palabras que no son unidades fraseológicas, iii) las diferentes acepciones que pueden presentar ciertas unidades fraseológicas (Meneses, 2012). Después, presentaremos algunos ejemplos de unidades fraseológicas en contexto identificadas de manera automática por medio de herramientas informáticas. Estos ejemplos nos permitirán mostrar que es posible, gracias al contexto, identificar rasgos morfosintácticos y sintáctico-semánticos que permitan “afinar” la identificación de unidades fraseológicas en corpus. Finalmente, mostraremos que la proyección de bases de datos de unidades fraseológicas en corpus es insuficiente. El contexto juega un papel importante en la identificación de unidades fraseológicas.

References

BUVET P-A., BLANCO Xavier (2001). "Classes d'objets et traduction automatique". L'Eloge de la différence: la voix de l'"autre" : VIe Journées

scientifiques du Réseau thématique de l'AUF, Lexicologie, terminologie, traduction, Beyrouth, Liban, 11.12.13 novembre 1999, pp. 345-353.
BUVET P-A. (2009a). Des mots aux emplois: la représentation lexicographique des prédicats. Le Français Moderne 77 (1), pp. 83-96.
CORPAS PASTOR, G. (1996). Manual de Fraseología española, Gredos, Madrid.
DANLOS L. (1884). «An algorithm for automatic generation», in Proceedings of the Tenth European Conference on Artificial Intelligence, T. O'Shea édition, Elsevier Science Publishers, Amsterdam.
DENICIA-CARRAL, C., MONTES-y-GOMEZ, M., VILLASENOR-PINEDA, L., and ACEVES-PEREZ, R. (2010). « Bilingual Document Clustering using Translation-Independent Features ». International Journal of Computational Linguistics and Applications, Vol. 1, No. 1-2.
GREVISSE M. (1975). Le bon usage. Dixième édition revue, Gembloux, Duculot.
GROSS Gaston. MASSOUSSI Taoufik (2011). "Figement et transparence". In ANSCOMBRE Jean-Claude, MEJRI Salah, (éds), Le figement linguistique: la parole entravée, pp. 95-108. H. Champion. Paris.
GROSS Gaston. (2010). La notion d'"emploi" dans une grammaire de prédicats. Cahiers de Lexicologie 96-1, pp. 97-115.
GROSS Gaston (2010). Sur la notion de contexte. Meta 55 (1), pp. 187-197.
GROSSMANN Francis et TUTIN Agnès. 2003. Quelques pistes pour le traitement des collocations. Les collocations. Analyse et traitement. Travaux de Recherches en linguistique appliquée, 1, pp. 5-21.
HABERT, B., NAZARENKO, A. et SALEM, A. (1997). Les linguistiques de corpus. A. Colin, Paris.
HARRIS Zellig. S. (1976). Notes du cours de syntaxe. Paris: Seuil.
ISSAC F., Hamon T., Fouquère C., Bouchard L., Emirkanien L. (2001). "Extraction informatique de données sur le web". DistanceS 5 (2), pp. 195-209.
ISSAC Fabrice (2011). "Figement et informatique". In MEJRI Salah, ANSCOMBRE Jean-Claude, (eds), Le figement linguistique: la parole entravée, pp. 413-431. H. Champion. Paris.
LAPORTE E. (1988). "La reconnaissance des expressions figées lors de l'analyse automatique", Langages 90, Les expressions figées, Laurence Danlos ed., Paris: Larousse, pp.117-126.
LE PESANT (Denis) et MATHIEU-COLAS (Michel): (1998). «Introduction aux classes d'objets», Langages 131, Larousse, Paris.
MEJRI Salah (2008). «Constructions à verbes supports, collocations et locutions verbales», Les constructions verbo-nominales libres et figées. Approches contrastives et traductologiques, p.191-200 Université d'Alicante.
MEJRI Salah (2000). "Figement lexical et renouvellement du lexique : quand le processus détermine la dynamique du système". Le français moderne LXVIII (1), p. 39-62.
MEJRI Salah (2009). Catégories linguistiques et étiquetage de corpus, L'information grammaticale, Peeters, Paris.
MEJRI Salah (2011). "Les Dictionnaires électroniques sémantico-syntaxiques". In CARDOSO Suzana Alice Marcelino, MEJRI Salah, MOTA Jacyra

Andrade, (eds), Os di.ci.o.ná.rios : fontes, métodos e novas tecnologias, pp. 159-187. Instituto de Letras da Universidade Federal da Bahia.
MENESES LERIN Luis (2011). "La polysémie et le réseau synonymique des prédicats polylexicaux". Neophilologica 23 , pp. 84-99.
MENESES LERIN Luis (et al.) (2011). "INAOE @ DEFT 2011: Using a plagiarism detection method for pairing abstracts-scientific papers". TALN 2011, Montpellier, 27 Juin-1er Juillet 2011.
MENESES LERIN Luis (2012). "Les emplois et les unités diatopiques mexicaines". Lenguas especializadas, fijación y traducción = Langues spécialisées, figement et traduction, Encuentros Mediterráneos; 4 pp. 49-63. Université d'Alicante. Espagne.
MENESES LERIN L. (2012). Dictionnaires électroniques monolingues coordonnés du verbe donner français (France) – espagnol (Espagne) – espagnol (Mexique) : approche syntactico-sémantique. Thèse de doctorat en Sciences du langage. Université Paris XIII (Pres Sorbonne Paris-Cité), 372 p.
MOGORRON HUERTA P. (2002). La expresividad en las locuciones verbales en francés y en español. Publicaciones Universidad de Alicante.
MOGORRON HUERTA P., MEJRI Salah, (eds) (2012). Lenguas especializadas, fijación y traducción. Langues spécialisées, figement et traduction, Encuentros Mediterráneos; 4 Université d'Alicante, Espagne.
NEVEU F. (2004). Dictionnaire des sciences du langage, Armand Colin.
RUWET N. (1983). «Du bon usage des expressions idiomatiques dans l'argumentation en syntaxe générative», Recherches linguistiques n° 11, p.5-84, Université Paris VIII, Vincennes. Syntaxe et sémantique n°5, 2003: Polysémie et polylexicalité.
SVENSSON, Maria Helena (2004). Critères de figement. L'identification des expressions figées en français contemporain (Doctoral thesis). Umeå Universitet, 2004.
ZULUAGA, A. (1980). "Introducción al estudio de las expresiones fijas", en Studia Romanica et Linguisticia, 10, Frakfurt-Berna-Cirencester, Peter D.Lang.

CONTRASTIVE ANALYSIS OF PHRASEOLOGICAL UNITS WITH SPECIFIC ANIMAL CONSTITUENTS IN ENGLISH, SPANISH AND GERMAN

Marta Morer Murcia

Universidad de Murcia

marta.morer@um.es

In the following lines I present a brief introduction about a contrastive analysis in which three different languages (English, Spanish, and German) are contrasted in order to examine the degree of equivalence that their phraseological units (PUs) with specific animal constituents present. More precisely, the analysis uses this comparison to classify the languages in different groups based on semantic connotations and several idiosyncrasies of the PUs following Corpas Pastor's (2003) degrees of equivalence. The PUs taken into consideration to build the corpus refer to those idioms and routine formulas that have the lexical element of specific animals. The animal constituents that I have selected are *cat, dog, horse* and *monkey*. After the analysis of these specific English PUs, the work shows the type of equivalence that they have in the other two languages and the classification of these units depending on their degrees of equivalence.

Due to its importance in the study of language as well as the difficulties it frequently presents in translation and second language teaching, phraseology plays a significant role in the field of linguistics. Some PUs can present special

elements which, in some cases, make these tasks easy and, in others, complex; this is the case with PUs referring to specific animals. As a result, this topic required further study. Furthermore, we must consider the fact that there are several contrastive studies which have analysed the PUs of two languages but there are not many phraseological studies in which several languages are compared in order to look for the equivalences of the PUs between them. That is the aim of the present study: to contrast three different languages with the purpose of collecting data about their relative degrees of equivalence of selected PUs.

The work is organized as follows: first, the theoretical base required to undertake and interpret the study is presented; then, the main aims of the study are explained, as is the purpose of analysing the PUs with specific animal constituents in the three different languages. The results, which are interpreted in the last part of the paper, state the specific findings that I get from comparing the PUs. Graphs and tables are used to illustrate selected examples as well as the percentages of equivalence found in the different types. The whole corpus of the PUs with the specific animal constituents used in the study is provided in the appendix.

Due to the study's strict aims, the study focuses on two specific types of PUs. These PUs are idioms and phraseological utterances as routine formulas. Therefore, the corpus was selected along highly-specific lines: those PUs as idioms or routine formulas with the specific animals' constituents of *dog, cat, horse* and *monkey* in English, Spanish and German selected from the Oxford English Dictionary (OED). The corpus is divided into three different columns depending on the languages of the PUs, as well as into three different sections, which classify them according to the degree of equivalence that these PUs have between the three languages.

References

Alexander, R. (1984). *Fixed expressions in English: reference books and the teacher.* ELT Journal, Vol. 38/2. Oxford: Oxford University Press.

Burger, H. (1998). *Phraseologie. Eine Einführung am Beispiel des Deutsches*. Berlin: Erichschmidt.

Corpas Pastor, G. (1997). *Manual de fraseología española*. Madrid: Editorial Gredos.

Corpas Pastor, G. (2003). *Diez años de investigación en fraseología, análisis sintáctico semánticos, contrastivos y traductológicos*. Madrid: Iberoamericana

Diccionario de María Moliner. (1998). 2nd Ed. Madrid: Editiorial Gredos

Diccionario práctico de locuciones y Frases hechas. (1998. Madrid: Editorial Everest.

Drosdowski, B. and Scholze-Stubenrecht, W. (1992). *DUDEN. Redewendungen und sprichwörtliche Redensarten. Band 11.* Mannheim: Dudenverlag.

Fernando, C. and Flavel, R. (1981). *On Idiom. Critical views and perspectives*. Vol. 5. Exeter: University of Exeter

Gelbrecht, A. (2011). *Phraseology in Intercultural Communication. A contrastive approach towards German and English phraseological units of fire and water.* München: GRIN.

Gläser, R. (1986). *Phraseologie der englischen Sprache.* Tübingen: Niemeyer.

Granger, S. and Meunier, F. (2008). *Phraseology. An interdisciplinary perspective.* Amsterdam: John Benjamins Publishing Company

Herbst, T., Faulhaber, S. and Uhrig, P. (2011). *The phraseological view of language. A tribute to John Sinchair*. Berlin: De Gruyter Mouton.

Katz, J. and Postal, P. (1963). *Semantic interpretation of idioms and sentences containing them. Quarterly Progress Report.* Massachusetts: MIT Press.

Makkai, A. (1972). *Idioms structure in English.* The Hague: Mouton.

Nash, R. (1973). *Reading in Spanish-English Contrastive Linguistics.* Puerto Rico: Inter American University Press.

Oxford English Dictionary Online. <http://www.oxforddictionaries.com>, Accessed 15th January 2015.

Zuluaga, A. (1980). *Introducción al estudio de las expresiones fijas.* Studia Romantica et lingüística; 10. Frankfurt am Main: Peter Lang.

"*PALE AS DEATH*" OR "*PÂLE COMME LA MORT*": FROZEN SIMILES USED AS LITERARY CLICHÉS

Suzanne Mpouli
LIP6 (Université Pierre et Marie Curie) & Labex Obvil (Paris IV)
mpouli@acasa.lip6.fr

Jean-Gabriel Ganascia
LIP6 (Université Pierre et Marie Curie) & Labex Obvil (Paris IV)
jean-gabriel.ganascia@lip6.fr

The present study is focused on the automatic identification and description of frozen similes in in British and French novels written between the 19th century and the beginning of the 20th century. Similes are figures of speech in which two or more nouns pertaining to different semantical fields are compared by virtue of a set of common shared properties that are implicit or explicitly stated. Since they are derived from comparative structures and are powerful tools to create striking images, similes are popular in everyday language, so much so that most languages possess a significant number of frozen similes used to describe an action, an entity or a phenomenon qualitatively. Even though literary style is mainly associated with creative writing and deviations from stereotyped expressions, it is generally acknowledged that clichés can be used for stylistic purposes, be it to add realism or local flavour in characterisation, to bring about humour, to create contrast inside the text or to make reference to popular sayings, other authors as well as the culture of the time.

In practice, the canonical simile structure of a sentence such as "*His face grew black like the night*" can be rendered by the following expression, A Ω *y* X B. More specifically, A, the source term (*face*) of the comparison is generally referred to as the tenor as opposed to B, its target term (*night*), which is called

the tenor. In addition, a comparator or simile marker (X), in this case, *like*, which states to what extent the tenor and the vehicle are similar, is a compulsory element of any simile, unlike Ω and *y* which stand respectively for the eventuality or verb (*grew)* and for the ground or shared property (*black)*.

In accordance with previous works, we defined two main patterns of frozen similes: adjectival ground + simile marker + nominal vehicle (e.g. *happy as a lark)* and event + simile marker + nominal vehicle (e.g. *sleep like a top*). Furthermore, apart from traditional simile markers such as the French adverb 'comme' and the English preposition 'like', were also considered as simile markers, comparatives as well as other conjunctions, like the French conjunction 'ainsi que' , which can be used to introduce a simile. In order to extract similes from our two corpora of novels, we used a rule-based algorithm which first singles out potential simile sentence candidates and then identifies each of its components. Alongside frequency, the semantic correlation between every identified nominal tenor and its corresponding nominal vehicle was taken as a relevant feature for differentiating frozen similes from literal ones. Furthermore, we proposed a framework to describe each extracted frozen similes based on two main criteria: (i) a scale of literary clichédness that takes into account its frequency and (ii) its recurrent combination with a specific tenor to form a frozen image.

References

BOLSHAKOV, I. A. (2003). Simile cliché phrasemes in colloquial language. *Proceedings of the First International Conference on Meaning-Text Theory*, [online] Available at <http://meaningtext.net/mtt2003/proceedings/10.Bolshakov.pdf> [Accessed 30 March 2015].

BOUVEROT, D. (1969). Comparaison et métaphore. *Le Français moderne*, 37 (2), pp. 132-147, 224-238, 301-316.

FISHELOV, D. (1993). Poetic and non-poetic simile: Structure, semantics, rhetoric. *Poetics Today*, 14 (1), pp. 1-23.

GLUCKSBERG, S. AND KEYSAR, B. (1990). Understanding metaphorical comparisons: Beyond similarities. *Psychological Review*, 97 (1), pp. 3-18.

HANKS, P. (2005). Similes and sets: the English preposition like. In: R. Blatna and V. Petkevič, eds. 2005. *Languages and Linguistics: Festschrift for Fr. Cermak*. Prague: Charles University, Philosophy Faculty.

HANKS, P. (2008). How to say new things: An essay on linguistic creativity. *Brno Studies in English*, 34, pp 39-50.

MOON, R. (2008). Conventionalized as –similes in English. *International Journal of Corpus Linguistics*, 13(1), pp. 3-37.

NICULAE, V. (2013). Comparison pattern matching and creative simile recognition. *Proceedings of the Joint Symposium on Semantic Processing, Textual Inference and Structure in Corpora*, pp. 110-114.

NICULAE, V. AND YANEVA, V. (2013). Computational considerations of comparisons and similes. *Proceedings of the ACL Research Student Workshop*, pp. 110-114.

PERRIN-NAFFKAH, A. (1985). *Le cliché de style en français moderne : nature linguistique et rhétorique, fonction rhétorique.* Bordeaux : PU Bordeaux.

RIFFATERRE, M. (1964). Fonctions du cliché dans la prose littéraire. *Cahiers de l'Association internationale des études françaises*, 16, pp. 81-95.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44-49.

SHIE, J. (2007). The semiotic structure and semantic composition of English similes. *Journal of Humanities and Social Sciences*, 3 (1), pp. 57-68.

VEALE, T. (2012). A computational exploration of creative similes. In: MacArthur, F. (Oncins-Martínez, J.L. (Sanchez-García, M. and Piquer-Píriz, A. M. (eds. *Metaphor in Use: Context, Culture and Communication*, pp. 329-344.

VEALE, T. AND HAO, Y. (2007). Learning to understand figurative language: From similes to metaphor to irony. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, pp. 683-688.

VEALE, T. (HAO, Y. (AND LI, G. (2008. Multilingual Harvesting of Cross-Cultural Stereotypes. *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics,* pp. 523-531.

WIKBERG, K. (2008). Phrasal similes in the BNC. In: S. Granger and F. Meunier, eds. 2008. *Phraseology: An Interdisciplinary Perspective*. Amsterdam and Philadelphia: John Benjamins*,* pp. 127-142.

EXTRACTING TERMS WITH EXTRA

Lucia C. Passaro

Università di Pisa

lucia.passaro@for.unipi.it

Alessandro Lenci

Università di Pisa

alessandro.lenci@unipi.it

The identification and extraction of terms play an important role in many areas of knowledge-based applications, such as automatic indexing, knowledge discovery and management. The main distinction among term recognition approaches is between algorithms that take into account the distributive properties of terms, such as frequency and TF/IDF (Salton and McGill, 1983), and extraction techniques that use contextual information (Frantzi and Ananiadou, 1999; Maynard, 2000; Bonin et al., 2010). A common trait to most of these strategies is the identification of a set of candidates from texts, and then the application of a filtering function that separates the terms from non-terms.

With EXTra, we propose a new approach that jointly considers contextual and distributive properties. To address this issue, instead of using flat POS-sequences, we extract the candidates using structured POS-sequences that take into account the internal syntactic structure of term phrases, and then we calculate a particular association measure that allows us to consider the term as a “composition” of sub-terms. The intuition is that the degree of termhood of a candidate pattern depends not only on the statistical association between its parts, but also on the fact that these parts are also terms. EXTra works in three basic steps:

1. Candidates. Candidate terms are identified using structured POS-sequences. For example, the pattern [[noun(-s), preposition(-e), noun(-s)], preposition(-ea), [noun(-s), adjective(-a)]] allows us to extract the candidate [politica-s di-e sviluppo-s] delle-ea [risorse-s umane-a]] (human resource

development policy). Ignoring prepositions and following the parentheses order, EXTra stores the statistical information of the sub-patterns starting from the nested pairs. In this example, at the first step EXTra evaluates the pairs (risorse-s,umane-a) and (politica-s, sviluppo). At the second step, it considers the aggregate pair (politica_di_sviluppo, risorse_umane).

2. Term weighting. Pattern structure is also used to guide the process of statistical term weighting by following the same order of incremental composition: in the base step, we measure the association strength σ of each 2-word term <w1, w2> using standard measures such as Local Mutual Information, Log-Likelihood, etc. In the recursive step, we calculate σ(c1, c2) as S(c1)*S(c2), where c1, c2 or both belong to terms T. S(ci)=1 if ci is a word and S(ci)=(log2(σci))/k if ci Î T. Using this algorithm, we are able to play up long terms, so as to balance the different frequency between bigrams and n-grams. The salience of sub-terms depends on the parameter k, which allows us to control the possibility to select long terms: the smaller is k, the higher weight is assigned to longer terms.

3. Term selection. The candidate terms whose score exceeds an empirically fixed threshold are added to the set of terms.

What we're proposing with EXTra, is a language-independent methodology, whose application and exploitability is reasonably economic and reliable. The final paper will mainly focus on illustrating the methodology we're proposing, and will describe in details the structure of the resource and its algorithms. We will also evaluate EXTra on an experiment to extract Italian terms from the domain of public administration.

References

BONIN, F., DELL'ORLETTA, F., MONTEMAGNI, S., VENTURI, G. (2010). *A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora*. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta (Malta), May 19-21 2010: European Language ResourcesAssociation (ELRA).

FRANTZI K.T. and ANANIADOU S. (1999). *The C-Value/NC-Value domain independent method for multi-word term extraction*. Journal of Natural Language Processing, 6(3):145–179.

MAYNARD D.G. (2000). Term *Recognition Using Combined Knowledge Sources*. PhD thesis, Manchester Metropolitan University, UK, 2000.

SALTON,G. AND MCGILL M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

STATISTICAL AUTOMATIC EXTRACTION OF V-N ITALIAN COLLOCATIONS FROM AN ACADEMIC SPOKEN CORPUS

Diana Peppoloni

DITALS, University for Foreigners of Siena

dianapeppoloni@gmail.com

The aim of the present paper is to describe the results of a method of semi-automatic extraction of Italian V-N collocations, taken from the annotated medium size corpus ASIC (Academic Spoken Italian Corpus) (Peppoloni, 2012).

This corpus consists of audio recordings of academic lectures on various subjects, including about 500,000 words.

Collocations, that represent a subclass of multi-word expressions, are widely recognized as playing an important role in various fields of language research; after the early interest in the domain of language teaching (Palmer, 1933), they occupy, ever more frequently, a central position in the field of lexicography (Benson, Benson & Ilson, 1986; Benson, 1990; Cowie, 1981, Granger & Meunier, 2008), natural language processing (Smadja, 1993; Calzolari et al. (2002; Sag et al. (2002), corpus linguistics (Sinclair, 1991) and language acquisition (Nesselhauf, 2005).

From the many different attempts to define these phenomena, in this research it is used the following definition of it: “A collocation is a word combination whose semantic and/or syntactic properties cannot be fully

predicted from those of its components, and which therefore has to be listed in a lexicon" (Evert, 2005: 17).

Many researches (Biber, 2006) have analysed academic lexicon starting from collocations, that is to say recurrent combinations of words that tend to occur together to form fixed expressions with a global and conventional meaning. A misuse of the terms that make up collocations, can lead to misunderstandings in communication, altering the message that words bring. To know a word, means to know the other terms with which this usually combines; Firth (1957) says "you shall know a word by the company it keeps". This process promotes an easier way in producing and understanding sentences and concepts. Native speakers do not store individual lexical entries, but rather entire fixed words sequences, not having to rebuild it all the time, but using it already formed.

In order to extract V-N collocations from the corpus ASIC, we have tested the suite of statistical tools offered by the CWB platform, brought up by the Institute for Natural Language Processing of the University of Stuttgart. Its central component is the flexible and efficient query processor CQP. Our script provides the possibility to insert optional linguistic constituents, placed between the possibly identified verbs and nouns, in the automatic research; the resulting research pattern is the following: VERB (ADVERB) (ARTICLE) (ADJECTIVE OR NUMBER) NOUN.

The result of this computational operation is a list of 344 V-N collocations, with a frequency of occurrence equal to or greater than 3. Not all the identified words combinations corresponded though in actual Italian collocations. These were then judged by 50 Italian native speakers non-linguistic experts, who evaluated respectively groups of 10 collocations, indicating which, according to them, were or were not valid for the Italian language. This crowd sourcing experiment has allowed us to obtain impartial linguistic data, not influenced by the professional background of the speakers as linguists, but only by their linguistic competence as native speakers.

References

BENSON, M. (1990). Collocations and general-purpose dictionaries. International Journal of Lexicography, pp. 3: 23-34.

BENSON, M. BENSON, E. AND ILSON, R. (1986). Lexicographic Description of English. Amsterdam/Philadelphia: Benjamins.

BIBER, D. (2006). University Language. A corpus-based study of spoken and written register. Amsterdam/Philadelphia: John Benjamin Publishing Company.
CALZOLARI, N. FILLMORE, C. GRISHMAN, R. IDE, N. LENCI, A. MACLEOD, C. AND ZAMPOLLI, A. (2002). Towards best practice for multiword expressions in computational lexicons. In Proceedings of LREC-2002, Las Palmas, pp. 1934-1940.
COWIE, A. P. (1981). The Treatment of Collocations and Idioms. Learners' Dictionaries Applied Linguistics II(4), pp.223-235.
FIRTH, J. (1957). Papers in Linguistics, 1934-1951. Oxford: Oxford University Press.
GRANGER, S. AND MEUNIER, F. eds. (2008). Phraseology in Foreign Language Learning and Teaching. Amsterdam: John Benjamins.
NESSELHAUF, N. (2005). Collocations in a Learner Corpus. Studies in Corpus Linguistics. Amsterdam: John Benjamins.
PALMER, H. E. (1933). Second Interim Report on Collocations. Tokyo: Kaitakusha.
PEPPOLONI, D. (2012). 'Linguistic and computational tools in support of non native Italian speaking students: the development of the Academic Spoken Italian Corpus'. In Llanes, A. (Astrid, L. (Gallego, L. e Mateu, R. (eds. 2012. La lingüística aplicada en la era de la globalización, Lleida: Edicions i Publicacions de la Universitat de Lleida. pp.322-330.
SAG, I.A. BALDWIN, T. BOND, F. COPESTAKE, A. AND FLICKINGER, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. Proceedings of CICLing-2002. Mexico City, Mexico, pp.1–15.
SINCLAIR, J. (1991). Corpus, Concordance and Collocation. Oxford: Oxford University Press.
SMADJA, F. (1993). Retrieving collocations from text: Xtract. Comput. Linguist. 19, 1 (Mar. 1993), pp.143-177.

A SEMI-AUTOMATIC ALGORITHM FOR THE IDENTIFICATION AND EXTRACTION OF MWUS IN BILINGUAL PARALLEL CORPORA

Éric Poirier

Université du Québec à Trois-Rivières

eric.poirier@uqtr.ca

Although probably universal, phraseological units or multi-word units (MWUs) are distinctly molded out of specific usages of words and expressions in languages. The alternance of the idiom and the open choice principles (Sinclair, 1991) at play in the monolingual production of texts is constantly challenged by the translator in the constrained production of an equivalent text in the target language. Except for borrowings, a source language MWU never emerges intact in the target text and cannot be translated word-for-word. In the translation of pragmatic texts, as opposed to aesthetic and literary texts, it is either substituted with a functionally equivalent MWU linguistically or phraseologically different in every respect, or translated freely (semantically) giving preference to its meaning based on its specific usage in the source text. In both cases, it should be noted that all odds are against any isomorphism between source language MWUs and their target language translations. The content word precision algorithm exploits the contrastive manifestation of MWUs in languages as shown in their heteromorphical translation in a second language.

Written in Python language, the content-word precision algorithm compares the number of content words that are calculated in source and target segments provided in aligned segments created by translation memory software or alignment tools. The algorithm automatically and efficiently compares the content word count in the source segments with that of the target segments in any bitext. This automatized operation of the algorithm allows for the mathematical comparison of the content word count in the source segment with the content word count of the target segment, which can be either equal (possibly no MWUs), lower (potential MWU in source segment) or higher (potential MWU in target segment). With this data, the user can pinpoint uneven segments for the manual extraction of MWUs in source or target segments (depending on the volume of the word count). Although its success rate is not perfect, this simple method makes it possible to identify rapidly and objectively, a significant number of MWUs, in any bitext, without consideration to their nature (idioms, compounds, lexical bundles, formulaic expressions or even phrasal explicitations and implicitations in translation). In our presentation, we present statistical data compiled after the application of the algorithm to pragmatic text types.

References

CARTIER, E. (2008). Repérage automatique des expressions figées: état des lieux, perspectives. In: P. Blumenthal and S. Mejri, ed. 2008. *Les séquences figées: entre langue et discours.* Stuttgart: Franz Steiner Verlag, pp.55-70.

COLSON, J.-P. (2008). Cross-linguistic phraseological studies: An overview. In: Granger, S. and Meunier F. eds. (2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins Pub.

DELISLE, J. AND FIOLA, M. (2013). *La traduction raisonnée.* 3rd ed. Ottawa: Presses de l'Université d'Ottawa.

ERMAN, B. and WARREN, B. (2000). The idiom principle and the open choice principle. *Text*, 20(1), pp. 29-62.

FAZLY, A., COOK, P. and STEVENSON, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1), pp. 61-103.

GANESAN, K. (2015). All About Stop Words for Text Mining and Information Retrieval. *Text Mining, Analytics & More* [blog]. < http://www.text-analytics101.com/2014/10/all-about-stop-words-for-text-mining.html> [Accessed 29 March 2015]

GRANGER, S. (2014). A lexical bundle approach to comparing languages: Stems in English and French, *Languages in Contrast*, 14(1), 58-72.

GRANGER, S. and MEUNIER F. eds. (2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins Pub.

KOEHN, P. (2010). *Statistical Machine Translation.* Cambridge: Cambridge University Press.

Mel'čuk, I. (1993-2000). *Cours de morphologie générale, Vol. 1-5*. Paris/Montreal: CNRS/Presses de l'Université de Montréal.

MELAMED, I. D. (1997). Automatic discovery of non-compositional compounds in parallel data. *arXiv preprint cmp-lg/9706027*.

MOON, R. (1998). *Fixed Expressions and Idioms in English. A Corpus-based Approach.* Oxford: Clarendon Press.

PAPINENI, K., ROUKOS, S. WARD, T., AND ZHU, W. J. (2002). BLEU: a method for automatic evaluation of machine translation". *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL), 311-318.

PECMAN, M. (2005). De la phraséologie à la traductologie proactive : essai de synthèse des fondements théoriques sous-tendant la recherche en phraséologie. *Meta*, 50(4), 402-410.

POIRIER, E. (2003). Conséquences didactiques et théoriques du caractère conventionnel et arbitraire de la traduction des unités phraséologiques. *Meta*, 48(3), 402-410.

SINCLAIR, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

SIYANOVA-CHANTURIA, A. and MARTINEZ, R. (2014). *The Idiom Principle Revisited*, Applied Linguistics, Advance Access published January 26, 2015, doi:10.1093/applin/amt054, pp. 1-22.

PORTUGUESE PROVERBS: TYPES AND VARIANTS

Sónia Reis
University of Algarve
a17930@ualg.pt

Jorge Baptista
University of Algarve
jbaptis@ualg.pt

Proverbs are an important part of most societies' culture and language. As micro-texts, brought into discourse from the common cultural repository, they are subject to many creative types of variation. On the other hand, functioning as in quotation mode, they integrate discourse in an almost disruptive way, challenging natural language processing (NLP) systems, and requiring their accurate identification and delimitation.

Concerning Portuguese proverbs, and though several, extensive collections of proverbs are available in printed form (Machado, 2011), to the best of our knowledge, no resources have been specifically produced for NLP purposes, even if some digitally available dictionaries (Almeida, 2014) include a few examples, interspersed between other type expressions, like different types of idioms and many forms of slang.

Recently, Rassi *et al*. (2014) have proposed a formal (syntactic) classification of Portuguese proverbs, based on a collection over 3,500 proverb variants, organized in 594 proverb types (Rassi, 2014), and taken from several dictionaries from the Brazilian variety of the language. The authors presented a finite-state based method for the automatic identification of proverbs in large-sized corpora, and experimented on a 29M tokens corpus of journalistic text (Bruckschein *et al*. 2008), taken from the daily online edition of the Brazilian newspaper *Folha de São Paulo*. The authors report a 60 to 73% precision, depending on the proverb class and the width of the insertion window between

the proverbs' keywords. In spite of the corpus size, but not surprisingly, only 137 types and 788 instances were matched, most likely because of the journalistic nature of the texts in this corpus. However, seen from this side of the Atlantic, results from Rassi and colleagues are surprising mostly for the fact that, in spite of some American idiosyncrasies, most proverbs seem to exist also in the European variety, quite unlike the mismatch that has been found for verbal idioms.

Drawing on this methodology and these previous results, this paper reports on an extension of that experiment, but now focused on the identification of the European Portuguese proverbs and their variants on a much larger, 190M words corpus of journalistic text (Rocha and Santos, 2000), taken from the Portuguese daily online edition of the newspaper *Público*. Our goal is to assess differences in frequency and lexical coverage, to determine the best-performing, class-specific, insertion window width. This intends to set up the basis for a large collection of Portuguese proverbs and their variants, specifically built for natural language processing, and to make it publicly available, along with the finite-state tools built for retrieving them from texts. These tools and resources will undoubtedly be deemed useful assets to other paremiology studies.

References

ALMEIDA, J.J. (2014). *Dicionário aberto de calão e expressões idiomáticas*. [online] Available at:<http://natura.di.uminho.pt/~jj/pln/calao/dicionario.pdf> [Accessed 14 March 2015].

BRUCKSCHEIN, M., MUNIZ, F., SOUZA, J., FUCHS, J.T., INFANTE, K., GONÇALEZ, P.N., VIEIRA, R. and ALUISIO, S.M. (2008). *Anotação linguística em XML do corpus PLN-BR. Série de Relatórios do NILC*, São Carlos (SP): NILC-ICMC-USP.

MACHADO, J.P. (2011). *O Grande Livro dos Provérbios*, 4ª ed., Lisboa: Casa das Letras.

RASSI, A.P. (2014). *List of Proverbs in Brazilian Portuguese*. <https://www.researchgate.net/ publication/269165152_Rassi2014> [Accessed 14 March 2015]. DOI: 10.13140/2.1.4907.7280

RASSI, A.P., BAPTISTA, J. AND VALE, O. (2014). Automatic Detection of Proverbs and their Variants. In: M. Pereira, J. Leal, J. and A. Simões, eds. *Proceedings of the Symposium on Languages, Applications and Technologies* (SLATE`14). Leibniz (Germany): Schloss Dagstuhl - Leibniz-Zentrum fur Informatik, Dagstuhl Publishing. pp. 235-249.

ROCHA, P. AND SANTOS, D. (2000). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: Nunes, M.G. *et al*., eds., *V Encontro para o processamento computacional da língua*

portuguesa escrita e falada (PROPOR'2000), São Paulo: ICMC/USP. pp. 131–140.

‘X ME NO XS’ – A CORPUS-BASED CASE STUDY

Gustavo A. Rodríguez Martín

Universidad de Extremadura

garoma@unex.es

Phraseological variation and modification are two of the most complex areas of study within general phraseology, particularly because of the challenges they pose (Cf. Jaki, 2014: 17 et passim). Studies in this interesting area have flourished in recent years, thanks in part to the advent of powerful digital media and large corpora to work with. In all, modification is especially significant when a given phraseological frame (Martin, 2008) or kernel (Norrick, 1985) is very productive and is exploited across genres over a long period.

The use of some of these phraseological frames has been attested in literary discourse for centuries, as exemplified by the structure ‘X me no Xs’ – a frame that has been used by authors like Shakespeare (“Grace me no grace,” Richard II), Henry Fielding (“Petition me no petitions,” Tom Thumb), and Jack Kerouac (“Tennessee me no Tennessees,” Visions of Cody), to name but a few. Indeed, this kernel received some scholarly attention in the past, to the extent that Potter (1915) published a gloss of some of the examples he had come across in literary works. His study, however, is little more than a curious note because it is restricted to literature and says nothing of the semantic, pragmatic, and stylistic functions that the different variations serve.

The purpose of this paper is to make up for the aforementioned deficiencies in the study of such a productive phraseological frame. It is my contention that “X me no Xs” is, on the one hand, used far more often in non-literary genres

than previous studies seem to suggest. In addition, the stylistic potential of this frame – together with the fact that it is commonly misattributed as having been coined by Shakespeare – makes this phraseological kernel worthy of further investigation in specific contexts. This study will make a corpus-based analysis of the data retrieved from selected corpora in order to arrive at conclusions on the semantic, pragmatic and stylistic behavior of this frame.

References

Jaki, S. (2014). *Phraseological Substitutions in Newspaper Headlines: "More than Meats the Eye".* Amsterdam: John Benjamins.

Norrick, N. (1985). *How Proverbs Mean*. New York: Mouton.

Martin, W. (2008). A unified approach to semantic frames and collocational patterns. In: S. Granger and F. Meunier, eds. *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins, pp. 51-66.

Potter, A. C. (1915). But Me No Buts. *Modern Language Notes*, 30(5), 160.

AUX ANGES OU IN SEVENTH HEAVEN: IDENTIFICATION D'UNITES PHRASEOLOGIQUES ET EQUIVALENCE SEMANTIQUE DANS LA TRADUCTION

Dorota Sikora
Université du Littoral – Côte d'Opale
Dorota.Sikora@univ-littoral.fr

Notre réflexion sera consacrée au mode de traitement linguistique et à l'encodage des phrasèmes de type locution forte (Mel'čuk, 2013) dans deux bases de données lexicales développées actuellement à l'Atilf (CNRS UMR 7118): Réseau Lexical du Français (RL-fr, cf. Gader et al. (2012, Polguère, 2014), et Réseau Lexical de l'Anglais (en-LN, Gader et al. (2014). Il s'agit de ressources construites en parallèle (bien qu'à des rythmes différents), selon les principes théoriques et méthodologiques homogènes, définis initialement dans le cadre de la Lexicologie Explicative et Combinatoire et développés récemment par Polguère (2009, 2014) sous le nom de systèmes lexicaux. Nous présenterons les méthodes d'identification de ces phrasèmes en vue de leur intégration dans un modèle du lexique de la langue L1 et de la langue L2.

Comme le montre Vaguer (2010), dans les ressources actuellement disponibles, la couverture reste faible en ce qui concerne les unités phraséologiques (UP). Elle est de plus hétérogène: même les locutions les plus fréquentes ne s'y retrouvent pas systématiquement. Dans une perspective bilingue, la probabilité de trouver l'équivalent d'une UP de la langue L1 dans la langue L2 est par conséquent faible.

Quand un logiciel de traduction automatique encode des locutions sémantiquement équivalentes, la fiabilité des résultats obtenus laisse à désirer. En effet, la traduction de la phrase (1a) est satisfaisante du point de vue d'équivalence sémantique, alors que celle en (2b) ne rend aucunement le caractère idiomatique de l'original en (2a) :

1. a. Mon amour, je vous aime de toutes mes forces et je suis tout aux anges de vous revoir bientôt. [FRANTEXT]

   b. My love, I love you of all my strengths and I am quite in seventh heaven to see again you soon. [traduction Reverso, 18/03/2015]

2. a. J'installe mes affaires et me sens aux anges.
   [FRANTEXT]

   b. I install(settle) my business(affairs) and feel to the angels. [traduction Reverso, 18/03/2015]

Notre communication se focalisera sur l'un des problèmes centraux de la traduction automatique, celui de l'identification des locutions et de la détermination précise de leur nature linguistique. En considérant le cas précis des phrasèmes être aux anges et to be in seventh heaven, nous présenterons le mode d'identification et d'encodage des locutions fortes dans les bases lexicales RL-fr et en-LN. Conformément aux critères adoptés, dans le premier cas, nous avons affaire à une collocation construite sur aux anges, locution prépositionnelle à valeur adjectivale qui forme sa base, et le verbe être sélectionné comme collocatif. Parallèlement, to be in seventh heaven s'avère une collocation construite autour d'une base nominale (seventh heaven).

Au-delà de ces cas précis, nous tenons à montrer qu'une identification appropriée des UP encodées dans une ressource lexicale permet de déterminer leur place dans le système lexical de L1, de préciser les liens paradigmatiques et syntagmatiques (Fontanelle, 2008, Sikora et Polguère, 2013) qui les rattachent à d'autres unités lexicales, notamment monolexémiques pour, à termes, établir des équivalences sémantiques (Svensén, 2009) entre des UP de langues différentes.

References

Fontanelle, T. (2008). Using a Bilingual Dictionary to Create Semantic Networks. In: T. Fontanelle, ed. 2008. Practical Lexicography. A Reader. Oxford, New York: Oxford University Press, pp. 169-189.

Gader, N., Lux-Pogodalla, V., and Polguère, A. (2012). Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor. *Proceedings of the Third Workshop on Cognitive Aspects of the Lexicon (CogALex III)*, The COLING 2012 Organizing Committee, Mumbai, pp. 109–125.

Gader, N., Ollinger, S., and Polguère, A. (2014). One Lexicon, Two Structures: So What Gives? In: H. Orav, Ch. Fellbaum, P. Vossen, eds 2014. *Proceedings of the Seventh Global Wordnet Conference (GWC2014)*. Tartu (Estonie), Global WordNet Association, pp. 163-171.

Granger, S. and Lefer, M.-A. (2012). Towards more and better phrasal entries in bilingual dictionaries. *Proceedings of the XV EURALEX Congress.* Available through European Association for Lexicography website <http://www.euralex.org/proceedings-toc/euralex_2012/> [Accessed 12 February 2015]

Gross, G. (1996). *Les expressions figées en français : noms composés et autres locutions*. Paris: Gap/Ophrys.

Lux-Pogodalla, V. and Polguère, A. (2011). Construction of a French Lexical Network. Methodological Issues. *Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011. An ESSLLI 2011 Workshop*, Ljubljana, pp. 54–61.

Mejri, S. (2013). Unité en sciences du langage et collocations. *Cahiers de lexicologie*, v. 1, n° 102.

Mel'čuk, I. (2008). Phraséologie dans la langue et dans le dictionnaire. *Repères & Applications VI*, *Actes des XXIV*^{e} *Journées Pédagogiques sur l'Enseignement du Français en Espagne*, Barcelone, 3–5 septembre 2007, pp. 187–200.

Mel'čuk, I. (2013). Tous ce que nous voulions savoir sur les phrasèmes, mais... *Cahiers de Lexicologie*, 102, 1, pp. 129-149.

Osherson, A. and Fellbaum, Ch. (2010). The Representation of Idioms in WordNet. *Proceedings of Global WordNet Conference 2002, CFILT, IIT*, Bombay, Mumbai, 2010.

Petit, G. (2004). La polysémie des séquences polylexicales. *Syntaxe et sémantique*, 1, n° 5, pp. 91-114.

Polguère, A. (2009). Lexical systems: graph models of natural language lexicons. *Language Ressources and Evaluation*, 43, pp. 41-55.

Polguère, A. (2014). From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, 27/4, pp. 396-418.

Sikora, D. and Polguère, A. (2013). Comment tisser son réseau lexical dans l'apprentissage d'une langue seconde ? Communication au colloque «Phraséo 2013» – Premières Rencontres Phraséologiques, Grenoble, Université de Grenoble 3 – Stendhal, 13-15/11/2013.

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-making*. Cambridge, New York: Cambridge University Press.

Vaguer, C. (2010). *Être aux anges*, *sortir de ses gonds*... Comment les langues traduisent-elles des états émotionnels ? *Cahiers Sens public*, 1, n° 13-14, pp. 253-269.

TEST MODEL FOR RICH SEMANTIC GRAPH REPRESENTATION FOR HINDI TEXT USING ABSTRACTIVE METHOD

Manjula Subramaniam
Mumbai University
manjula_p5@yahoo.co.in

Vipul Dalal
Mumbai University
vipul.dalal@vit.edu.in

In this paper we present a method for summarizing Hindi Text document by creating semantic graph of original document and identifying substructures of graph that can extract meaningful sentences for generating a document summary. This paper contributes the idea to summarize Hindi text document using abstractive method. We extract a set of features from each sentence that helps identify its importance in the document. It uses Hindi WordNet to identify appropriate position of word for checking SOV (Subject-Object-Verb) qualification. Therefore to optimize the summary, we find similarity among the sentences and merge the sentence which represented using Rich Semantic Sub graph which in turn produces a summarized text document.

References

Ibrahim F. Moawad, Mostafa Aref (2012). “Semantic Graph Reduction Approach for Abstractive Text Summarization” IEEE

M. Aref, I. Moawad, S. Ibrahim (2010). "Rich Semantic Graph Generation System Prototype", The tenth Conference on Language Engineering,Cairo, Egypt.

Chetana Thaokar, Dr.Latesh Malik (2013). “Test Model for Summarizing Hindi Text using Extraction Method” Proceedings of 2013 IEEE Conference on Information and Communication Technologies.

J. Leskovec, M. Grobelnik, N. Milic-Frayling (2004). "Learning Sub-structures of Document Semantic Graphs for Document Summarization", in KDD2004 Workshop on Link Analysis.

J. Leskovec, M. Grobelnik, N. Milic-Frayling (2000) "Learning Semantic Graph Mapping for Document Summarization".

Kedar Bellare, Anish Das Sarma, Atish Das Sarma, Navneet Loiwal,Vaibhav Mehta, Ganesh Ramakrishnan, Pushpak Bhattacharyya “Generic Text Summarization using WordNet”.

I. Moawad, M. Aref, S. Ibrahim (2011). "Ontology-based Model for Generating Text Semantic Representation", the International Journal of Intelligent Computing and Information Sciences “IJICIS”, Vol. 11, No. 1, pp. 117-128, January.

D. Evans, K. McKeon, J. Klavans (2005). "Similarity-based Multilingual Multi-Document Summarization", Technical Report CUCS-014-05,Department of Computer Science, Columbia University.

A. Stergos, K. Vangelis, S. Panagiotis (2005). "Summarization from medical documents: a survey", Artificial intelligence in medicine, Vol. 33, No. 2,pp. 157-77.

Hovy, E. H. (2005). “Automated Text Summarization”. Ox-ford Handbook of Computational Linguistics, pages 583-598.Oxford University Press.

Dehkordi, P.K., Khosravi, H., Kumarci (2009). “Text Summarization Based on Genetic programming”. International Journal of Computing and ICT Research, 3(1), 57-64.

C. Fellbaum (1998). "WordNet: An Electronic Lexical Database", MIT Press.

G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller (1990). Five Papers on WordNet. Cognitive Science Laboratory, Princeton University, Princeton.

INTEGRATING VERB+NOUN COLLOCATIONS INTO A FRENCH - ROMANIAN LEXICAL ALIGNMENT SYSTEM FOR LAW DOMAIN

Amalia Todirascu
LiLPa, Université de Strasbourg
todiras@unistra.fr

Mirabela Navlea
LiLPa, Université de Strasbourg
mirabela_abe@yahoo.com

In this paper, we present several methods of completing alignments of MWEs into a French - Romanian lexical alignment system. We also present the evaluation of these methods and we discuss the MWE alignment errors. We focus on a specific class of MWEs, Verb+Noun collocations.

MWEs represent a major part of translation errors in Statistical Machine Translation (SMT) (Ramisch et al. (2013; Kordoni and Simova, 2014). They might be classified by their fixedness (idiomatic expressions are fixed : kick the bucket, prendre la fuite 'run away') or by their non-compositionality (jeter l'éponge 'throw the towel').

SMT systems use lexically aligned parallel corpora. The lexical alignment is a crucial step for SMT systems and generally fails to build MWE alignments due to their syntactic variability and to their non-compositional sense. Lexical alignments are completed by specific algorithms (Melamed, 1997), by applying external bilingual dictionaries (Okita et al. (2013), or by using specific heuristic rules to complete chunk alignment (Tufis et al. (2006). Other methods identify

MWE candidates in monolingual corpus (Ren et al. (2009; Ramisch et al. (2013) and find mappings between candidates using parallel data.

We propose to identify MWEs in our lexical alignment system by two approaches:

- the use of an external dictionary (Todirascu et al. (2008), containing 250 Verb+Noun collocations. This dictionary also contains a rich description of the morphological and syntactic properties of MWEs;
- the use of a MWE extractor (Todirascu et al. (2009), implementing a hybrid method (statistical extraction and linguistic patterns to filter the candidates) to detect them before alignment. In this method, we identify candidates in each monolingual corpus and we find some possible mappings using simple lexical alignments.

In both cases, we complete the list of alignments by adding MWE multiple links, using linguistic information (POS tags, lemmas).

For our experiments, we use law French - Romanian parallel corpora, extracted from DGT-TM (Steinberger et al. (2012), which are tagged, lemmatized and sentence-aligned. We first apply GIZA++, the most popular statistical aligner, on a 1 000 pairs of parallel sentences in order to build the baseline alignment system. Next, we apply the two methods to identify the MWEs (dictionary-based and MWEs extractor methods) and we complete the baseline alignments via our alignment algorithm, in both cases. Then, we evaluate the results against a reference corpus (Navlea, 2014) which is manually aligned and also contains Verb+Noun multiple alignments. The evaluation measure is done by computing the AER score (Och and Ney, 2003). Finally, we discuss the MWE alignment errors provided by our alignment algorithm.

References

KORDONI, V. ANDSIMOVA, I. (2014). Multiword Expressions in Machine Translation. In *Proceedings of the International Conference on Language Resources and Evaluation,* Reykjavik, Iceland: ELRA, pp. 1208-1211.

MELAMED D. I. (1997). Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing,* RI, USA: Providence, pp. 97-108.

NAVLEA, M. (2014). *La traduction automatique statistique factorisée : une application à la paire de langues français - roumain.* Ph.D.Thesis,

Université de Strasbourg, Strasbourg, 374 pages.

OCH, F. J., AND NEY, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Journal of Computational Linguistics*, 29(1), pp. 19-51.

OKITA, T, GUERRA, A. M., GRAHAM, Y.,AND WAY, A. (2010). Multi-Word Expression-Sensitive Word Alignment. In *Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010*. Beijing, China, pp. 26–34.

RAMISCH, C., BESACIER, L., AND KOBZAR, A. (2013). How hard is it to automatically translate phrasal verbs from English to French?. In J. Monti, R. Mitkov, G.. Corpas Pastor, V. Seretan (Eds.), *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology,* , Nice (France), pp. 53-61.

REN, Z, LÜ, CAO, J., LIU, Q, AND HUANG, Y. (2009). Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions. In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009,* pp. 47-54.

STEINBERGER, R., EISELE, A., KLOCEK, S., PILOS, S. AND SCHLÜTER, P. (2012). DGT-TM: A freely Available Translation Memory in 22 Languages. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012),* . Istanbul (Turquie): ELRA, pp. 454-459.

TODIRASCU, A. (HEID, U. (STEFANESCU, D. (TUFIS, D. (GLEDHILL, C. (WELLER M. AND ROUSSELOT F. (2008. Vers un dictionnaire de collocations multilingue. *Cahiers de Linguistique*, 33(1), pp. 171- 185.

TODIRASCU A., GLEDHILL C. AND STEFANESCU D. (2009). Extracting Collocations in Contexts. In Z. Vetulani, H. Uszkoreit (Eds.), *Responding to Information Society Challenges: New Advances in Human Language Technologies, LNAI 5603* . Berlin Heidelberg: Springer-Verlag, pp. 336-349.

TUFIŞ, D., ION, R., CEAUŞU, A. AND ŞTEFĂNESCU, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. *In* Ishida, T. (Fussell, S. R. (Vossen T. J. M. (eds). *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)*, avril 2006, Trento (Italie). Stroudsburg (USA, PA):Association for Computational Linguistics, 2006, pp. 153-160.

Computer-aided and/or corpus-based analysis of phraseological units Análisis de unidades fraseológicas basado en corpus o asistido por ordenador

IDIOMS IN SPOKEN CORPUS. A SAMPLE OF CZECH DATA

František Čermák
Charles University in Prague
Frantisek.Cermak@ff.cuni.cz

Marie Kopřivová
Charles University in Prague
Marie.Koprivova@ff.cuni.cz

It is a generally accepted view that at least some idioms (phrasemes) occur in the spoken communication primarily. Though this old idea needs a corroboration on extensive data, spoken corpora that are being built up in many languages seem to be an ideal resource for them. The situation of the Czech language appears to be favourable in that there exists a large multi-volume dictionary of idioms while a series of spoken corpora is now being built, too, some of them being already available.

Even a cursory look into a volume of the Czech idiom dictionary containing sentence-type idioms offers an interesting part in that many of its very first records start in a "iambic" manner, namely with an unstressed component or word (usually a particle) *a.* There are some 27 such idioms including such as

A hele!, A vida!, A co/což/jak potom..., A co?/A co!, A co by ne?, A proč ne?, A co teď?, A což tohle/ tamhle/takhle?, A co ty?, A co víc, A hrome/hergot/sakra!, A jéje/ jéjej/jejej!, A jo/jó! etc.

All of these (and other) have been used for searching the Prague Spoken Corpus; the corpus is based on spoken unscripted and unprepared conversation between equal partners. It has been found out that these idioms are, as a rule, very short (see the examples above), and are often made up of grammar (synsemantic) words. Functionally and pragmatically, these are heavily loaded in the sense that they express various types of often emotional

reactions. These idioms belonging to propositional idioms, i.e. they form a sentence in each case, have a feature that has not been explored in phraseology much, namely a specific intonation, which is recorded in the dictionary, too.

The corpus exploration of the idioms, based on a sample of the idiom dictionary will be concerned, next to finding their frequency and verification of form, in which they occur, with inspection of their use. That will focus on types of reactions of people using them (positive or negative) and a specification of this use, including a discrimination of polysemy which may occur here, too. For example *A co?* (literally "And what?") means both (a) a question and encouragement to finish the description that has been started and (b) a threat combined with a warning (usually against continuing of one´s speech). Obviously, each meaning is strictly distinguished by a particular intonation type. It seems that, in so far, as the corpora used allow this, there is a tendency for some of these idiomatic reactions to be used in specified types of spoken text, such as, for example, formalized texts, etc.

References

ČERMÁK, F. (2006). Mluvené korpusy. In: F. Čermák, R. Blatná, ed. *Studie z korpusové lingvistiky 1: Korpusová lingvistika, stav a modelové přístupy.* Praha: NLN. pp 53-67.

ČERMÁK, F. ET AL. (2009). *Slovník české frazeologie a idiomatiky: Výrazy větné.* Praha: Leda.

HNÁTKOVÁ, M. (2002). Značkování frazémů a idiomů v ČNK s pomocí SČFI. *SaS,* 63(2).

HNÁTKOVÁ, M. AND KOPŘIVOVÁ, M. (2013). Identifikacion of idioms in Spoken Corpora. In: *Proceedings of the Seventh International Conference Slovko.* Bratislava: Slovenská akadémia vied. pp 92-99.

KOPŘIVOVÁ, M. (2008). Frazeologie v mluvených korpusech na základě PMK. In: M. Kopřivová, M. Waclawičová, ed. *Čeština v mluveném korpusu.* Praha: NLN. pp 149-160.

PHRASÉOLOGIE ET TRADUCTION: PERSPECTIVE CONTRASTIVE À BASE D'UN CORPUS BILINGUE FRANÇAIS-ARABE TUNISIEN

Abdellatif Chekir
Institut Supérieur des Langues de Nabeul
Tunisie. TIL
chquirlotfi@yahoo.fr

L'interférence linguistique entre le français et l'arabe tunisien est perceptible à travers l'emprunt de termes mais également à travers le transfert des expressions idiomatiques qui sont censées être spécifiques à chaque langue. En effet, le locuteur tunisien recourt fréquemment à la langue française pour s'exprimer soit par le recours au code switching, soit par la traduction littérale d'expressions figées ou de collocations. Dans ce cas le locuteur use du calque pour transférer le sens opaque des phrasèmes et des semi-phrasèmes. Cependant, ce type de calque se distingue de celui qui est utilisé en arabe littéral car dans cette langue on transfère les expressions avec leur sens particulier dans la langue de départ en puisant les mots dans la langue arabe. En arabe tunisien le procédé de transfert est un peu différent puisqu'il permet d'effectuer une double opération : on emprunte et le sens de l'expression dans L1 et certains mots français. C'est ce qu'on peut observer à travers des exemples comme :

Tourner la veste qlib il fista

Faire monter la barre trop haut tallaʕ lbarra lfuq

Serrer le vis jikbis lvis

Serrer la ceinture jikbis essintu:ra

Ces exemples montrent que l'arabe dialectal retient certains mots français dans cet acte de transfert toutefois ces mots sont souvent adaptés au système phonétique et morphologique de l'arabe tunisien car le [v] de *veste* devient [f] en arabe tunisien alors que le [ə] dans les noms veste, barre, ceinture devient [a] en arabe tunisien qui est la marque du féminin dans cet idiome. Cette adaptation est une forme d'adoption de l'expression dans la langue d'arrivée.

Cette opération de traduction peut produire en arabe tunisien des expressions fidèles qui restituent la structure sémantico-syntaxique des expressions de départ comme elle peut générer des expressions qui s'écartent du modèle d'origine. En effet, certaines expressions traduites en arabe tunisien sont infidèles par rapport au moule d'origine. Une expression comme *prendre une veste* a pour équivalent ðа kabbut littéralement prendre un manteau. *Crever l'abcès* est traduit par fqa iddimma:la littéralement crever le bouton. Ces expressions ne constituent que quelques exemples mais ils sont emblématiques des problèmes de transfert de L1 vers L2.

Notre travail est le fruit du dépouillement d'un corpus de deux mille expressions traduites du français vers l'arabe dialectal et relevées à travers les échanges verbaux entre les locuteurs tunisiens. Il consiste à relever, à travers un regard croisé sur le français et l'arabe tunisien la fidélité et l'écart par rapport au moule de départ et les motifs qui dictent ces écarts et plus particulièrement les exigences du système linguistique de l'arabe tunisien.

References

BACCOUCHE T. (1994), *L'emprunt en arabe*, Beït Al-Hikma-Carthage, IBLV-Université Tunis I.

GROSS G. (1996), *Les expressions figées en français : noms composés et locutions*, Ophrys.

HOBEIKA-CHAKROUN F. (2010), «Les collocations arabes intensives N+Adj dans deux romans Les Filles de Ryad et l'Immeuble Yacoubian », *Revue Interdisciplinaire "Textes & contextes,* n° 5

MEJRI S. (1997). *Le figement lexical. Descriptions linguistiques et structuration sémantique*, Publications de la Faculté des Lettres de la Manouba.

MEJRI S., GROSS G., BACCOUCHE G. & CLAS A., (dir) (2003). *Traduire la langue, Traduire la culture,* Sud Editions, Maisonneuve & Larose, Tunis, Paris.

MEJRI S. (2004). «L'idiomaticité, problématique théorique » in MEJRI S., (dir), *L'espace euro-méditerranéen : Une idiomaticité partagée,* Tunis, Ceres, p.231-243.
NEVEU F. (2004), *Dictionnaire des Sciences du langage*, Armand Colin, Paris.
TUTIN A., GROSSMANN (2002). « Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif, *Revue française de linguistique appliquée* Vol. VII p. 7- 25

THE LEXICO-PHRASEOLOGY OF THE AND A/AN IN SPOKEN ENGLISH: A CORPUS-BASED STUDY

Stephen James Coffey

Università di Pisa

stephen.james.coffey@unipi.it

The English articles (*the*, *a*, *an*) are normally described in terms of the grammar of the language. This is only natural, since they are extremely frequent, fit into certain well-defined syntactic slots, and usually help to communicate only very broad aspects of textual meaning.

However, as John Sinclair has pointed out (1999, pp.160-161), the articles are also found as components of many lexico-phraseological units, and in such cases a normal grammatical description may not be of relevance. An example he gives is the presence of *a* in the phrase *come to a head*, where '*a* has little more status than that of a letter of the alphabet' (p.161). Sinclair also makes the observation that, 'I do not know of an estimate of the proportion of instances of *a*, for example, that are not a realisation of the choice of article but of the realisation of part of a multi-word expression.' (p.161).

The present paper addresses the questions raised by Sinclair, and does so with reference to both the definite and the indefinite article. It focuses, in particular, on the spoken language, and presents the results of analyses of random samples of the articles in the spoken component of the British National Corpus (hereafter BNC-spkn). According to the data in Leech et al (2001, p.144), *the* is the most frequent word in BNC-spkn and *a* is the sixth most

frequent (a rank position which remains unaltered when the frequencies of *a* and *an* are combined). Using the BNCweb interface, and specifying that the relevant word forms should be 'articles', the total numbers of tokens are: *an* 19,049; *a* 200,004; *the* 409,060. Since the numbers are very high, the samples investigated also contained a reasonably large number of tokens (500). The relative samples corresponded to the following proportions of tokens in BNC-spkn: *an* 2.62%, *a* 0.25%, *the* 0.12%. The latter two are very low percentages, and for this reason, three separate samples of each were investigated, in order to see the extent to which the samples differed.

Analyis of article usage was carried out in the first instance by reading right-sorted concordance lines. Whenever doubts arose, larger contexts were retrieved from the corpus. Various reference works were also consulted, including Berry (1993), Francis et al (1998), and various corpus-based dictionaries and grammars.

The data presented will include: description of the various types of lexico-phraseological unit found; the proportions of the samples judged to involve the different lexico-phraseological phenomena identified; the problems encountered when deciding whether or not phraseology is an important factor in specific instances of article usage; and the number of tokens in each sample which were in some way irrelevant, for example because they involved speaker repetition of the article, or the non-completion of a noun phrase.

References

BERRY, R. (1993). *Collins Cobuild English Guides 3: Articles*. London: HarperCollins Publishers.

FRANCIS, G. HUNSTON, S. AND MANNING, E. (1998). *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins Publishers.

LEECH, G. RAYSON, P. AND WILSON, A. (2001). *Word Frequencies in Written and Spoken English, based on the British National Corpus*. Harlow: Pearson Education.

SINCLAIR, J. (1999). A Way with Common Words. In: H. Hasselgård and S. Oksefjell eds. *Out of Corpora: Studies in Honour of Stig Johansson*. Amsterdam: Rodopi. pp.157-179.

A CORPUS-BASED STUDY FOR EXPLORING ADJECTIVE-NOUN COMBINATIONS IN THE ADVENTURE TOURISM IN SPANISH AND ENGLISH

Isabel Durán Muñoz
Universidad de Málaga
iduran@uma.es

This paper explores adjective-noun combinations in the adventure tourism and it proposes a method of studying the frequent word combinations of English and Spanish by automatically extracting and comparing the adjectives and adjective+noun combination in both languages. The purposes of this study are mainly two: firstly, using Spanish and English comparable corpora in the domain of adventure tourism, the most frequent adjectives and adjective+noun combinations are extracted from the Spanish-English comparable corpora by means of a novel tool described herein and, secondly, a comparison analysis is carried out to shed some light on the differences and similarities of the most frequent adjectives, recurrent adjective+noun combinations and, eventually, possible collocations in the domain.

To do so, the first part of the paper briefly discusses some general characteristics of the adventure tourism segment, mainly focused on the use of adjectives and positive language, and highlights some examples in both languages. Adventure tourism is one of the most growing segments in tourism at the moment, since more and more people are becoming involved in sport, nature and sustainability, and they are continuously looking for active holidays

instead of looking for more traditional holidays. Although passive tourist activities related to this traditional tourism still occupy a relevant position in the global tourism economy, several types of alternative tourism, such as the one we are concerned with about in this paper (i.e. the adventure tourism), are gaining popularity among tourists. In this vein, the study of this type of tourism from a linguistic viewpoint will be of great importance for translators and terminologists, since they both will improve their results. On the one hand, translators will gain knowledge on frequent word combinations and collocates as well as typical formulaic language used in this specialised domain; and on the other hand, terminologists will be able to enrich their databases and dictionaries with frequent combinations taken from real texts in this domain. The second part of the paper thoroughly describes the specialised comparable corpora employed for the experiments as well as the process followed to automatically POS tagging it, and it depicts the novel web-based tool developed in the framework of this paper. This tool aims to detect and extract adjectives and adjective+noun combinations from POS tagged specialised comparable corpora. Finally, a comparison study of the results in both English and Spanish and the categorisation of possible domain collocations is analysed. As concluding remarks, the similarities and differences in the use of adjectives between Spanish and English in the adventure tourism will be stated and future work in this line will be proposed.

References

COLSON, J. P. (2008). Cross-linguistic phraseological studies: An overview. In Granger, Sylviane and Fanny Meunier (eds.), *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.

STUBBS, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics,* 7(2), 215-244, [online] Available at: <http://www.corpus4u.org/forum/upload/forum/2005072404143981.pdf>

STUBBS, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

ELABORATION OF A NEW SCORE: LOG-R FOR CHARACTERIZING THE TYPES OF COLLOCATIONS COMPARISON WITH MUTUAL INFORMATION

Itsuko Fujimura

Nagoya University, Japan

fujimura@nagoya-u.jp

Nobushige Aoki

Gunma University, Japan

aoki@si.gunma-u.ac.jp

The more or less usual combination of words, called “collocations”, are not yet well defined, though it is recognized that they play an important role in the linguistic activity of the speakers of any language. The two principal characteristics often mentioned are the frequency of occurrence and the degree of association between their components (Ellis, 2012; Gries, 2012; Wray, 2012). While the frequency is simple and clear, the measurement of the degree of association is controversial (François & Manguin 2006). Mutual Information (MI) is one of the most frequently used measurements (Church & Hank, 1990; Ellis, 2012).

Inspired by the model for the characterization of the types of collocations by Wray (2012), the current study proposes a new score: Log-r as a measure for the degree of association of the components of bigrams. We are to affirm the effectiveness of Log-r compared to the MI, by showing examples of 1.1 million English bigrams taken from corpora of 1.1 billion words, and of 0.4 million French bigrams taken from corpora of 0.1 billion words.

The Log-r is the decimal logarithm of the coefficient of correlation of Pearson r applied to the bigrams. We employ the approximate expression by supposing a distribution of Poisson. The Log-r formula is shown below along with MI.

$$\text{Log-r} = \log_{10} \frac{f_{xy}}{\sqrt{f_x f_y}} \qquad \text{IM} = \log_2 \frac{f_{xy} N}{f_x f_y}$$

The Log-r is solid and clear, which measures only the strength of association between two words of the bigram, while the MI measures the frequency of bigram and the strength of association between the two components of a bigram at the same time. The two figures below illustrate the distributions of the 1.1 million English bigrams using Log-r and MI, showing that the Log-r represents linguistic reality more clearly. Figure 1 is the scatter chart made up of the Log-r and of the Log( ) (=frequency of the bigram xy). Figure 2 is the chart made up of the MI and the Log( ). The comparison of the positions of the bigrams in the two figures leads us to affirm that the diagram with the MI is the result of the deformation of that with the Log-r. In Figure 1, “jai alai” and “Hong Kong”, for example, have the same degree on the axis y. In Figure 2, howerver, “Hong Kong” stays lower on the axis y compared to “jai alai” and close to “gender gap”. The top side of the triangle is downward-sloping in Figure 2 (MI), as the frequency increases. This is because of the characteristics of the formula of the MI. This is not the case in Figure 1 (Log-r).

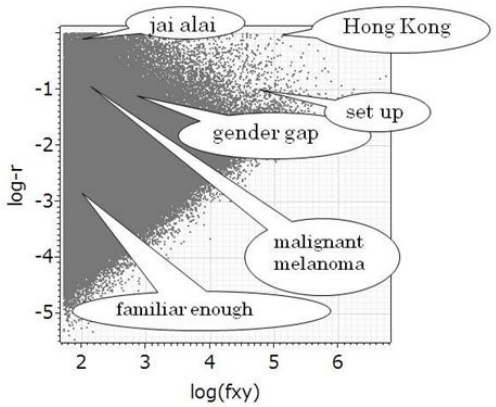

Fig 1: Log-r et Log(fxy)

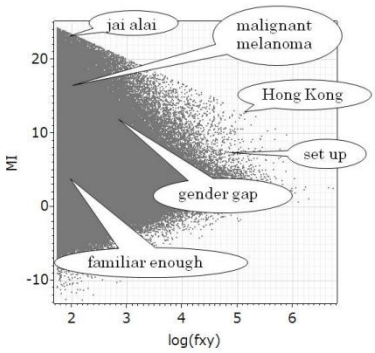

Fig 2: IM et Log(fxy)

Although the MI is a practical tool for searching rare and strongly associated collocations, it does not seem to be appropriate to describe phraseological phenomena of a language in general. In the presentation, we will show that the Log-r is more useful for characterizing the types of collocations with some more examples

References

CHURCH, K. & HANKS, P. (1990). Word Association Norms, Mutual Information and Lexicography, *Computational Linguistics*, 16(1), pp.22-29.

ELLIS, N.C. (2012). Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual Review of Applied Linguistics*, 32, pp.17-44.

FRANÇOIS, J. & MANGUIN, J.-L. (2006). *Dispute théologique*, *discussion oiseuse* et *conversation téléphonique*: Les collocations adjectivo-nominales au cœur du débat, *Langue Française*., 150, pp.50-66.

GRIES, S.Th. (2012). Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language,* 36(3), pp.477-510.

WRAY, A. (2012). What Do We (Think We) Know About Formulaic Language? An Evaluation of the Current State of Play, *Annual Review of Applied Linguistics*, 32, pp.231-254.

FRASEOLOGÍA ESPECIALIZADA, VARIACIÓN Y TRADUCCIÓN ECONÓMICA. ANÁLISIS BASADO EN CORPUS

Daniel Gallego Hernández

Universidad de Alicante

daniel.gallego@ua.es

En este trabajo presentamos un estudio comparativo basado en corpus en torno a la variación de la fraseología especializada en el ámbito de la economía. En primer lugar, repasamos someramente el complejo concepto de *unidad fraseológica especializada* y su variación (Gouadec, 1994; L'Homme, 1997; Lorente Casafont, 2001; Aguado de Cea, 2007, entre otros). A continuación, tras definir el tipo de unidades con el que trabajamos, presentamos un estudio de casos en que analizamos la variación de tales unidades a partir de un corpus genérico comparable de cuentas anuales o estados financieros en español peninsular y en español de México. En concreto, el corpus tiene 40 estados financieros correspondientes a ejercicios económicos de 2012 y 2013 (20 archivos de sendas sociedades que cotizan en el IBEX español y 20 archivos de sendas sociedades que hacen lo propio en la BMV de México) y suma un total de 1.350.000 palabras distribuidas de manera equitativa en ambas variedades lingüísticas. Llevamos a cabo el análisis de la variación tanto en el interior de cada variedad diatópica como entre ambas variedades lingüísticas. Para ello, ideamos una estrategia de extracción de

fraseología especializada relativa a una serie de eventos en el ámbito de la contabilidad. En concreto, nos centramos en las expresiones utilizadas por los contables para determinar el coste o el valor con el que se registran en los libros contables los activos, así como en las expresiones que introducen un determinado tipo de amortización o depreciación al que se somete el valor de tales bienes. Los resultados muestran la existencia de numerosas variantes con diferentes frecuencias de uso en ambas variedades, lo que lleva a pensar, por ejemplo, en la necesidad de elaborar productos fraseográficos adecuados a las necesidades de los traductores de cuentas anuales que trabajan hacia el español, así como en las dificultades de encontrar expresiones que, en cierta medida, permitan, si cabe, *neutralizar* cualquier tipo de variación diatópica. En cualquier caso, conviene seguir describiendo la variación que afecta a la fraseología especializada no solo estudiando más casos, sino también abriendo el marco de actuación a otras variedades diatópicas del español, así como a otros campos de especialización económica, pues se trata de un campo que, como muchos autores ya han señalado (Rynne, 2001; Houbert, 2001; Fuertes Olivera *et al.,* 2002; entre otros), es muy propenso a la variación no solo fraseológica, tal como vemos en el presente trabajo, sino también terminológica.

References

AGUADO DE CEA, G. (2007). La fraseología en las lenguas especializadas. In: E. Alcaraz Varó and J. Mateo Martínez and F. Yus Ramos, eds. 2007. *Las lenguas profesionales y académicas*. Madrid: Ariel, pp. 53-65.

FUERTES OLIVERA, P., ARRIBAS BAÑO, A., VELASCO SACRISTÁN, M. AND SAMANIEGO FERNÁNDEZ, E. (2002). La variación y la metáfora terminológicas en el dominio de la economía. *Atlantis,* 24(1), pp. 109-128.

GOUADEC, D., 1994. Nature et traitement des entités phraséologiques. In: D. Gouadec, ed. (1994). *Terminologie et phraséologie. Acteurs et aménageurs.* Paris: La maison du dictionnaire, pp. 165-193.

HOUBERT, F. (2001). Problématique de la traduction économique et financière. *Translation Journal,* [e-journal] 5(2). Available through: <http://accurapid.com/journal/> [Accessed 27 February 2015].

L'HOMME, M. (1997). Méthode d'accès informatisé aux combinaisons lexicales en langue technique. *Meta,* 42(1), pp. 15-23.

LORENTE CASAFONT, M. (2001). Terminología y fraseología especializada: del léxico a la sintaxis. In: M. Pérez Lagos and G. Guerrero Ramos, eds. 2001. *Panorama actual de la terminología.* Málaga: Comares, pp. 159-180.

RYNNE, J. (2001). Approaching the Translation of Spanish Financial Statements. *The ATA Chronicle,* 30(6), pp. 33-36.

LAS LOCUCIONES VERBALES EN EL ESPAÑOL DE MÉXICO

Enrique A. González Álvarez

Universidad Nacional Autónoma de México

eglezal@yahoo.com

El trabajo hace referencia a las condiciones que se requieren para que un grupo de palabras pueda ser considerado unidad fraseológica y, concretamente, locución; esto con el fin de mostrar lo que en México se ha hecho para profundizar en el estudio de este tema, tratando de unificar la terminología fraseológica que en diversos autores todavía es variada al referirse a los mismos fenómenos lingüísticos.

Después de una breve introducción se abordarán las condiciones, a saber: expresión formada por varias palabras, frecuencia de uso y de coaparición, estabilidad, unidad de significado y lexicalización, idiomaticidad, variación y, por último, se hablará del contorno que se usa en la construcción técnica lexicográfica. Se estudian los aspectos particulares que hacen que una locución logre la fijación total o parcial, como parte de un proceso o de la libertad que tiene el hablante.

Posteriormente se realizará un análisis sintáctico de un corpus de 2437 locuciones verbales utilizadas en México, resaltando las variaciones que se dan en el uso de estas expresiones. Para este análisis dividiré las locuciones en tres grupos: locuciones verbales simples, locuciones verbales pronominales y locuciones verbales negativas. Se menciona brevemente el uso de los verbos soporte. En este apartado se van analizando los tipos de construcciones sintácticas que se utilizan para la elaboración de las locuciones verbales, se analizan las semejanzas y las diferencias, los elementos sintácticos más

recurrentes y los menos utilizados, se da una primera respuesta al por qué se considera que el hablante prefiere un tipo de construcción sobre los demás.

En este apartado además se ven las preferencias del uso, o no, de los pronombres y de la negación como parte integrante de la locución, cuándo y por qué se utiliza una negación y los pocos casos de las locuciones que pueden utilizarse en forma positiva y negativa.

En un tercer momento se hace un análisis semántico partiendo de hiperónimos que engloban el sentido de las locuciones verbales más utilizadas en México. Se resaltan los semas que producen la sinonimia y los semas diferenciadores que pueden hacer que una locución tenga un sentido o matiz diferente de acuerdo al uso y al contexto de uso, por lo que puede compartir un sentido genérico con otra locución pero también tiene un sentido específico (*v. g. lloverle a uno en su milpita* vs *irle a alguien como en feria*). Así como también se analizan las locuciones que tienen el mismo significado (*v. g. estirar la pata* vs *irse alguien al otro barrio*).

Finalmente, se presentan algunas conclusiones del análisis realizado y se abren algunas propuestas de investigación que aún quedan por explorar en el español hablado en México.

References

ACADEMIA MEXICANA DE LA LENGUA (2000). *Índice de mexicanismos.* México: Fondo de Cultura Económica.

ALBAIGES, J. y HIPÓLITO, M. D. (2010). *400 frases que uno dice a menudo y no sabe por qué*. Madrid: Ediciones Martínez Roca.

ALEXANDROVA, S. (1987). *English Syntax (Collocation, Colligation and Discourse)*. Moscú: Universidad de Moscú.

BARKER, G. AND SORHUS, H. (1975). *The Importance of Fixs Expressiones in Oral Spontaneity*. Otawa: Pubish Service Commision.

CABRÉ, M.T., ESTOPA, R., y LORENTE, M. (2008). "Terminología y fraseología", en www.riterm.net/actes/5simposio/cabre5.htm

CARDERO G. A. (2011). *Estudios de neología y terminología en México. Formas creativas de dibujar la realidad del español*. México: Facultad de Estudios Superiores Acatlán, UNAM.

CEIA, C. (2005). "Fraseología", en: www.fcsh.unl.pt/edtl/verbetes/F/fraseologia.htm

CIFUENTES, J. L. (2003). *Locuciones prepositivas. Sobre la gramaticalización preposicional en español.* Alicante: Universidad de Alicante.

CORPAS, G. (1996). *Manual de fraseología*. Madrid: Gredos.

----------, (2003). *Diez años de investigación en fraseología: Análisis sintáctico-semánticos, contrastivos y traductológicos.* Madrid: Iberoamericana.
DIAMANTE COLADO, G. (2008). “Fraseología del español”, en www.mec.es/redete/biblioteca/diamante.shtml
FORMENT FERNÁNDEZ, M. (2008). “Del aprendizaje memorístico al agrupamiento de los repertorios de funciones comunicativas” en www.ucm.es/info/especulo/numero10/did_fras.html
GARCÍA MEDALL, J. (2006). *Fraseología e ironía. Descripción y contraste*. Lugo, España: Axac.
GARCIA-PAGE SÁNCHEZ, M. (2008). *Introducción a la fraseología española. Estudio de las locuciones.* Barcelona: Anthropos.
GLÄSER, R. (1986). *Phraseologie der englischen Sprache.* Tubinga: Max Niemeyer.
LORENTE, M. (2001). “Terminología y fraseología especializada del léxico a la sintaxis” en www.iulaterm.com
LUQUE TORO, L. (2010). *Manual práctico de usos de la fraseología española actual.* Madrid: Verbum.
MARTÍNEZ LÓPEZ, J. A. (1996). *La fraseología del español. Acercamiento morfosintáctico, semántico y pragmático.* Tesis doctoral. Granada: Facultad de Filosofía y Letras, Universidad de Granada.
PENADÉS MARTÍNEZ, I. (2012). *Gramática y semántica de las locuciones*. Alcalá de Henares: Universidad de Alcalá de Henares.
RAE. 2009. *Nueva Gramática de la Lengua Española* Madrid: Espasa.
SEGURA MUNGUÍA, S. (2000). *Lexicogénesis. Derivados y compuestos en la creación del vocabulario latino y castellano.* Bilbao: Universidad de Deusto.
TIMOFEEVA, L. (2012). *El significado fraseológico. En torno a un modelo explicativo y aplicado*. Madrid: Liceus.
VINOGRADOV, V. V. (1947). *Ob osnovnij Tipaj frasologicheskij iedinits v ruskom yasike*. Moscú.
WOTJAK, G. (1998). *Estudios de fraseología y fraseografía del español actual*. Madrid: Iberoamericana.

UNIDADES DISCURSIVAS CON CARÁCTER FRASEOLÓGICO: SU FUNCIÓN EN LOS DISCURSOS DE ÁLVARO URIBE VÉLEZ

Henry Hernández Bayter
Université d'Artois
henry.hernandez.bayter@gmail.com

Nos proponemos describir, analizar y caracterizar un cierto número de secuencias empleadas por el ex-presidente Álvaro Uribe Vélez durante sus intervenciones en los Consejos Comunales de Gobierno, CCG, durante sus dos mandatos, entre 2002 y 2010. Es bien sabido que el discurso político corresponde a un género discursivo propicio al empleo de fórmulas (concepto empleado por A. Krieg-Planque, 2009). Éstas hacen parte importante de la estructura del discurso de los locutores políticos. Nuestro objetivo principal es, con la ayuda de un programa lexicométrico, Lexico 3, resaltar la presencia de ciertas secuencias, formulas o asociaciones de palabras en los discurso pronunciados por el ex-presidente colombiano durante los CCG. Lo que nos interesa, en particular, es describir y analizar estas unidades desde un punto de vista discursivo con la ayuda de diferentes métodos lexicométricos propuestos por el programa Lexico 3.

Nuestra investigación se centra en unidades discursivas de un corpus de textos, que forman una serie cronológica (A. Salem, 2003). El corpus está compuesto por 277 discursos pronunciados durante los CCG en Colombia por el ex-presidente A. Uribe Vélez entre el mes de agosto de 2002 y el mes de

julio de 2010. Los CCG corresponden a un dispositivo de comunicación innovador propuesto por el ex-presidente para la constitución del Plan Nacional de Desarrollo de la Nación y para la creación del Estado Comunitario. Se trata de un corpus de gran talla y que corresponde a una característica particular de los corpus empleados para los estudios lexicométricos. Lo que nos interesa entonces es el empleo de las secuencias con carácter fraseológico en estos discursos y su función dentro de un dispositivo de comunicación política de un género innovador.

Esta investigación busca poner en relación las investigaciones del campo de la fraseología y las del campo del análisis del discurso político. Abordamos la fraseología en un sentido amplio del término como el conjunto de hechos lingüísticos y pragmáticos que conciernen ciertas unidades polilexicales, que contienen un cierto grado de fijación a nivel estructural, semántico, pero también a nivel del uso en un contexto dado, en el discurso. Cabe señalar que nos interesamos en unidades que son recurrentes en un cierto tipo discursivo, pero también nos interesamos en unidades que toman un significado particular en un contexto discursivo particular.

References

AMOSSY, RUTH (1999). *Images de soi dans le discours. La construction de l'ethos*, textes réunis et présentés par Ruth Amossy. Lausanne : Delachaux et Niestlé, collection Sciences des discours.

BALLY, CHARLES (1957 [1909]). *Traité de stylistique française*, Volume I, troisième édition. Paris : Librairie C. Klincksieck, rue de Lille 11.

CHARAUDEAU, PATRICK (2007). De l'argumentation entre les visées d'influence de la situation de communication. *Argumentation, manipulation, persuasion*. Paris: L'Harmattan.

_____ (2005). *Le discours politique. Les masques du pouvoir*. Paris: Vuibert.

CORPAS PASTOR, GLORIA (1996). *Manual de fraseología española.* Madrid: Gredos Biblioteca Románica Hispánica.

FERNÁNDEZ LAGUNILLA, MARINA (1999). *La lengua en la comunicación política tomo I et II: El discurso del poder*. *La palabra del poder*. Madrid: Arco Libros.

GONZÁLEZ-REY, MARÍA ISABEL (2002). *La phraséologie du français, Linguistique et didactique*. Toulouse: Presses universitaires du Mirail.

GRECIANO, GERTRUDE (2000). Phraséologie, ses co(n)textes et ses contrastes. *Paremia*, 9: 91-102. Madrid. Disponible en: http://www.paremia.org/paremia/P9-11.pdf

GROSS, GASTON (1996). *Les expressions figées en français. Noms composés et autres locutions.* Paris : Éditions Ophrys.

KRIEG-PLANQUE, ALICE (2009). *La notion de « formule » en analyse du discours. Cadre théorique et méthodologique*. Besançon: Presse Universitaires de Franche-Comté.

LEBART LUDOVIC ET SALEM ANDRE (1994). *Statistique textuelle,* [En ligne]. Paris: Éditions DUNOD.

MAINGUENEAU, DOMINIQUE (1991). *L'analyse du discours. Introduction aux lectures de l'archive.* Paris: Hachette.

PÉCAULT, DANIEL (2013). *La experiencia de la violencia: los desafíos del relato y la memoria*. Medellín: La Carreta Editores E.U.

PECMAN, MOJCA (2004). P*hraséologie contrastive anglais-français: Analyse et traitement en vue de l'aide à la rédaction scientifique*. Th: Ling.: Nice, Université de Nice-Sophia Anitpolis. Directeur de thèse : Henri Zinglé.

PÉREZ GUEVARA, NADIA JIMENA (2009). *El sistema de partidos colombianos hoy la pervivencia y persistencia de la personalización política*. Instituto de Iberoamérica. Universidad de Salamanca.

REY, ALAIN (1977). *Le lexique: images et modèles du dictionnaire à la lexicologie.* Paris: Libraire Armand Colin: 188-200.

SALEM ANDRE (1998) Approches du temps lexical, statistique textuelle et séries chronologiques. *Revue Mots*, vol. 17 : 105-143.

_____ (2003) *Lexico 3. Outils de statistique textuelle. Manuel d'utilisation,* SYLED – CLA^{2}T, Université de la Sorbonne Nouvelle-Paris 3.

KORPUSBASIERTE INTRA- UND INTERLINGUALE KOLLOKATIONEN

Zita Hollós

Károli Gáspár University of the Reformed Church in Hungary

hollos.zita@kre.hu

Der geplante Konferenzbeitrag stellt den Begriff der intralingualen Kollokation vor und analysiert exemplarisch intralinguale Kolloktionen – gewonnen aus KOLLEX, dem *Korpusbasierten Wörterbuch der Kollokationen* – unter dem Aspekt der Kontrastivität im phraseologischen Bereich. Dazu ist es unabdingbar, den integrativen, lernerlexikographisch orientierten Kollokationsbegriff und das korpus- und datenbankbasierte *Deutsch-ungarische KOLLokationsLEXikon* (Hollós 2014b) kurz zu charakterisieren. Nach der Vorstellung dieses neuen Wörterbuchtyps wird an verschiedenen ausgewählten Aspekten der Wörterbuchkonzeption – wie die äußere und innere Selektion und das Datenangebot – vorgeführt, wie die Korpusbasiertheit umgesetzt wurde.

Bei der Untersuchung der Kollokationen wird die Aufmerksamkeit einem stiefmütterlich behandeltem Kollokationstyp gewidmet, nämlich Kollokationen mit Adverbien. Sie entsprechen von den allgemein bekannten Strukturtypen den folgenden zwei: ADV+VERB oder ADV+ADJ. Während bei meinen letzten Untersuchungen der Kollokationen das Augenmerk zunächst dem intralingualen Aspekt, z.B. Syntax/Morphosyntax (Hollós 2014a) oder dem interlingualen Aspekt wie der Interferenz (Hollós 2013) galt, richtet sich die Aufmerksamkeit diesmal mit Hilfe der korpusermittelten Daten von KOLLEX auf beide Aspekte. Es ist ein Versuch, in die Untersuchung eines komplexen phraseologischen

Gegenstandes intra- und interlinguale Aspekte zu integrieren, um dadurch neue Erkenntnisse über dieses Phänomen, genauer über diesen Kollokationsyp zu gewinnen. Außerdem sollte eine effektive Methode für die integrative Untersuchung der Kollokationen erarbeitet werden. Integrativ soll heißen, dass das Phänomen mit einer Methodenkombination aus der Kollokationsforschung (Strukturtypen), Korpuslinguistik (statistische Signifikanz), kontrastive Linguistik (Interferenz) und Syntax (Valenz) näher eingegrenzt wird. Vorrangiges Ziel ist es, den intralingualen und den interlingualen Kollokationsbegriff i.S. von Hollós (2004) in Bezug auf einen Strukturtyp ADV+VERB weiter zu spezifizieren. Im Vortrag werden durch konkrete Wörterbuchartikel aus KOLLEX gezeigt, wie vielfältig dieses Phänomen ist und man kann anhand intra- und interlingualer Kollokationen zu Verblemmata auch Einblicke in die Schwierigkeiten der Äquivalentfindung gewinnen.

Im letzten Teil des Beitrags wird gezeigt, wie Interferenz anhand von konkreten Beispielen, den sogenannten Interferenzkandidaten im KOLLEX berücksichtigt wurde. Dazu ist die kurze Vorstellung der Interferenz auf mehreren sprachlichen Ebenen, vor allem auf der morphosyntaktischen, semantischen und stilistischen Ebene unumgänglich (vgl. u.a. Forgács 2007, Heine 2006, Réder 2006). Zum Schluss werden intralinguale ADV+VERB-Kollokationen der Lemmastrecke G und H ausgewählt, auf ihren Status als interlinguale Kollokation geprüft und eine mögliche, auf der Datenbank von KOLLEX basierende, automatisch erstellte lexikographische Präsentation vorgestellt. Mit dieser exemplarischen Bestandsaufnahme soll ein Schritt in Richtung einer phraseologisch orientierten Systematik im Bereich der Kollokationen und Interferenz getan werden, mit der Zielsetzung, den didaktisch und lexikographisch geprägten Kollokationsbegriff weiter zu präzisieren.

Zum Schluss werden Möglichkeiten der Onlinestellung, der Weiterentwicklung und der Erstellung weiterer (Teil-)Wörterbücher von KOLLEX diskutiert.

References

Forgács 2007. Erzsébet Forgács: Kontrastive Sprachbetrachtung. Szeged 2007.

Heine 2006. Antje Heine: Ansätze zur Darstellung nicht- und schwach idiomatischer verbonominaler Wortverbindungen in der zweisprachigen

(Lerner)-Lexikografie Deutsch-Finnisch (Beschreibung eines Forschungsvorhabens). In: Linguistik online 27, 2/06. Internetseite: http://www.linguistik-online.de/27_06/heine.html [am 27.03.15].

Hollós, Zita 2004. Lernerlexikographie: syntagmatisch. Konzeption für ein deutsch-ungarisches Lernerwörterbuch. Tübingen 2004. (Lexicographica. Series Maior 116).

Hollós, Zita 2013. Interferenzkandidaten in zweisprachigen Lernerwörterbüchern, insbesondere im deutsch-ungarischen Kollokationslexikon KOLLEX. In: Lexicographica 29. Tübingen 2013, 92-116.

Hollós, Zita 2014a. Syntagmatik im KOLLEX: Die lexikographische Darstellung grammatischer Syntagmatik in einem zweisprachigen Kollokationslexikon für Deutschlerner. In: Zweisprachige Lexikographie zwischen Translation und Didaktik. Hrsg. von María José Domínguez Vázquez, Fabio Mollica und Martina Nied Curcio. Berlin/Boston 2014, 113-129. (Lexikographica. Series Maior 147).

Hollós, Zita 2014b. SZÓKAPTÁR: Német–magyar SZÓkapcsolatTÁR. Korpuszalapú kollokációs tanulószótár. KOLLEX: deutsch-ungarisches KOLLokationsLEXikon. Korpusbasiertes Wörterbuch der Kollokationen. Deutsch als Fremdsprache. Szeged 2014.

Steyer 2002. Kathrin Steyer: Wenn der Schwanz mit dem Hund wedelt. Zum linguistischen Erklärungspotenzial der korpusbasierten Kookkurrenzanalyse. In: Haß-Zumkehr, Ulrike / Kallmeyer, Werner / Zifonun, Gisela (Hgg.): Ansichten zur deutschen Sprache. Festschrift für Gerhard Stickel zum 65. Geburtstag. Tübingen: Narr 2002, 215-236. (Studien zur deutschen Sprache 25).

Réder 2006. Anna Réder: Kollokationen in der Wortschatzarbeit. Wien 2006.

MIT BEDACHT: KORPUSLINGUISTISCHE UNTERSUCHUNGEN ZU STRUKTUREN *[PRÄPOSITION + SUBSTANTIV]* MIT ADVERBIALER FUNKTION

Herbert J. Holzinger
Universitat de València
herbert.holzinger@uv.es

Im Rahmen der Forschungsgruppe FRASESPAL werden unter Projekt Nr. FFI2013-45769-P Strukturen des Typs *[Präposition + Substantiv]* des heutigen Deutsch analysiert und mit dem Spanischen kontrastiert.

Als Methodologie dient dabei eine Korpusauswertung. Für das Deutsche wird das Deutsche Referenzkorpus (DeReKo) des Instituts für Deutsche Sprache verwendet, für das Spanische das *Corpus del Español* (M. Davies) und das CREA. Anhand dieser Korpora werden Untersuchungen zu Frequenz und Kookkurrenz vorgenommen, um die so genannten usuellen Wortverbindungen herauszufiltern. Unter „usuellen Wortverbindungen“ versteht man rekurrente, polylexikalische, habitualisierte sprachliche Zeichen (Steyer 2013: 23), die nicht nur kombinatorischen Restriktionen unterworfen sind, sondern auch einen semantisch-pragmatischen Mehrwert aufweisen.

Neben dem Grammatikalisierungsgrad der Präposition (Kiss (2008); Müller u.a.) und dem Lexikalisierungsgrad der Wortverbindung werden zudem die

innere und äußere Festigkeit sowie der semantisch-pragmatische Mehrwert der Verbindung untersucht.

Zur Untersuchung des inneren Festigkeitsgrades dient der Einschubtest, mit dem überprüft werden kann, ob zwischen Präposition und Substantiv weitere lexikalische Einheiten eingeschoben werden können und wenn ja, ob sich diese lexikalischen Füller semantisch gruppieren lassen. Bei *mit Bedacht* sind als lexikalische Füller z.B. intensivierende Adjektive zu beobachten: *mit viel/mehr/äußerstem/... Bedacht*. Die Tatsache, dass derartige Erweiterungen möglich sind, ist deshalb bemerkenswert, weil *Bedacht* den Charakter eines unikalen Elements (Holzinger 2012, 2013) oder phraseologisch gebundenen Wortes hat, d.h. nur innerhalb eines Phrasems gebraucht werden kann und nicht frei kombinierbar ist. Der Stabilitätsgrad dieser phraseologischen Einheiten variiert stark und erreicht in bestimmten Fällen 100%, wie etwa bei *im Handumdrehen* und *mit Windeseile* (Stumpf 2014).

Unter der „äußeren Festigkeit“ verstehen wir die bevorzugte Kombinatorik mit bestimmten Lexemen, wie z.B. *mit Bedacht wählen/auswählen/aussuchen*. Als zusätzlicher Schritt soll festgestellt werden, ob sich in den Erweiterungen Muster erkennen lassen, die in semantisch-pragmatische Gruppen eingeordnet werden können.

Die Kontrastierung mit dem Spanischen geht vom prototypischen Gebrauch der deutschen Vorkommen aus und versucht, Entsprechungen zu finden. Bei dieser Äquivalenzsuche ist zu überprüfen, ob die Verbindung in den einschlägigen zweisprachigen Wörterbüchern überhaupt enthalten ist, und wenn ja, ob die angebotenen Äquivalente bzw. Übersetzungen zutreffend und ausreichend sind, oder ob sie durch aus der Korpusanalyse gewonnene neue Möglichkeiten ergänzt werden müssen.

Die in den Schlussfolgerungen dargebotenen Ergebnisse können sowohl der ein- als auch der zweisprachigen Lexikografie neue Impulse geben und gliedern sich in das Gesamtkonzept der Forschungsgruppe ein, deren Ziel darin besteht, eine umfassende lexikografische Darstellung dieser Kombinationen in einer Online-Plattform zur Verfügung zu stellen.

References

CREA. Corpus de referencia del español actual. [online] Available at: <http://corpus.rae.es/creanet.html> [Accessed 30 March 2015].

DAVIES, M. Corpus del español. [online] Available at: < http://www.corpusdelespanol.org/> [Accessed 30 March 2015].

DEUTSCHES REFERENZKORPUS (DeReKo). [online] Available at: <https://cosmas2.ids-mannheim.de/cosmas2-web/> [Accessed 30 March 2015].

HILPERT, M.: *The German mit-predicative construction*. [online] Available at: <http://members.unine.ch/martin.hilpert/GMPC.pdf> [Accessed 30 March 2015].

HILPERT, M. (2014): *Collostructional analysis. Measuring associations between constructions and lexical elements*. In: D. Glynn & J. Robinson (eds.) *Polysemy and Synonymy. Corpus Methods and Applications in Cognitive Linguistics.* Amsterdam: John Benjamins. pp. 391-404.

HOLZINGER, H.J. (2013). Unikale Elemente: Eine Herausforderung für Lexikologie und Lexikografie. *Aussiger Beiträge* 7, pp. 53-66.

HOLZINGER, H.J. (2012). Unikale Elemente: Apuntamentos sobre as palabras ligadas fraseoloxicamente do alemán actual. *Cadernos de Fraseoloxia galega* 14, pp. 165-173. [online] Available at: <http://www.cirp.es/pub/docs/cfg/cfg14.pdf> [Accessed 30 March 2015].

KISS, T. (2008). Towards a Grammar of Preposition-Noun Combinations. In: St. Müller (ed.): *Proceedings of the HPSG08 Conference NICT, Keihanna, Japan*. CSLI Publications. [online] Available at: <http://csli-publications.stanford.edu/HPSG/2008>, pp. 116–130. [Accessed 30 March 2015].

MÜLLER, A., ROCH, C., STADTFELD, T. AND KISS T. *The annotation of preposition senses in German*. [online] Available at: <http://www.linguistics.ruhr-uni-bochum.de/~kiss/publications/LingEvidence_neu.pdf > [Accessed 30 March 2015].

STEYER, K. (2013). *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Tübingen: Narr.

STUMPF, S. (2014). *Mit Fug und Recht*. Korpusbasierte Erkenntnisse zu phraseologisch gebundenen Formativen. *Sprachwissenschaft* 39, 1, pp. 85-114.

ASPECTOS CONCEPTUALES Y CULTURALES DE ALGUNOS FRASEOLOGISMOS DEL KAMAIURÁ

Enrique Huelva Unternbäumen

Universidad de Brasilia

enriquehuelva@gmail.com

En este trabajo presentamos el análisis de algunos fraseologismos del kamaiurá, lengua indígena brasileña, perteneciente al grupo étnico-lingüístico tupi-guaraní del Alto Xingú. Los casos analizados revelan una alta complejidad semántica resultante, especialmente, de la integración conceptual (blending) de la metáfora REALIZAR UNA ACCIÓN ES TRANSMITIR UNA PARTE DEL CUERPO y de la metonimia PARTE DEL CUERPO POR LA ACCIÓN para conceptuar acciones físicas y perceptuales/cognitivas. A diferencia de muchas lenguas europeas, la combinación de estos dos elementos conceptuales es utilizada en Kamaiurá tanto para la conceptuación de acciones abstractas (‘yo te mando mi ojo’ = yo me acuerdo de ti) como de acciones concretas (‘yo te mando mi brazo’ = yo te abrazo). Los datos presentados muestran además la gran importancia cultural que posee en kamaiurá el cuerpo humano como dominio fuente para la configuración de la estructura conceptual codificada por unidades fraseológicas.

References

HUELVA UNTERNBÄUMEN, E. (2015).From primary metaphors to the complex semantic pole of grammatical constructions. Language and Cognition, v. 7, p. 68-97.

HUELVA UNTERNBAUMEN, E. (2012). Cómo hacer palabras con cosas haciendo cosas con palabras: acerca de la experienciación y conceptualización de actos de habla. Language Design, v. 13, p. 29-72.

LONGEST-COMMONEST MATCH

Milos Jakubicek

Sketch Engine. Lexical Computing Ltd.

milos.jakubicek@sketchengine.co.uk

The prospects for automatically identifying two-word multiwords in corpora have been explored in depth, and there are now well-established methods in widespread use. (We use 'multiwords' as a cover-all term to include collocations, colligations, idioms, set phrases etc.) But many multiwords are of more than two words and research into methods for finding items of three and more words has been less successful.

We present an algorithm for identifying candidate multiwords of more than two words called longest-commonest match. We start from a two-word collocation, as identified using well-established techniques (dependency-parsing, followed by finding high-salience pairs of lexical arguments to a dependency relation.) We then explore whether a high proportion of this data is accounted for by a particular string.

We currently use a 'one quarter' (n/4) threshold, and a minimum frequency of 5 hits for l-c matches. These were set on the basis of informal reviewing of output. If we can find a more objective way of setting the thresholds, we shall of course do so (and we plan to revise the minimum-frequency threshold so it varies with corpus size).

An earlier version of the longest-commonest algorithm was already presented in Kilgarriff et al (2012). We (re-)present the work because it was only covered very briefly in the earlier presentation, and in the meantime we have developed a version of the algorithm that works very fast even for multi-

billion word corpora, and is fully integrated into our corpus query system, the Sketch Engine.

Acknowledgement

This work has been partly supported by the Ministry of Education of the Czech Republic within the LINDAT-Clarin project LM2010013 and by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

References

Church, K. W., Hanks, P. (1989). Word association norms, mutual information, and lexicography. Proc 27th ACL, Vancouver, Canada. Pp. 76–83.

Daudaravičius, V., Marcinkevičienė, R. (2004). Gravity counts for the boundaries of collocations. Int Jnl of Corpus Linguistics 9(2) pp. 321–348.

Dias, G. (2003). Multiword unit hybrid extraction. Proc. ACL workshop on Multiword expressions: analysis, acquisition and treatment. Pp 41–48.

Evert, S., Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. Proc 39th ACL, Toulouse, France. Pp. 188–195.

Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. Proc. EURALEX. pp. 105–116.

Kilgarriff, A., Rychlý, P., Kovář, V., Baisa, V. (2012). Finding multiword of more than two words. Proc. EURALEX. Oslo, Norway.

Kilgarriff, A., Rychlý, P., Jakubicek. M., Kovář, V., Baisa, V. and Kocincová, L. (2014). Extrinsic Corpus Evaluation with a Collocation Dictionary Task. Proc LREC, Rejkyavik, Iceland.

Petrovic, S., Snajder, J., Basic, B.D. (2010). Extending lexical association measures for collocation extraction. Computer Speech & Language 24(2) pp. 383–394.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. Proc. RASLAN workshop, Brno, Czech Republic.

Wermter, J., Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction. Proc. 44th ACL, Sydney, Australia. Pp. 785–792.

FRASEOLOGUISMOS DE COLOR EN ESPAÑOL Y RUSO: ESTUDIO DE FRASEOLOGISMOS SIN ANÁLOGOS EN LA OTRA LENGUA

Anastasia Kovaleva

Universidad de Málaga

akovaleva@uma.es

El estudio del aspecto contrastivo de la fraseología siempre ha provocado cierto interés en los investigadores dada la gran variedad de vías de investigación que se presentan en este ámbito. En los últimos años, dentro de este enfoque, se destaca la tendencia a aplicar los avances de la lingüística de corpus con el fin de investigar sobre las similitudes y las diferencias entre fraseologismos de lenguas distantes, basándose en una comparación del funcionamiento de los fraseologismos, desde el punto de vista pragmático-discursivo, realizada a través de la aplicación de técnicas y herramientas de PLN (Procesamiento de Lenguaje Natural), con especial referencia al uso de corpus. Nuestro trabajo se enmarca, precisamente, en estos nuevos enfoques de los estudios fraseológicos basados en corpus. Siguiendo estas tendencias, el presente trabajo se centra en un estudio de fraseologismos acromáticos (unidades fraseológicas con blanco o negro entre sus componentes) y fraseologismos cromáticos (unidades fraseológicas con colores básicos entre sus componentes) en el par de lenguas español-ruso. Se definen y presentan fraseologismos de color y se estudia el grupo de los fraseologismos que disponen de equivalentes fraseológicos que no denotan colores en la otra

lengua. Se trata de los fraseologismos del tipo: *черный монах* (“monje negro”), *белая горячка* (resultado del síndrome de abstinencia del alcohol), *увидеть белый свет в клеточку* (“ver la luz blanca en cuadraditos”), *зеленый змей* (“serpiente verde”) en ruso y *quedarse en blanco* y *mangas verdes* en español. Frecuentemente estas unidades presentan dificultades para traducción ya que reflejan realidades u objetos propios de una de las dos culturas.

El estudio se realiza de acuerdo a los principios y técnicas de investigación propias de la fraseología computacional, con especial referencia a la metodología de corpus. La base para el análisis de los fraseologismos acromáticos y cromáticos la constituyen los estudios de G. Corpas Pastor dedicados a esta rama de investigación fraseológica. Los datos se obtienen mediante la consulta a corpus y sistemas en línea que proporcionan más información que los diccionarios bilingües y monolíngües. Se utiliza el CREA (Corpus de referencia del español actual), el CNR (Corpus Nacional Ruso) y el corpus basado en la red Internet y gestionado por el programa WebCorp, herramienta de gestión de corpus que resulta imprescindible para analizar las combinaciones que se caracterizan por su baja frecuencia de aparición en el discurso. La aplicación de la metodología de corpus para el análisis de los resultados obtenidos ofrece una información fraseológica contrastiva relativa a los dos universos fraseológicos más completa dado que este método proporciona datos acerca de uso y el funcionamiento pragmático-discursivo de los fraseologismos de color.

References

ALEFIRENKO, N. AND SEMENENKO N. (2009). *Fraseologuija i Paremiologuija* [Фразеология и паремиология]. Moscú: Flinta, Nauka.

BARANOV, A. AND DOBROVOL’SKIJ, D. (2008). *Aspecti teorii fraseologuii* [Аспекты теории фразеологии]. Moscú: Znak.

CORPAS PASTOR, G. (1996). *Manual de fraseología española*. Biblioteca Románica Hispánica. Manuales. Madrid: Gredos.

CORPAS PASTOR, G. (2008). Traducir con corpus: los retos de un nuevo paradigma. Fráncfort: Peter Lang.

CORPAS PASTOR, G. (2013). Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. En I. Olza and E. Manera, ed. 2013. *Fraseopragmática*, Berlín: Frank & Timme. pp. 335-373.

CORPAS PASTOR, G. (2014). El fraseólogo internauta: cómo *pasarlo pipa* en la red. En J. Sevilla Muñoz, ed. 2014. *Fraseología y paremiología: enfoques y aplicaciones.* Madrid: Instituto Cervantes pp. 133-152.

KÜPPERS, H. (1971). *Fundamentos de la teoría de los colores*. Barcelona: Gustavo Gili SA.
SECO, M., RAMOS, G. AND ANDRES, O. (1999). *Diccionario del español actual*. Madrid: Aguilar.
SECO, M., RAMOS, G. AND ANDRES, O. (2004). *Diccionario fraseológico documentado del español actual*. Madrid: Aguilar.
SZAŁEK, J. (2005). Los colores y su semántica en las expresiones fraseológicas españolas. In: Studia Romanica Posnanienska. 32. pp. 87-96.
TELIA, V. (1996). *Russkaja fraseologuija* [Русская фразеология]. Moscú: Shkola “Yazyki russkoj kultury”.
TUROVER, G. AND NOGUEIRA, J. (2000). *Gran Diccionario Ruso-español* [Большой русско-испанский словарь]. Madrid: Rubiños-1860.
VARELA, F. AND KUBARTH, H. (1994). *Diccionario fraseológico del español moderno*. Madrid: Gredos.

N-GRAMS IN MULTILINGUAL CORPORA: EXTRACTING AND ANALYZING LEXICAL BUNDLES IN CONTRASTIVE STUDIES

Marie-Aude Lefer

Marie Haps School of Translation and Interpreting, Brussels

marie-aude.lefer@ilmh.be

Natalia Grabar

Université de Lille 3

natalia.grabar@univ-lille3.fr

This presentation introduces a new extraction method for the analysis of lexical bundles in corpus-based contrastive studies and reports on preliminary findings for the English-French language pair. Lexical bundles are "sequences of word forms that commonly go together in natural discourse" (Biber et al. 1999: 990ff). These recurrent sequences of words, which are often semantically compositional and not cognitively salient, include referential expressions (e.g. *in the European Union*, *weapons of mass destruction*), discourse organizers (e.g. *and that is why*, *when it comes to*) and stance expressions (e.g. *it is very important that*, *it seems to me that*) (see e.g. Biber et al. 2004). Bundles can be extracted from corpora by relying on the n-gram method (i.e. the automatic extraction of sequences of *n* contiguous words). Most monolingual studies so far have focused on 4-word sequences. However, as pointed out by Ebeling and Ebeling (2013) and Granger (2014), equivalent bundles across languages can differ in length (e.g. *you will see that* vs. *vous verrez que*). Limiting the analysis to one bundle length may therefore jeopardize the cross-linguistic comparability of the data.

In the present paper, we aim at tackling this issue by devising a method that automatically extracts and groups bundles into 'bundle families', i.e. sets of trigrams plus longer n-grams containing them (e.g. *on the other*: *on the other side*, *on the other side of*, *on the other side of the*, *on the other hand*, *on the other hand the*, *on the other hand there*). Interestingly, the method also makes it possible to group bundle variants (e.g. *de même pour*: *il en est de même pour* vs. *il en va de même pour*) and phrase-frames, i.e. recurrent n-grams "with one variable lexical slot" (Stubbs 2007: 166) (e.g. *la mise en*: *la mise en place/oeuvre/concurrence de*; *it would be*: *it would be wrong/irresponsible/interesting/helpful to*).

The extraction method is used in an illustrative case study devoted to English and French. The study relies on comparable data extracted from four corpora, representing four genres: Europarl (transcripts of EU parliamentary debates; Koehn 2005, Cartoni and Meyer 2012), KIAP (research articles in medicine, economics and linguistics; Fløttum et al. 2006), Multed (editorials) and PLECI (news). Together they total 7+ million tokens (ca. 900,000 tokens per genre and per language). The study focuses on bundles that are shared across genres (i.e. found in at least three of the four genres investigated), as they are more likely to reflect the typical phraseological features of the two language systems under scrutiny, and relies on rather low frequency thresholds (min. 5 occurrences per genre for the trigrams and min. 2 occurrences per genre for the accompanying, longer n-grams). This corresponds to 3,251 and 1,600 bundle families in French and English, respectively, with varying numbers of family members.

In our presentation we will assess the potential of this new extraction method by means of a detailed contrastive analysis of the genre-shared bundles in English and French (in terms of length, function and structure). More generally, we will show the benefits of the frequency-driven approach to contrastive phraseology.

References

BIBER, D., CONRAD, S. AND CORTES, V. (2004). *If you look at* … Lexical Bundles in University Lectures and Textbooks. *Applied Linguistics*, 25, 371-405.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S. AND FINEGAN, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

CARTONI, B. AND MEYER, T. (2012). Extracting directional and comparable corpora from a multilingual corpus for translation studies. *8th International Conference on Language Resources and Evaluation (LREC)*. Available at:< https://www.idiap.ch/~tmeyer/res/Cartoni-LREC-2012.pdf>

EBELING, J. AND OKSEFJELL EBELING, S. (2013). *Patterns in Contrast*. Amsterdam & Philadelphia: John Benjamins.

FLØTTUM, K. DAHL, T. AND KINN, T. (2006). *Academic Voices – across languages and disciplines*. Amsterdam & Philadelphia: John Benjamins.

GRANGER, S. (2014). A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast*, 14(1), 48-72.

KOEHN, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT Summit X*, 79-86. Available at:<http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf>

STUBBS, M. (2007). Quantitative data on multi-word sequences in English: the case of the word *world*. In HOEY, M., MALHBERG, M., STUBBS, M. AND TEUBERT, W. (eds.). *Text, Discourse and Corpora: Theory and Analysis*. London: Continuum, 163-189.

A TIROS Y *A BALAZOS*: ANÁLISIS CONSTRUCCIONAL

Belén López Meirama
Universidade de Santiago de Compostela
belen.meirama@usc.es

La presente comunicación se enmarca en el proyecto de investigación FFI2013-45769 *Combinaciones fraseológicas del alemán de estructura [PREP. + SUST.]: patrones sintagmáticos, descripción lexicográfica y correspondencias en español*, en el cual defendemos la existencia de "construcciones fraseológicas" o, en términos de Taylor (2014), *constructional idioms*, patrones construccionales que, como indica el autor, presentan unas propiedades sintácticas, semánticas, pragmáticas e incluso fonológicas que no derivan de principios generales (2014: 11). Las "construcciones fraseológicas" se caracterizan por ser esquemas morfosintácticos recurrentes que aparecen con ciertos constituyentes fijos (en la que nos ocupa, la preposición *a*) y otros que, aunque son casillas vacías (*free slots*), sufren determinadas restricciones semántico-combinatorias (en nuestro caso, los sustantivos escuetos plurales *tiros* y *balazos*, a los que cabría sumar otros muchos, como *golpes*, *palos*, *balonazos*, *guantazos*, *empujones*, *pisotones*, *cuchilladas*, *pedradas, mordiscos*, etc.).

Tras ejecutar la orden de búsqueda "a + sustantivo" en el *Corpus del español* (Davies) para determinar las combinaciones usuales más frecuentes con este esquema, se ha puesto de manifiesto que las secuencias *a tiros* y *a balazos* se encuentran entre las de mayor frecuencia absoluta (70 y 52 ocurrencias en textos del siglo XX, respectivamente). Estas combinaciones

usuales forman parte de la construcción fraseológica [*a* + $S_{plural/acción\ violenta}$], que constituye el marco de la presente investigación.

Tal construcción es sobradamente conocida en la fraseología española (aunque generalmente, cuando se alude a su significado, suele emplearse el término 'golpe'), que se refiere a ella bien en esta forma más general o bien en la más específica de [*a* + $S_{-azo/plural}$], probablemente por la enorme productividad que el sufijo aumentativo *–azo* tiene en español –y en esta construcción en particular–, sufijo que, como indicó Monge (1972), ha desarrollado en nuestra lengua de manera muy significativa el sentido de 'acción' o, más concretamente, de «acción momentánea, fuerte e inesperada» (1972: 243).

Su frecuencia en Davies, así como el hecho de que en principio las dos realizaciones mencionadas podrían considerarse sinónimas (además de que una de ellas contiene el sufijo derivativo *–azo*), me han sugerido la conveniencia de realizar un estudio detallado de corpus, con el objetivo de lograr una descripción minuciosa de sus características semántico-pragmáticas. Para ello me serviré del procedimiento de análisis empleado en el proyecto citado, aplicando los parámetros de las que denominamos 'fijación interna' y 'fijación externa' de la construcción. De acuerdo con la primera, intentaré determinar si existe o no, y en qué grado, tendencia a la expansión, sea a través de la coordinación (como en *a patadas y mamporros*) o de la adyacencia de algún modificador (como en *a golpes de machete, a grandes mordiscos*); en relación con la fijación externa, el objetivo perseguido es detectar posibles preferencias combinatorias verbales (parece que *a balazos* se construye frecuentemente con el verbo *acribillar*); también, si hay algún valor pragmático que pueda asociarse a la construcción; etc. Para ello, además del de Davies, emplearé los dos corpus de referencia de la Real Academia Española dedicados al español contemporáneo: el CREA y el CORPES XXI.

References

CORPAS PASTOR, G. (1996). *Manual de fraseología española*. Madrid: Gredos.

DE BRUYNE, J. (1978). Acerca del sufijo *–azo* en español contemporáneo. *Iberorromania*, 8, pp. 54-81.

GARCÍA-PAGE SÁNCHEZ, M. (2007). Esquemas sintácticos de formación de locuciones adverbiales. *Moenia*, 13, pp. 121-144.

GARCÍA-PAGE SÁNCHEZ, M. (2008). *Introducción a la fraseología española: estudio de las locuciones*. Barcelona: Arthropos.

GOLDBERG, A.E. (1995). *Constructions*. *A Construction Grammar Approach to Argument Structure.* Chicago: University of Chicago Press.
MELLADO BLANCO, C. (2015) (forthcoming). Antiphrasis-based comparative constructional idioms in Spanish. *Journal of Social Sciences*, 11:1.
MONGE, F. 1972. Sufijos españoles para la designación de 'golpe'. In: *Homenaje a Francisco Yndurain*. Zaragoza: Universidad de Zaragoza. pp: 229-247.
PENADÉS MARTÍNEZ, I. (2012). *Gramática y semántica de las locuciones*. Alcalá de Henares: Universidad de Alcalá.
RUIZ GURILLO, L. (2001). *Las locuciones en español actual*. Madrid: Arco/Libros.
TAYLOR, J. R. (2014). Cognitive linguistics (for Routledge Handbook of Linguistics), draft. Available at: <https://www.academia.edu/7179973/Taylor_2014_Cognitive_linguistics_for_ Routledge_Handbook_of_Linguistics_>.
ZULUAGA, A. (1980). *Introducción al estudio de las expresiones fijas*. Frankfurt am Main: Peter D. Lang.

Corpus de referencia

REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual*. Available at: <http://www.rae.es>.
REAL ACADEMIA ESPAÑOLA: Banco de datos (CORPES XXI) [en línea]. *Corpus del español del siglo XXI.* Available at: http://www.rae.es
DAVIES, M. (2002-) *Corpus del Español: 100 million words, 1200s-1900s*. Available at: <http://www.corpusdelespanol.org>.

AN EXPLORATION OF THE PHRASEOLOGY OF A LARGE CORPUS OF ACADEMIC ENGLISH

John Anthony McKenny

British University in Dubai

john.mckenny@buid.ac.ae

This paper grows out of a corpus project which began with the compilation of the British University in Dubai (BUiD) corpus of written academic English. 600 Masters dissertations submitted by BUiD students since 2004 were collected and the corpus stands around 8 million running words. This corpus is the only sizeable corpus of tertiary English writing from the Middle East.

The most useful application of our corpus seemed to be the production of a resource for writers in the same position and context as our corpus authors. Writers of English for Academic Purposes need to develop phraseological competence, formerly referred to as collocational competence (Howarth (1998). We made available to student writers and to teachers of EAP writing in our Academic Success Unit some of our findings from the corpus. We provided lists of N-grams (8,7,6,5,4,3 grams) in descending order of frequency following an idea used in stylistics by Mahlberg (2007). We sought to show that these holophrases performed local textual functions and characterised the texts they were found in. To go further than Ngrams and lexical bundles we tried experiments in workshops with student or teacher participants to filter out word sequences which were truncated, ending e.g. in a preposition or not in any way formulaic and which were generated by the grammar of the language and at the

same time, transparent. We used the system of De Cock, Granger, Leech and McEnery (1998:75)

1. the elimination of most combinations of closed-class items that are only phrase or clause fragments; e.g. *in the, and it* and the subjective elimination of those word combinations which are fragmentary in nature; e.g. *are a lot of, don't know if you.*

2. The next phase of the filtering process is an assessment of whether the remaining candidate formulae have the potential to serve any pragmatic or discourse functions.

3. The final stage examines each occurrence of a potential formula in its co-text.

(De Cock, S., Granger, S., Leech, G. and McEnery, T. 1998)

This began a quest for relevant academic tropes related to, for example, metadiscourse, hedging and boosting (Hyland 2005). We looked at reporting verbs, such as *suggest, state, claim,* and *shows* in our corpus. By posting on the University VLE the outcome of each of these searches and the subsequent discussion a useful resource emerged. A lively debate among students about their favourite way of saying things ensued. One of the strengths of our corpus is its large size which generates sufficient examples of locutions and also whenever users wish they can switch to a smaller sub-corpus comprised only of students awarded a distinction for their dissertation. Other things being equal, it might be presumed that the quality of writing in this smaller sub-corpus will be high and the phraseology well-wrought. I would like to synthesize these phraseological explorations and bring a succinct report of the findings to EUROPHRAS2015. We would like to share our corpus with participants and would welcome suggestions for further study of our corpus.

References

DE COCK, S., GRANGER, S., LEECH, G. and MCENERY, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (ed.) *Learner English on Computer.* London: Longman. 67-79.

HOWARTH, P. (1998). Phraseology and second language proficiency. *Applied Linguistics* 19(1): 24-44.

HYLAND, K. (2005). *Metadiscourse*. London: Continuum.

MAHLBERG, M. (2007) Corpus stylistics: bridging the gap between linguistics and literary studies. 191-208. In HOEY, M. MAHLBERG, M., STUBBS, M.

and TEUBERT, W. (2007) *Text, Discourse and Corpora. Theory and Analysis.* London: Continuum.

RECURSOS FRASEOLÓGICOS DE ATENUACIÓN EN EL CORPUS PRESEEA-GRANADA

Esteban Montoro del Arco

Universidad de Granada

montoro@ugr.es

La atenuación es una estrategia pragmática de gran complejidad, que los hablantes usan para conseguir un acercamiento social, bien evitando amenazar la propia imagen personal, la del interlocutor o las de terceras personas, o bien reduciendo el compromiso ante lo dicho o hecho. Su uso en la interacción oral es bastante habitual y, como demuestran trabajos recientes (Briz 2007, Albelda & Cestero 2011, Albelda 2013, Cestero 2014) se sirve de gran variedad de recursos lingüísticos y paralingüísticos que pueden agruparse según los niveles habituales de estudio de las lenguas: fónicos (alargamientos, vacilaciones, etc.); morfológicos (uso de diminutivos, modificación temporal del verbo, etc.), sintácticos (uso de adverbios modales de pensamiento, duda o probabilidad, impersonalizaciones, etc.), semánticos (eufemismos, ironía, etc.) y pragmático-discursivos (petición de disculpas, acotación de la opinión, marcadores del discurso de consecuencia lógica, etc.).

En este trabajo presentamos en primer lugar los recursos de atenuación que pueden considerarse específicamente *fraseológicos*, pues estos no se suelen considerar y, de hacerlo, se encuentran diseminados en el resto de los niveles analizados hasta la fecha, sin constituir un grupo de fenómenos definido frente a los demás en los listados creados a tal efecto (Albelda *et al* 2014). En segundo lugar, analizamos su distribución social y para ello nos servimos del

corpus de español hablado en Granada (PRESEEA-Granada) (Moya, coord., 2007, 2008 y 2009), incluido en el proyecto coordinado PASOS (*Patrones sociolingüísticos del español de España*), e inscrito a su vez en el macroproyecto panhispánico PRESEEA. Siguiendo la metodología común de este último (Moreno Fernández 1996), la muestra oral del corpus se obtuvo mediante un muestreo por cuotas de afijación uniforme en el cual se divide el universo relativo en estratos según las tres variables sociales básicas: sexo, edad y nivel de instrucción.

References

ALBELDA, Marta (2013). "La atenuación: tipos y estrategias" en José Ramón GÓMEZ MOLINA (coord.), *El español de Valencia. Estudio sociolingüístico.* Frankfurt am Main: Peter Lang, 315-343.

ALBELDA, Marta, Antonio BRIZ, Ana M. CESTERO, Dorota KOTWICA y Cristina VILLALBA (2014). "Metodología para el análisis sociopragmático de la atenuación en corpus discursivos del español", *Oralia,* 17: 7-62.

ALBELDA, Marta, y Ana M. CESTERO (2011). "De nuevo, sobre los procedimientos de atenuación lingüística", *Español Actual* 96: 121-155.

BRIZ, Antonio (2007). "Para un análisis semántico, pragmático y sociopragmático de la cortesía atenuadora en España y América", *Lingüística Española Actual* 29,1: 5-40.

CESTERO, Ana M. (2011). "Las estrategias de atenuación: estudio sociolingüístico". *Actas del IX Congreso Internacional de Lingüística General*. Valladolid: Universidad de Valladolid, 525-542.

CESTERO, Ana M. (2014). "Estudio coordinado de la atenuación en el marco del PRESEEA: propuesta metodológica" en D. da HORA, J. LOPES RIBEIRO y R. MARQUES DE LUCENA (org.), *Estudos Linguísticos e Filológicos. ANAIS. XVII Congreso Internacional Asociación de Lingüística y Filología de América Latina*. Joao Pessoa: ADALTECH-ALFAL, 1-13.

MORENO FERNÁNDEZ, Francisco (1996). "Metodología para el 'Proyecto para el Estudio Sociolingüístico del Español de España y América (PRESEEA)", *Lingüística*, 8, 257-287.

MOYA CORRAL, Juan Antonio (coord.) (2007). *El español hablado en Granada. Corpus oral para su estudio sociolingüístico. I Nivel de estudios alto*, Granada, Editorial Universidad de Granada.

MOYA CORRAL, Juan Antonio (coord.) (2008). *El español hablado en Granada II. Corpus oral para su estudio sociolingüístico. Nivel de estudios medio*, Granada, Editorial Universidad de Granada.

MOYA CORRAL, Juan Antonio (coord.) (2009). *El español hablado en Granada III. Corpus oral para su estudio sociolingüístico. Nivel de estudios bajo*, Granada, Editorial Universidad de Granada.

VERBALE KOLLOKATIONEN: *JEDER KENNT SEINEN PLATZ. JEDER WEIß, WO SEIN PLATZ IST*

Nikoleta Olexová

Universität der Hl. Cyril und Methodius in Trnava

nika.olex@gmail.com

Wir kommunizieren nicht nur mit einzelnen Wörtern, sondern mit festen, bzw. typischen oder usuellen Wortkombinationen, die wir zur Bildung der sinnvollen Sätze verwenden. Im Mittelpunkt des vorliegenden Beitrags stehen Kollokationen, die zum Bereich der festen Wortverbindungen gehören und einen aktiven Bestandteil des Wortschatzes jeder Sprache bilden. Der Grund für die Auswahl des Themas im Hinblick auf aktuelle Situation der wissenschaftlichen Forschung der Kollokationen ist, die Ursachen und die Faktoren des kollokationellen Verhaltens der Wörter und präferierte Kombinatorik einzelner Kollokationskomponenten weiter zu erforschen. Es handelt sich um eine empirisch basierte korpusgestützte Untersuchung von ausgewählten kognitiven Verben *kennen* und *wissen* im Deutschen, mit dem Ziel, neue Erkenntnisse über ihre Semantik und Kollokabilität zu gewinnen und eventuelle Korrektur und Ergänzungen der lexikografischen Beschreibung vorzuschlagen. Es handelt sich um experimentale Forschung, die primär die Methoden und Mitteln der Korpuslinguistik ausnutzt. Die Forschung beruht auf der Annahme, dass mit Hilfe von einer detaillierten Korpusanalyse des gewonnenen empirischen Materials ein detailliertes Bild über die Kollokabilität der untersuchten kognitiven Verben gegeben werden kann. Nach der

Gewinnung der Informationen über Kollokationspotenzial der Verben können wir ihre Bedeutungsstruktur anhand von den lexikografischen Standardwerken in der Konfrontation mit gewonnenen Datenmengen aus Korpora überprüfen. Wir brauchen umfangreiche empirische Basis, um zuverlässige Ergebnisse für die Erstellung der Kollokationsprofile zu gewinnen. Die Ermittlung der Kollokabilität der Lexeme erfolgt auf Grund der Datengewinnung aus Korpora und lexikographischen Quellen. Die methodologische Ausgangsbasis der Forschung und Ermittlung der Kollokationen im Rahmen der Lexikographie stellen Erklärungswörterbüchern, Kollokationswörterbüchern und Valenzwörterbüchern als primäre Informationsquelle über das Aktatntenpotential und Rektion der untersuchten Verben. Bei den korpusbasierten Unterschungen stützen wir uns auf die Korpora und Kookkurrenzdatenbank.

Die experimentale Erforschung der Kollokabilität der verbalen Kollokationen stützt sich auf die frequenzbasierten und musterbasierten Methoden der Korpuslinguistik. Im Vordergrund stehen die mathematisch-statistischen Modellen, die die Kookkurenz und die Kollokabilität der Verben im Text ermitteln können.

Als sehr effektiv zur Unterscheidung zwischen festen und freien Wortverbindungen haben sich die kombinierten Verfahren der “7K-Methode“ und die Anwendung des Frequenzkriteriums bewährt. Am objektivsten dokumentiert die Typikalität der Kollokationen die Frequenzdistribution mit vordefinierten paradigmatischen und syntagmatischen Filtern. Die Analyse der Frequenzdistribution identifiziert Kollokationen und zeigt klar und deutlich die typische Umgebung zum untersuchten Basislexem. Das Verhalten der Kollokationen kann man nicht nur mittels der statistischen Maße, sondern auch in strukturellen Kettenmodellen erforschen. Eine manuelle Sortierung und linguistische Analyse sind trotzdem immer notwendig. Empirische Basis unserer Untersuchung stellt die Ausarbeitung der detaillierten Kollokationsprofile der ausgewählten kognitiven Verben dar.

Die gewonnen Erkenntnisse können hilfreich für die detaillierte Darstellung, Forschung und Erfassung des deutschen Wortschatzes sein. Neue Erkenntnisse aus der Kollokationsforschung stellen auch eine solide, empirisch erprobte materielle Basis für eine vertiefte kontrastive Beschreibung und sie

bilden eine neue Basis für eine detailliertere Beschreibung der morphologischen und syntaktischen Eigenschaften der Lexik und ihrer Verwendung, was natürlich auch zur Verbesserung der Muttersprache- und Fremdsprachendidaktik beitragen kann.

References

ĎURČO, P. ET AL. (2014). *Valenz Und Kookkurrenz.* Münster: Lit Verlag.

ĎURČO, P., BANÁŠOVÁ, M., HANZLÍČKOVÁ A. (2010). *Feste Wortverbindungen im Kontrast*. Trnava: Univerzita sv. Cyrila a Metoda.

BÚSQUEDA Y ANÁLISIS DE LA FRASEOLOGÍA DEL DESACUERDO EN UN CORPUS MULTIMODAL DE TELEVISIÓN

Inés Olza
Universidad de Navarra
iolzamor@unav.es

Laura Amigot Castillo
Universidad Complutense de Madrid
lamigot@filol.ucm.es

Elvira Manero Richard
Universidad de Murcia
emanero@um.es

En este trabajo nos aproximamos a la fraseología que expresa desacuerdo en español e inglés con base en los datos ofrecidos por la Biblioteca NewsScape de noticias de televisión. Alojada por la University of California, Los Angeles (UCLA)[2], la Biblioteca NewsScape constituye el mayor corpus multimodal buscable y etiquetado para el inglés y el español[3]. Este corpus permite realizar búsquedas automáticas sobre las más de 200.000 horas de programas informativos que almacena; programas que integran una gran variedad de géneros del lenguaje hablado en los ámbitos periodístico y político (informativos, entrevistas, debates, tertulias, etc.) de diversas lenguas europeas[4]. Estas más de 200.000 horas de televisión se corresponden con unos 3.000 millones de palabras en forma de subtítulos –*closed captioning*[5]–

[2] Ver http://newsscape.library.ucla.edu/.
[3] En este momento, los archivos textuales de NewsScape han sido lematizados y etiquetados morfosintácticamente, además de haber sido también procesados a través de otras herramientas de Procesamiento de Lenguaje Natural (*sentiment detection*, *named entity recognition* y FrameNet, entre otras). Ver el apartado "Current state of tagging" de la página web informativa de NewsScape: http://bit.ly/1Ny5QQX.
[4] Las lenguas mejor representadas en NewsScape son el inglés –sobre todo– y, en segundo lugar, el español, de ahí que las hayamos escogido para nuestro análisis.
[5] El subtitulado oculto o *closed caption* es el sistema de subtítulos de programas de televisión destinado a que personas con problemas de audición puedan ver por escrito lo que se

que han sido alineados con los archivos audiovisuales, de modo que en NewsScape se hace posible llevar a cabo búsquedas que llevan al momento preciso de emisión en el que una palabra o una secuencia multiverbal fueron pronunciadas.

El análisis fraseológico que llevamos a cabo con base en este recurso resulta innovador en dos aspectos fundamentales: por un lado, tenemos acceso a un corpus de intervenciones e intercambios orales *reales* sin precedentes en cuanto a su tamaño y validez ecológica, lo cual permite acceder a una vastísima cantidad de datos de uso espontáneo de unidades fraseológicas (UFS), sobre todo, de expresiones de valor pragmático, modal e interactivo; y, por otra parte, NewsScape ofrece la posibilidad, hasta ahora inexplorada, de llevar a cabo un *análisis multimodal sistemático* del empleo real de los fraseologismos.

En esta comunicación, nuestra atención se dirige hacia varias UFS utilizadas conversacionalmente en español e inglés para mostrar rechazo o desacuerdo hacia el contenido proposicional y actitudinal de un acto de habla anterior (cf. Herrero Moreno 2002; Olza 2011): nos referimos a expresiones como *de ninguna manera*, *(y) (unas/las) narices*, *(y) una leche/mierda*, *qué* + […] + *ni qué narices/leche(s)/mierda(s)*, en español; y *(like) hell*, *my eye*, *my foot*, *no way* o *nonsense*, en inglés. El examen de las concordancias proporcionadas por NewsScape para este conjunto de expresiones persigue los siguientes fines:

(a) caracterizar de modo general los *patrones prosódicos* que acompañan al empleo de este conjunto de fraseologismos;

(b) definir los *moldes formales* que adquieren más habitualmente estas estrategias de disensión, pues de modo preliminar se observa que basculan entre realizaciones fraseológicas más convencionales y estilísticamente neutras (*de ninguna manera*, *no way*) y otras crecientemente creativas (por ejemplo, *qué […] ni qué ocho cuartos*);

(c) trazar un panorama de los *géneros discursivos* en que estas UFS son proclives a aparecer;

---

retransmite oralmente. Estos subtítulos suelen constituir una transcripción –altamente fiable, aunque no perfecta– de las intervenciones e informaciones orales incluidas en los programas.

(d) por último, y muy fundamentalmente, definir los *gestos* y comportamientos no verbales[6] ejecutados de modo simultáneo a la emisión de estos fraseologismos de desacuerdo.

References

HERRERO MORENO, G. (2002). Los actos disentivos. *Verba*, 29, pp.221–242

MCNEILL, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago/London: The University of Chicago Press.

MCNEILL, D. (2013). Gesture as a window onto mind and brain, and the relationship to linguistic creativity and ontogenesis. En: C. Müller *et al.* (eds). 2013. *Body – Language – Communication*. Berlin: Mouton de Gruyter, pp.28-54.

OLZA, I. (2011). On the (meta)pragmatic value of some Spanish idioms based on terms for body parts. *Journal of Pragmatics*, 43, pp.3049-3067.

[6] Nos centraremos en lo que habitualmente se entiende como *gesto*, es decir, en los movimientos de manos –sobre todo– y otras partes del cuerpo –cabeza, ojos– que intervienen en el proceso de habla, y que habitualmente suelen poseer un carácter figurativo o imaginístico (McNeill 1992, 2013).

FÜHLEN ODER EMPFINDEN? EIN VERGLEICH DER KOOKKURRENZPROFILE DER PARTIELLEN SYNONYME

Gabriela Orsolya

Universität der Hl. Kyrill und Method in Trnava

lengabi@yahoo.com

In der Kollokationsforschung gibt es viele Betrachtungsweisen, wie der Begriff „Kollokation“ definiert wird. Kollokationen sind eine Art feste Wortverbindungen, die ihren Platz aus der Sicht des Idiomatizitätsgrades in der Peripherie des phraseologischen Bestandes haben, jedoch sie bilden einen zentralen Bereich des Wortschatzes jeder Sprache. Für einen Muttersprachler ist ihre Verwendung unterbewusst und meistens problemlos, für einen L2-Lerner können sie aber oft im Sprachgebrauch Schwierigkeiten verursachen.

Der heutige Stand der Kollokationsforschung spricht über zwei Kollokationsauffassungen. Die lexikologisch-lexikographische Auffassung sieht die festen Wortverbindungen im engeren Sinne, wobei Kriterien wie kombinatorische Möglichkeiten von lexikalischen Einheiten mit anderen Wörtern, die semantische Kompatibilität und lexikalische Kollokabilität eine wichtige Rolle spielen. Hausmann (1984: 398) beschreibt Kollokationen als Verbindungen von Wörtern mit begrenzter Kombinierbarkeit, die auf „differenzierten semantischen Regeln und einer gewissen zusätzlichen Üblichkeit“ basieren. Er spricht über „Halbfertigprodukte der Sprache“ (1984:

398) die von einem Sprecher nicht kreativ gebildet, sondern als eine Einheit aus dem Gedächtnis abgerufen werden. Die zweite, breitere Auffassung von Kollokationen ist die korpuslinguistische Auffassung, die unter Kollokationen charakteristische, häufig auftretende Wortverbindungen versteht, die aufgrund der Frequenz und statistischen Kriterien recherchiert werden. Hier spricht man über sog. Kookkurrezen, d.h. über das auffällige gemeinsame Auftreten zweier lexikalischer Einheiten.

In diesem Beitrag werden Kollokationen als sinnvolle Verbindungen von Wörtern verstanden, deren Entstehung sowohl durch die Kollokabilität (quantitatives Kriterium) als auch Kompatibilität (qualitatives Kriterium) bedingt ist. (Vgl. Ďurčo, 2010: 15).

Bei der Analyse werden mehrere linguistischen Korpora, Kookkurrenzdatenbank und online Wörterbüchern verwendet. Der Schwerpunkt der Recherche liegt in der korpusbasierten Methodik der Kollokationsforschung. Das Kollokationsprofil des Basislexems wird aufgrund von Häufigkeitsvorkommen mit mathematisch-statistischen Modellen und Methoden erstellt, die die Kookkurrenz und die Kollokabilität der Verben ermitteln können. Nach der Analyse des kollokationellen Verhaltens der untersuchten Verben wird versucht, die Bedeutungsstrukturen der Verben *„fühlen*“ und „*empfinden“* vor dem Hintergrund ihrer aktuellen lexikografischen Bearbeitung zu (re)interprätieren.

Ich gehe bei dem Vergleich von der Basis-Kollokator-Beziehung (vgl. zum Begriff Hausmann 1984, Konecny 2010) aus und versuche, die Kollokationsparadigmen von beiden Verben herauszuarbeiten. Der Vergleich von Kookkurrenzen beider Verben kann die Gemeinsamkeiten und Unterschiede in ihren Bedeutungsstrukturen exakt und empirisch erprobt zeigen. Bei meiner Untersuchung dient das Matrixmodell als Basis für die Delimitation und Erfassung der Kollokationen/Kookkurrenzen von Ďurčo (2007), die alle Kombinationen von binären Strukturen umfasst. Zum Gegenstand wird eine kontrastive Analyse der Kollokabilität von Verben *„fühlen“* und *„empfinden“* gestellt. Die Erstellung von Kollokationsmatrizes dient zur Vertiefung der Erkenntnisse über die Kollokabilität der untersuchten lexikalischen Einheiten.

References

HAUSMANN, F.J. (1984): Wortschatzlernen ist Kollokationslernen. In: Praxis des neusprachigen Unterrichts Jg. 31, 395 – 406.
ĎURČO, P./BANAŠOVÁ, M./HANZLÍČKOVÁ, A. (2010): Feste Wortverbindungen im Kontrast. Trnava: Univ. der hl.. Kyrill u. Method, Philosophische Fakultät.
KONECNY, CH. (2010): *Versuch einer semantisch-begrifflichen Annäherung und Klassifizierung anhand italienischer Beispiele*, Lang Peter Frankfurt.

DICCIONARIO FRASEOLÓGICO DEL ESPAÑOL DE MÉXICO: OBTENCIÓN DE UNIDADES FRASEOLÓGICAS EN CORPUS ORALES REGIONALES

Niktelol Palacios Cuahtecontzi

El Colegio de México

niktelolpalaciosc@gmail.com

Gabriela Vidauri González

Universidad Autónoma de Baja California

gvidauri@uabc.edu.mx

La redacción del *Proyecto para la elaboración del Diccionario fraseológico del español de México* nos ha llevado a estudiar los corpus mexicanos actuales que se han sumado, al menos metodológicamente, a la recolección de datos del *Proyecto para el Estudio Sociolingüístico del Español de España y de América* (PRESEEA).

En esta ponencia queremos centrar nuestro estudio en el reconocimiento —y propuesta del tratamiento lexicográfico— de unidades fraseológicas de la zona noroeste de México a partir de los datos del *Corpus del Habla de Tijuana* (Baja California, México). La importancia de estudiar esta región es, por una parte, su lejanía con la Ciudad de México (principal centro de irradiación de la norma lingüística mexicana) y, por la otra —más social—, que se trata de la quinta ciudad más poblada del país (según el Censo de Población y Vivienda 2010 tiene 1, 300, 983 habitantes), su cercanía con Estados Unidos de América (30.7 Km de San Diego) y su permanente flujo migratorio.

En este primer acercamiento queremos probar nuestro método cuantitativo y cualitativo para el reconocimiento de unidades fraseológicas (locuciones y

colocaciones) en corpus orales. En esta etapa de la investigación no podemos asegurar que estas unidades sean exclusivas del habla de Tijuana, pero sí que se documentan en esta variedad y que no se usan en el habla del Altiplano Central. La elección de posibles unidades fraseológicas será obtenida de acuerdo a los siguientes criterios: 1) análisis cuantitativo a partir de un programa automatizado de generación de concordancias; 2) análisis cualitativo (idiomaticidad, fijación y unidad de significado) y 3) contraste de las unidades que aparecen en este corpus con los del *Corpus sociolingüístico de la ciudad de México* y el *Corpus sociolingüístico de la ciudad de Puebla*.

Sírvannos de ejemplo, las siguientes unidades coloquiales:

1. *¡a la bestia!* ¡Caramba!: "¡A la bestia!, ¿viste cómo cayó la morra?"
2. *¿Qué traes, bato?* Esta frase se usa en dos contextos, el primero es entre amigos con el significado de ¿te pasa algo? Y el segundo es forma de reto: ¿tienes algún problema (con x situación)? “¿Qué traes, bato?, ¿por qué esa cara?”
3. *¿Cuál es tu cura?* Qué haces o en qué te ocupas. “Vamos al fut, ¿o cuál es tu cura?”

Nuestro trabajo busca identificar los criterios cuantitativos y cualitativos que permiten reconocer las unidades fraseológicas en un corpus de habla y cuáles de las unidades identificadas en el corpus del habla tijuanense deben llevar una marca de uso regional.

References

MOLINA, R., J. CRHOVA AND M.R. Domínguez (2013). El habla de Tijuana: material para el análisis de la variante regional. *Plurilinkgua*, 9(1). Available at: <http://idiomas.ens.uabc.mx/plurilinkgua/docs/v9/1/MCD.pdf> [Accessed 26 March 2015].

PRESEEA (2012). *Metodología del Proyecto para el Estudio Sociolingüístico del Español de España y de América* (PRESEEA). Versión revisado octubre 2003. Available at: <http://preseea.linguas.net/Portals/0/Metodologia/METODOLOG%C3%8DA%20PRESEEA.pdf> [Accessed 26 March 2015].

MARTÍN P. AND Y. LASTRA (2011). *Corpus sociolingüístico de la ciudad de México-Preseea*. México: El Colegio de México.

AUSBAU DES PHRASEOLOGISCHEN BILDES. EINE KORPUSBASIERTE UNTERSUCHUNG

Irina Parina

Nishny Novgorod State Linguistic University

parinai@yandex.ru

Gegenstand der vorliegenden Untersuchung sind Idiome des semantischen Feldes VERRÜCKTHEIT. Zunächst wurde anhand des Deutschen Referenzkorpus der Gebrauch von deutschen Idiomen untersucht, deren Bedeutung in (Duden, 2002) als „nicht (recht) bei Verstand sein“ umschrieben wird. Im Laufe der Untersuchung wurde festgestellt, dass einige Idiome dieses semantischen Feldes zum Ausbau des phraseologischen Bildes neigen. Das heißt, das Bild, das in einer festen Wendung verankert ist, wird im Gebrauchskontext weiter ausgebaut (vgl. Ptashnyk, 2005, S.84).

Zum Beispiel kommen in 18 Korpusbelegen mit den Phrasemen *nicht alle Tassen im Schrank haben* und *einen Sprung in der Schüssel haben* die Lexeme *Porzellan-Syndrom* oder *Porzellankrankheit* vor. Während in den meisten Belegen diese Lexeme sprachspielerisch verwendet werden, um die Bedeutung der Phraseme humorvoll zu verstärken, kommen auch vereinzelt Kontexte vor, wo sie selbständig gebraucht werden.

Die Lexeme *Porzellan-Syndrom* oder *Porzellankrankheit* sind in Online-Wörterbüchern (Duden online, 2015; Langenscheidt, 2015; PONS, 2015) nicht vertreten. Um zu überprüfen, ob diese Lexeme usuell sind, wurden Belege im Internet gesucht. Am 29.03.2015 lieferte die Suchmaschine Google für

Porzellankrankheit über 20500 Belege (wobei es in mehreren um eine bei Garnelen vorkommende Krankheit ging) und für *Porzellansyndrom* 818 Belege. Letztere waren oft Witze oder Fragmente der Diskussionen in Internet-Foren, wo die Bedeutung des Lexems besprochen wurde. Also sind diese Lexeme intersubjektiv gebräuchlich und können Neologismen genannt werden.

Außerdem wurde während der Suche nach Varianten von Phrasemen festgestellt, dass der Ausbau des Bildes auch zur Entstehung neuer Phraseologismen führen kann. So wurde zum Beispiel bei der Suche nach Varianten des Phrasems *einen Sprung in der Schüssel haben* die Korpusanfrage gekürzt auf: *einen Sprung in*. Die Anfrage lieferte 1369 Treffer. Die Belege enthielten neben dem Ausgangsphrasem und freien Wortverbindungen *einen Sprung in der Vase haben, einen Sprung in der Schale haben, einen Sprung in der Tasse haben* auch Wendungen *einen Sprung in der Marille haben, einen Sprung in der (Schall)platte haben.* Bei dem Phrasem *einen Sprung in der Marille haben* handelt es sich um ein Synonym von *einen Sprung in der Schüssel haben,* das in österreichischer Presse vorkommt. Fernere Analysen haben gezeigt, dass dieses Phrasem, das weder in (Duden, 2002) noch in (Redensarten-Index.de, 2015) vertreten ist, eine Variante hat (*einen Wurm in der Marille haben*) und, ähnlich wie *nicht alle Tassen im Schrank haben,* ein Derivat – *Wachauer Krankheit*.

Das Phrasem *einen Sprung in der (Schall)platte haben* wird in einigen Belegen auch als Synonym für *einen Sprung in der Schüssel haben* verwendet. Allerdings handelt es sich in diesen Fällen wahrscheinlich um Kontamination, weil in den meisten Belegen das Phrasem die Bedeutung „sich ständig wiederholen“ hat und es auch Kontexte gibt, in denen diese zwei Wendungen bewusst voneinander abgegrenzt werden. Die Wortverbindung fehlt in (Duden, 2002) und bei (Redensarten-Index.de, 2015). Es kann vermutet werden, dass es sich um einen Neologismus handelt.

Da die Derivation durch Ausbau des phraseologischen Bildes sich als ein produktiver Mechanismus erweist, können korpusbasierte Untersuchungen von bestimmten Basiskomponenten des Bildes bereits bekannter Phraseme ein Weg zur Auffindung phraseologischer Neologismen sein, die bekanntlich eine komplizierte Aufgabe darstellt.

References

DUDEN ONLINE (2015). [online] <http://www.duden.de/> [Letzter Zugriff: 29.3.2015].

DUDEN (2002). *Redewendungen und sprichwörtliche Redensarten / Wörterbuch der deutschen Idiomatik.* Hrsg. und bearbeitet von Drosdowski, G. / Stolze-Stubenrecht, W. Mannheim u.a.

LANGENSCHEIDT ONLINE-WÖRTERBUCH (2015). [online] <https://woerterbuch.langenscheidt.de/ssc/search/free.html> [Letzter Zugriff: 29.3.2015].

PONS ONLINE-WÖRTERBUCH (2015). [online] <http://de.pons.com/übersetzung> [Letzter Zugriff: 29.3.2015].

PTASCHNYK, S. (2005). „Unstabile“ feste Wortverbindungen: Zur Dynamik des phraseologischen Sprachbestandes. *Hermes - Journal of Linguistics*, 35. Aarhus: Aarhus School of Business, S. 77-95.

REDENSARTEN-INDEX.DE (2015). *Online-Wörterbuch für Redensarten, Redewendungen, idiomatische Ausdrücke, feste Wortverbindungen*. [online] <http://www.redensarten-index.de/suche.php> [Letzter Zugriff: 29.3.2015].

DEFINICIÓN, ANÁLISIS Y CLASIFICACIÓN DE LAS UNIDADES FRASEOLÓGICAS DESDE UNA PERSPECTIVA HISTÓRICA: LOS CORPUS DIACRÓNICOS Y SU IMPORTANCIA EN EL ESTUDIO DEL SISTEMA LOCUCIONAL PREPOSITIVO

David Porcel Bueno

Universitat de València

davidiezma@hotmail.com

En los últimos años la Fraseología histórica ha puesto en evidencia la necesidad de analizar las Unidades fraseológicas de la misma manera que analizamos un texto del pasado de la propia lengua, «en la que la sintaxis se ha vuelto rígida, con el consiguiente bloqueo de los principio de percepción y reformulación en su sentido gramatical» (Echenique 2003: 548). Así pues, la aprehensión de los mecanismos gramaticales que han ido operando históricamente hasta llegar a la situación actual ha permitido ahondar en la motivación inicial de determinadas Unidades fraseológicas y determinar, siempre en un plano gramatical, las circunstancias específicas en las que surgieron.

Hemos llevado a cabo un análisis histórico-lingüístico que nos ha permitido definir, clasificar y caracterizar un amplio conjunto de locuciones prepositivas,

teniendo en cuenta los tres estadios en los que operan los cambios lingüísticos: un estadio original anterior al cambio, una fase en la que triunfa la nueva estructura y una etapa intermedia en la que coexisten el sistema innovador y el original, siendo el paso intermedio el más interesante para determinar los factores que favorecen la aparición de las formas innovadoras (Enrique-Arias 2009: 12).

Para ello nuestro análisis se nutre de algunos de los corpus informatizados más importantes para el estudio histórico de la fraseología: *Archivo Digital de Manuscritos y Textos Españoles* (ADMYTE), *Corpus Diacrónico del Español* (CORDE), *Corpus Informatitzat del Català Antic* (CICA), *Corpus do portugués* (de Mark Davies) y *Tesouro Informatizado do Galego Medieval.*

Especial relevancia para nuestro estudio ha tenido también la cuestión de los universales lingüísticos, sobre todo teniendo en cuenta los vínculos de orden cultural que se establecen entre lenguas a lo largo de su evolución. Es por eso que la mayor parte de la documentación textual procedente de los corpus diacrónicos pertenece al siglo XIII, momento en que el castellano medieval está en pleno desarrollo y se está convirtiendo (gracias al impulso cultural alfonsí) en una lengua científica en contacto con lenguas plenamente consolidadas (árabe y latín) y con otras variedades románicas en gestación (catalán y galaico-portugués).

References

CIFUENTES HONRUBIA, J. L. (2003). *Locuciones prepositivas. Sobre la gramaticalización preposicional en español*. Alicante: Universidad de Alicante.

ECHENIQUE, M. T. (2003). Pautas para el estudio histórico de las unidades fraseológicas. In: J. L. Girón Alconchel et al., eds. 2003. *Estudios ofrecidos al profesor José Jesús de Bustos Tovar*. Madrid: UCM, vol. I, pp. 545-560.

ENRIQUE-ARIAS, A. (2009). *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid: Iberoamericana.

KABATEK, J. (2005). Las tradiciones discursivas del español medieval: historia de textos e historia de la lengua. *Iberoromania: Revista dedicada a las lenguas y literaturas iberorrománicas de Europa y América*, 62, pp. 28-43.

MARCHELLO-NIZIA, C. (2005). A NLP-driven approach to historical linguistics. In: C. D. Pusch, J. Kabatek, W. Raible, eds. 2005. *Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*. Tübingen: Gunter Narr. pp. 11-30

MONTORO DEL ARCO, E. (2006). *Teoría fraseológica de las locuciones particulares. Las locuciones prepositivas, conjuntivas ymarcadoras en español*, Frankfurt: Peter Lang.

PENADÉS, I. (2012). *Gramática y semántica de las locuciones*. Madrid: Universidad de Alcalá.

WANNER, D. (2005). The corpus as a key to diachronic explanation. In: C. D. Pusch, J. Kabatek, W. Raible, eds. 2005. *Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*. Tübingen: Gunter Narr. pp.

VICENTE LLAVATA, S. (2011). *Estudio de las locuciones en la obra literaria de don Íñigo López de Mendoza (Marqués de Santillana). Hacia una fraseología histórica del español*. València: Universitat de València.

ZULUAGA, A. (1975). La fijación fraseológica. *Thesaurus* 30(2), pp. 225-248.

ATEMBERAUBEND, BREATHTAKING, DECHBEROUCÍ: A WORD OR LEXICAL PHRASEME? CORPUS-BASED EXPLORATION OF THE BOUNDARIES OF PHRASEOLOGY ON THE EXAMPLE OF CZECH AND GERMAN PARTICIPLE I FORMS

Olga Richterová

The Institute of Czech National Corpus

richterova.olga@gmail.com

The paper explores the boundaries and meeting points of phraseology, morphology and word-formation while looking into the following questions: Can we distinguish where compounds end and lexical phrasemes begin? And is it necessary to draw a line between these two terms? The underexplored issue of *single-word idioms* (sometimes called *Einwortidiome*) is approached from the perspective represented by Čermák's definition of phrasemes / idioms as units that "exhibit a restricted and anomalous combinatorial capacity of their constituents with no rule behind them" (2007: 23).

The study draws on the idea that "recognizing an idiomatic combination of morphemes within the boundaries of a single lexeme, is an important field of concern" (Čermák 2007: 21), further elaborated on by drawing attention to the differences between various language types: "compounding is a rich field to

look for lexical idioms, in direct proportion to the language's typological preference for compounds" (2007: 22).

For the purposes of this study, two languages are chosen as the basis for the investigation of the phraseological potential of a specific type of compounds: the analysed lexemes are derived from participle I forms of verbs and used as adjectives. In highly flective Czech, the *-oucí/ící* group is exemplified by the adjective *dechberoucí* while the isolating German disposes with a multifunctional end-morpheme *-end* found in *atemberaubend*, both words meaning 'breathtaking'. Does the fact that German is very prone to compounding have an impact on the level of phraseology found in the given type of morpheme combinations? This is the last question the study addresses.

The data available through the KonText corpus manager (accessible online via www.korpus.cz) go into billions: Czech data used for the study are based on the SYN Corpus (2.2 billion word forms), the research into German makes use of the following corpora: InterCorp (74 million), Aranea (1 billion) and deWaC (1.3 billion word forms).

Preliminary results show that out of 37 participle forms comprising the *-oucí* group for Czech and constituting 162 lemmata in total, at least six can be classified as genuine lexical phrasemes (e.g. *pokrkujdoucí*, attested three times and functioning as a nominalization of the phraseme *jít (někomu) po krku – to be after someone, to seek someone´s life*). But how to categorize less clear-cut cases, bordering on simple compounding? For example, the extremely limited nominal paradigm which goes together with the form *-beraubend* (*atemberaubend* being by far the most frequent combination, *freiheits-, sinn(es)-, ohren-, lebens-,* and *kräfteberaubend* forming the complete list) is a perfect example of the above-described "restricted and anomalous combinatorial capacity". The analysis of individual (collocational) paradigms will form the basis of the paper while drawing into question the complementation of separately written participial forms, too.

References

ARANEA - COMPARABLE WEB CORPORA. V. BENKO. *The Institute of Czech national corpus*. [online] Available at: <http://www.korpus.cz> [Accessed 10 March 2015].

ČERMÁK, F. (2007). Idioms and morphology. In: H. Burger et al., eds, 2007. *Phraseologie / Phraseology: An International Handbook of Contemporary Research: Volume I*. Berlin / New York: de Gruyter. pp.20-26.

M. BARONI, S. BERNARDINI, A. FERRARESI AND E. ZANCHETTA. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation 43(3)*: 209-226.

INTERCORP - CZECH NATIONAL CORPUS. *The Institute of Czech national corpus*. [online] Available at: <http://www.korpus.cz> [Accessed 30 March 2015].

SYN - CZECH NATIONAL CORPUS. *The Institute of Czech national corpus*. [online] Available at: <http://www.korpus.cz> [Accessed 10 March 2015].

CONTRASTIVE ANALYSIS OF VERB-NOUN COLLOCATIONS OF 'UTTERANCE' IN FRENCH AND KOREAN

Sunock Shin
Paris 13 University
sunockshin@gmail.com

Pierre-André Buvet
Paris 13 University
pierreandre.buvet@gmail.com

This study concerns, in a contrastive analysis of verb-noun collocations of 'utterance' in French and Korean, how to set up proper method and theoretical basis in making a list of them and which Korean typology could be considered as being corresponded to French 'utterance'. We chose mainly noun including collocations, like a [verb-noun] for example, to be more selective about this study's scope. That's why various other collocation types like a [verb-adverb], an [adverb-adjective], a [noun-adjective] etc. won't be included in our case.

Basically, this study will be composed of two main analyses: the first one is about a collocation (collocations of 'utterance' in particular); the second one is a contrastive analysis between French and Korean.

In a contrastive analysis between two languages in question, an irregularity problem in the actualization phase could be often detected. Some previous studies about this problem would be discussed to see how it is usually treated.

Our main argument is to define collocations as lexically restrictive combination by the influence of combination restricting and semantic transparency. As having chosen a [verb-noun] as our main object of study, a selection of 'base-noun' (collocations-related) and an analysis of 'verbs-

collocate' (with the said noun) will be pursued as a prime mode of setting a list of collocations in French and Korean.

Based on the result of contrastive analysis, we propose five collocations-corresponding types between French (source language) and Korean (target language) like: collocations vs. collocations; collocations vs. free combination; collocations vs. one lexical element; collocations vs. idiomatic expression; collocations vs. no equivalents.

References

CRUSE, D. A. (1986). *Lexical Semantics.* Cambridge : Cambridge University Press.

GIRY-SCHNEIDER, J. (1987). *Les prédicats nominaux en français: les phrases simples à verbes supports.* Genève : Droz.

GROSS, M. (1996). *Les expressions figées en français.* Paris: Ophrys.

LAMIROY, B. ET AL. (2010). *Les expressions verbales figées de la francophonie : Les variétés de Belgique, de France, du Québec et de Suisse*. Paris : Ophrys.

LEE, S.H. AND PARK, S.Y. (2007). How to Design a Multilingual Database of Korean Collocations. *Eoneohag* 48, pp.46-64.

LEE, S.H. IM, H.P. AND HONG, J.S. (2009). Intérêt des classes sémantiques dans la construction d'une base de données en vue de l'étude constrative plurilingue des collocations : la cas de l'étude contrastive coréen-français. *Société Coréenne d'Enseignement de Langue et Littérature Françaises*, 31, pp.197-220.

LIM, G.S. (2002). *A Study on Korean Collocation*. Ph. D. Seoul National University.

MEJRI, S. (2008). Construction à verbes supports, collocations et locutions verbales. In: P. Mogorron AND S. Mejri, dir. *Las construcciones verbo-nominales libres y fijas. Aproximacion contrastiva y traductologica*. Université d'Alicante & Université Paris 13. pp.191-202.

MEL′ČUK, I. (1996). Lexical functions: a tool for the description of lexical relations in a lexicon in a lexicon. In L. Wanner, ed. *Lexical functions in lexicography and natural language processing*. Amsterdam: John Benjamins. pp.37-102.

MEL′ČUK, I. (1998). Collocations and Lexical Functions. In A. P. Cowie, ed. *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press. pp.23-53.

PARK, M. G. (2005). Analyse contrastive des constructions avec verbes supports en français et en coréen. *Société Coréenne d'Enseignement de Langue et Littérature Françaises*, 20, pp.189-224.

SINCLAIR, J. (1991). *Corpus, Concordance, and Collocation*. Oxford: Oxford University Press.

TONI GONZÁLEZ RODRÍGUEZ (2004). Dictionnaire des collocations. [online] Available at: < http://www.tonitraduction.net/> [Accessed 1 mars 2015]

VIVÈS, R. (1984). L'Aspect dans les constructions nominales prédicatives: avoir, prendre, verbe support et extension aspectulle. *Linguisticae Investigationes* 3(1), pp.161-185.

THE PARALLEL POLISH-BULGARIAN-RUSSIAN CORPUS: PROBLEMS AND SOLUTIONS

Wojciech Paweł Sosnowski
The Institute of Slavic Studies of the Polish Academy of Sciences
wpsosnow@uw.edu.pl

The parallel Polish-Bulgarian-Russian corpus we are currently developing as part of CLARIN-PL framework will become an essential tool for translators producing both traditional and digital translations. The electronic tools developed within the project facilitate fast search for and retrieval of multilingual equivalents of lexemes, phrases and sentences. Selected sentences and texts have been semantically annotated for the quantification of nomen, time and aspect. Our definition of equivalent stems from the contemporary contrastive linguistics theory. The guiding principle in the construction of the corpus was to proceed from meaning to form; the principle was first introduced in Koseska-Toszewa (2006).

During our work on the Polish-Bulgarian-Russian corpus, we have come across a number of issues, which we regard as characteristic of multilingual corpora: (1) the selection and procurement of texts, (2) the development of computer tools used for the construction of the corpus, (3) multilingual equivalence, and (4) semantic annotation.

Multilingual corpora have proved to be exceptionally helpful in language teaching, traditional and digital lexicography, as well as traditional and digital

translations. The usefulness of multilingual corpora in each of these areas will be demonstrated through example corpus queries.

References

Koseska-Toszewa, V. (2006). Gramatyka konfrontatywna bułgarsko-polska (V. Koseska-Toszewa, J. Penčev (Eds.) (vol. 7. Semantyczna kategoria czasu). Warszawa: Slawistyczny Ośrodek Wydawniczy.

IN-DEPTH STUDY OF THE PHRASEOLOGICAL UNITS IN ISLAMIC AND CHRISTIAN RELIGIONS IN SAMPLES (CORPORA) OF RELIGIOUS TEXTS

Madian Souliman **Ali Ahmad**

Over the last two decades there has been a great deal of interest in lexical studies, particularly in the combinations of words in natural languages. Conventionalized forms, frames, idioms, and collections have proven to be chiefly appealing in the areas of phraseology. The actuality of present research is conditioned by necessity of studying of the characteristics of the phraseological units in-depth and as they are expressed in the “Holy Bible and Holy Quoran” which will reveal many methods and approaches in translating them basically in the religious texts and will help us to bind up the whole religious texts (Bible & Quran) in one computerized corpus which will provide us with different translations of those holy texts for a comparative study. We cannot neglect that some of the earliest efforts at grammatical description were based at least in part on corpora of particular religious as the early Arabic grammarians paid particular attention to the language of the Quran which can prove that corpus linguistics adherents have reliable language analysis which best occurs on field-collected samples, in natural contexts and with minimal experimental interference. Within corpus linguistics there are divergent views as to the value of corpus annotation , from John Sinclair advocating minimal

annotation and allowing texts to ‘ speak for themselves’, to others, such as the Survey of English Usage team advocating annotation as a path to greater linguistic understanding and rigour. We cannot deny that a computerized corpus of the religious texts was found many years ago. An example is the Andersen-Forbes database of the Hebrew Bible, developed since the 1970s, in which every clause is parsed using graphs representing up to seven levels of syntax, and every segment tagged with seven fields of information. Another example is the Quranic Arabic Corpus which is an annotated corpus for the classical Arabic Language of the Quran. The subject of this presentation is to consider some peculiarities and problems in translating the religious texts especially after taking into consideration that there are many phraseological units in their sacred contexts and the results will be discussed after examining the translation of sixteen examples from both the Holy Qur’an and the Holy Bible. The main goal of the research is to create a computerized corpus for the phraseological units in the religious texts which later on can provide us with not only the state of the language in samples but also different translations with notes on their differences.

References

Ahmad, A. Итоговая научно-образовательная конференция студентов Казанского Федерального университета 2014 года:*In-depth study of the paremiological units in Islamic and Christian religions with some contemporary interpretations in their translations*. p. 157.

The observatory of language sciences, 2013, the Dep. of Vernacular Languages & the Graduate Program in Linguistics at the Federal University of Ceará. 1 p. 09/05/2014

Ayupova (2012). *English Phraseology* ,Kazan. 7p. 09/05/2014

Арутюнова Н.Д. Дискурс // *Лингвистический энциклопедический словарь*. – М., 1990. – 260с.

http://en.wikipedia.org/wiki/Set_phrase [Accessed 09 May 2014]

Anita Naciscione, *Phraseological units in discourse: towards applied stylistics*, 2001.- 5p. 09/05/2014

Anita Naciscione, *Phraseological units in discourse: towards applied stylistics*, 2001.- 5p. 09/05/2014

http://www.rusnauka.com/10._ENXXIV_2007/Philologia/21605.doc.htm [Accessed 09 May 2014]

Anita Naciscione , *Phraseological units in discourse:* 2001 *towards applied stylistics*,- 5p. 09/05/2014

The Observatory of Language Sciences, 2013 the Dep. of Vernacular Languages & the Graduate Program in Linguistics at the Federal University of Ceará . 1 p. 09/05/2014

http://en.wikipedia.org/wiki/Proverb#Paremiology. [Accessed 17 March 2014]

Wolfgang Mieder. 1990. *Not by bread alone: Proverbs of the Bible*. New England Press, 12p. 17/03/2014
The Holy Bible: *The old and New Testament* No. of books 66 & 1,189 chapters 1,281 ps. 1p. 14/03/2014
Wolfgang Mieder. 1990. *Not by bread alone: Proverbs of the Bible*. New England Press, 12p. 17/03/2014
http://www.americancorpus.org [Accessed 17 March 2014]
http://corpus.byu.edu/bnc/ [Accessed 17 March 2014]
http://corpus.byu.edu/bnc/ [Accessed 17 March 2014]
http://www.americancorpus.org [Accessed 17 March 2014]
http://en.wikipedia.org/wiki/Religious_views_of_Albert_Einstein [Accessed 19 March 2014]
http://www.biographyonline.net/scientists/albert-einstein-quotes.html [Accessed 19 March 2014]
А.В. Кунин. *О переводе английских фразеологизмов в англо-русском фразеологическом словаре*. Тетради переводчика. М.,1964№2. 12/06/2014
Shakespeare —*Much Ado About nothing*‖, act 5, scene 1. 12/06/2014
А.В. Кунин. *О переводе английских фразеологизмов в англо-русском фразеологическом словаре*. Тетради переводчика. М.,1964№2. 12/06/2014
The Holy Qur'an, *Surat Al-Baqarah* 15 . 20/06/2014
The Holy Qur'an, *Surat Al-Baqarah* 27 . 20/06/2014
The Holy Qur'an, *Surat Al-Baqarah* 63 . 20/06/2014
The Holy Qur'an, *Surat Al-Imran* 32. 20/06/2014
The Holy Qur'an, *Surat Al-Imran* 92 . 20/06/2014
The Holy Qur'an, *Surat Al-Imran* 103 . 20/06/2014
Imam Zain ul Abideen , *As-Sahifa Al-Kamilah Al-Sajjadiyya, supplication* 53 verse 1. 20/06/2014
The Holy Qur'an, *Al-Anaam* Verse No:78. 20/06/2014
The Holy Bible, *Psalms* 69, 2. 20/06/2014
The Holy Bible, *Proverbs* 16,18. 20/06/2014
The Holy Bible, I *Samuel* 23, 16. 20/06/2014
The Holy Bible, *Proverbs* 1, 14. 20/06/2014
The Holy Bible, *Psalms* 58, 4. 20/06/2014
The Holy Bible, *Luke* 12, 3. 20/06/2014
The Holy Bible, *Proverbs* 25, 13. 20/06/2014
The Holy Bible, *John* 5, 35. 20/06/2014

COMBINACIONES USUALES DE PALABRAS EN ALEMÁN DE VALOR ADVERBIAL: PATRONES SINTAGMÁTICOS COMO PARÁMETRO DE EQUIVALENCIA EN ESLOVACO Y ESPAÑOL

Kathrin Steyer
Institut für Deutsche Sprache, Mannheim
steyer@ids-mannheim.de

Carmen Mellado
Universidade de Santiago de Compostela
c.mellado@usc.es

Peter Ďurčo
Univerzita sv. Cyrila a Metoda v Trnave
peter.durco@ucm.sk

En nuestro trabajo abordamos el análisis de las combinaciones usuales de estructura [Prep. + S] desde un punto de vista contrastivo, las cuales apenas han sido objeto de estudio hasta el momento. Se trata de combinaciones binarias en alemán del tipo nach Belieben, aus Versehen, in Wahrheit, bei Gelegenheit, caracterizadas por llevar artículo cero y por su función adverbial, ya sea a nivel sintagmático u oracional.

Nuestro estudio se apoya en la teoría sobre los patrones sintagmáticos de las combinaciones usuales de Kathrin Steyer (cfr. Steyer 2013), que consiste en la identificación y descripción inductiva de perfiles combinatorios recurrentes de dichas combinaciones (por ejemplo nach Belieben + VERB: dominieren / beherrschen / kontrollieren / diktieren) [DOMÄNE: SPORT]) a partir de la valoración de datos extraídos de grandes corpus.

Para el análisis contrastivo se seleccionan combinaciones usuales del alemán que cuenten en las lenguas meta eslovaco y español con al menos un equivalente adverbial suficientemente consolidado desde un punto de vista de

la frecuencia en los corpus respectivos. En el caso de nach Belieben – Eslovaco: podľa ľubovôle ("nach Belieben"), según el contexto también podľa (svojej, vlastnej) vôle ("nach dem [eigenen"] Willen), podľa (svojej, vlastnej) chuti ("nach [eigenem] Geschmack"), podľa (svojho, vlastného) želania ("nach [seinem, eigenem] Wunsch"), entre otros; – Español: según el contexto a placer, a mi* antojo, al gusto, a mi* gusto, a capricho, a discreción, a voluntad, entre otros.

Uno de nuestros principales objetivos es mostrar como el método de ánalisis de los patrones sintagmáticos del nodo (KWIC: key word in context), así como del tipo de saturación léxica de los slots que componen dichos patrones en cada una de las tres lenguas, puede contribuir de modo significativo a la descripción detallada del uso distintivo de las combinaciones usuales en cada una de las lenguas. Los resultados de nuestra investigación pueden ser implementados en el campo de la didáctica de lenguas extranjeras, ya que con este método de análisis se facilita la adquisición y uso adecuados de las combinaciones usuales en su contexto. Asimismo, este procedimiento empírico basado en corpus arroja una nueva luz en la tarea de determinar si estamos ante unidades equivalentes de facto entre varias lenguas, sobre todo si tenemos en cuenta que el componente pragmático de las unidades lingüísticas (p. ej. connotaciones, implicaturas y dominios de uso) constituye un parámetro fundamental en el tema de la equivalencia interlingüística, más allá de lo que se suponía hace unos años, antes del análisis masivo de datos de corpus.

References

Ďurčo, Peter (2013): Extensionale und intensionale Äquivalenz in der Phraseologie am Beispiel von deutschen und slowakischen Sprichwörtern“. In: J. M. Benayoun, N. Kübler, J. P. Zouogbo (eds.): Parémiologie. Proverbs et formes voisines. Tome 2, Presses Universitaires de Sainte Gemme 2013, 49-64.

Mellado Blanco, Carmen (2010): Die phraseologische Äquivalenz auf der System- und Textebene. In: Korhonen, Jarmo et al. (eds.): Phraseologie global - areal - regional. Akten der Konferenz EUROPHRAS 2008 vom 13.-16.8.2008 in Helsinki. Tübingen, 277-284.

Steyer, Kathrin (2013): Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht . (=Studien zur Deutschen Sprache 65). Tübingen.

APRENDER FRASEOLOGÍA MEDIANTE CORPUS: UN CASO APLICADO A LA ENSEÑANZA DEL ALEMÁN

María Rosario Bautista Zambrano

Universidad de Málaga

mrbautista@uma.es

El presente trabajo da cuenta de la experiencia realizada en la asignatura Idioma Moderno II (Alemán), de nivel A1.2, perteneciente al título de Graduado en Estudios Ingleses. Partiendo de la premisa de la utilidad del corpus para la enseñanza de lenguas extranjeras (tema abordado, por ejemplo, en Flowerdew (1996) y López Sanjuán (2008)), hemos diseñado una serie de actividades basadas en corpus para el aprendizaje de elementos gramaticales y léxicos de la lengua alemana; dentro del componente léxico nos hemos centrado en el aprendizaje de unidades fraseológicas, especialmente colocaciones. Previamente al trabajo en el aula, compilamos un corpus acerca del tema elegido, *Orientierung in der Stadt* (orientación en la ciudad), con textos basados en la descripción del camino para llegar a un sitio. Siguiendo las recomendaciones de Leal Riol (2013), tratamos de que el corpus fuera un fiel reflejo del alemán en uso y que ofreciera muestras lingüísticas adecuadas; en ese sentido, hemos tratado de escoger textos sencillos y que contuvieran colocaciones habituales e incluidas en el libro de texto empleado en clase, *DaF kompakt A1* (Braun et ál., 2010). Así mismo, hemos seleccionado las unidades fraseológicas basándonos en criterios de frecuencia, facilidad, productividad y

necesidad (Leal Riol, 2013); de esta forma, al igual que en el caso del corpus, hemos tenido en cuenta el nivel de competencia de los alumnos (aún básico) y el objetivo de este aprendizaje, que es el de que los estudiantes las empleen en producciones escritas y orales. Una vez compilado el corpus, pasamos a utilizarlo en clase con los alumnos. El tema de *Orientierung in der Stadt* ya había sido tratado brevemente en clase. Primero, se les impartió una breve introducción del concepto de corpus y de su utilidad para el aprendizaje de lenguas; a continuación, se les enseñaron nociones básicas de utilización del programa de concordancias *AntConc*. El siguiente paso consistió en trabajar con el corpus preparado: en primer lugar, se realizó un ejercicio de extracción de las palabras más frecuentes, a fin de que los alumnos se familiarizasen con las palabras más comunes usadas en este ámbito; en segundo lugar, se realizaron ejercicios para descubrir, por medio del corpus, unidades fraseológicas útiles en relación con la descripción de un camino. Se realizaron dos tipos de ejercicios: tareas de rellenar huecos (donde tenían que encontrar una de las palabras pertenecientes a una colocación) y ejercicios de deducir un patrón a partir de unas concordancias dadas. Finalmente se les pidió que, con lo aprendido, realizaran una breve redacción donde explicaran a un viandante cómo llegar de un punto A al B. Ofreceremos las conclusiones a las que hemos llegado con esta actividad, aportando los resultados y comentando las dificultades con las que nos hemos encontrado.

References

ANTHONY, L. (2015). *AntConc Homepage*. [en línea] Disponible en: <http://www.laurenceanthony.net/software/antconc/> [Fecha de acceso 30 marzo 2015].

BRAUN, B., DOUBEK, M., FRATER-VOGEL, A., TREBESIUS, U., VITALE, R. Y SANDER, I. (2010). *DaF kompakt A1*. Klett.

FLOWERDEW, J. (1996). Concordancing in language learning. In: M. Pennington, ed. 1996. *The power of call*. Houston, TX: Athelstan. pp. 97-113.

LÓPEZ SANJUÁN, V. (2008). Integración de los corpus como herramienta de apoyo en la enseñanza de ESP. *Porta Linguarum*, 10, pp. 115-136.

LEAL RIOL, M. J. (2013). Estrategias para la enseñanza y aprendizaje de la fraseología en español como lengua extranjera. *Paremia*, 22, pp. 161-170.

¿DAR O ECHAR UN PIROPO? ME QUEDO LOCO, NUNCA ACIERTO. COLOCACIONES VERBALES EN ESPAÑOL Y PORTUGÉS

Javier Martín Salcedo

Universidade Federal do Ceará, Brasil

javims29@hotmail.com

El proceso de enseñanza y aprendizaje de las Unidades Fraseológicas, y en el caso concreto de las colocaciones verbales, es bastante complejo, ya que sólo el propio uso determina la configuración de las mismas. Entre los aspectos que dificultan, por parte de los estudiantes extranjeros, la comprensión de los significados y, por consiguiente, la asimilación y uso de estas unidades, está la amplia variedad de funciones, diversas formas y, en muchos casos, su opacidad y sentido figurado o metafórico. Por lo tanto, el objetivo de esta comunicación plantea la necesidad de pensar en estrategias para que los alumnos sepan identificar estas unidades fraseológicas en diferentes géneros orales y escritos con la finalidad de mejorar la calidad de la enseñanza en lenguas extranjeras tan próximas. En este sentido, este estudio pretende presentar algunos usos lingüísticos que difieren en ambas lenguas en cuanto al uso y a la elección del colocativo en estas UFs, como por ejemplo: en español europeo, por una parte, sería más común, usual o natural, la elección “echar una bronca” que “dar una bronca”, mientras que en portugués sería “dar uma bronca”, ya que echar o algún equivalente del mismo no se admitiría; por otra,

en portugués, valga decir, la estructura usual es “dar um elogio” para “echar un piropo”, y que, sin duda, para un español sonaría extraño “dar un piropo” como colocación consagrada por el propio uso.

En cualquier caso, cabe destacar que la investigación se desarrolló en la Universidad Federal de Ceará, en Brasil, basándose en producciones de alumnos brasileños de español en dicha universidad y alumnos españoles de portugués de la E.O.I de Castellón a través de un grupo de Facebook, en el que se proporcionaron intercambios lingüísticos que se llevaron a cabo durante los meses de octubre, noviembre y diciembre de 2014. Como fuentes para la misma, podríamos tener en cuenta a diversos autores como Molina Plaza y su libro: *La traducción de las unidades fraseológicas inglés-español: el caso de las colocaciones y frases idiomáticas*; Zuluaga y su publicación: “*Los enlaces frecuentes" de María Moliner. Observaciones sobre las llamadas colocaciones”*; y el propio *Manual de Fraseología Española* de Corpas Pastor, entre otros etc… De este modo, la elaboración de un compendio de colocaciones verbales en contraste es más que necesario, seleccionando las unidades con más frecuencia de uso en la lengua coloquial e intentando resolver los principales problemas de interlengua de los estudiantes que desconocen el uso correcto o habitual de este aspecto fraseológico entre dos lenguas tan próximas, pero a la vez tan distantes, como son el español y el portugués.

LA ENSEÑANZA DE REFRANES EN EL CORPUS DE TRADUCCIONES A CHINO DE *EL QUIJOTE*

Li Mei Liu Liu

Universidad Complutense de Madrid

limeigomez@gmail.com

Los refranes como formas literarias de carácter folclórico pueden construir una vía para ayudar al estudiante en clase de ELE a alcanzar los conocimientos lingüísticos e sculturales sobre el pueblo que los emplea. Por lo tanto, este trabajo pretende analizar el corpus de refranes en *El Quijote* mediante sus traducciones al chino desde el 1978 hasta hoy en día; en vista de poder concluir un acercamiento o un alejamiento a lo largo de estos años sobre los conocimientos del español por parte de los traductores chinos.

Además, a través de dos métodos traductológicos que define Venuti: el método de domesticación y el método de extranjerización, se precisarán una serie de cuestiones como ¿cuál es el método, por parte de los traductores chinos, que predomina en transmitir estas expresiones populares de la lengua española?, ¿En las traducciones hechas en diferentes épocas se hallaría una resistencia a la hegemonía cultural occidental? o ¿Se encontraría el fenómeno del etnocentrismo de la propia cultura?

Para poder concluir cuál de las traducciones es la más idónea para el aprendizaje de los refranes; el método del presente estudio consistirá en exponer los ejemplos incluidos en el *Quijote* con sus correspondientes traducciones. Sin embargo, puesto que son numerosos los refranes contenidos

en la obra maestra de Cervantes, seleccionaremos solo aquellos que están en uso vigente según el refranero en línea del Centro Virtual Cervantes. También se citarán las notas de edición incluidas por sus propios traductores para así apreciar qué competencia extralingüística dificultaría la enseñanza del español en el caso de los refranes para los alumnos chinos.

References

CATFORD, J. C. (1907). *A linguistic Theory of Translation: An Essay in Applied Linguistics,* Londres: Oxford University Press, 1965 (*Una teoría lingüística de la traducción: ensayo de lingüística aplicada.* Carasca: Universidad Central de Venezuela.

CHEN, CHIA-YING (2009). *Estudio sobre la traducción al chino de los refranes de Sancho Panza en Don Quijote de Yang Chiang.* Tesina de Máster de la Universidad Providence de Taiwan.

Don Quijote de la Mancha. Edición del Instituto Cervantesdirigida por Francisco Rico en <http://cvc.cervantes.es/literatura/clasicos/quijote/> [Accessed 16 de marzo de 2015].

HURTADO ALBIR, A. (2001). *Traducción y traductología.* Madrid: Cátedra.

NIDA, E.A. (2012). *Sobre la traducción.* Madrid: Cátedra.

NIDA, E.A. (2001). *Language and Culture-Contexts in Translating,* Shanghai: shanghai foreign language education press.

SEVILLA MUÑOZ, J. (2005). Presupuesto paremiológico de una propuesta metodológica para la enseñanza de los refranes a través de El Quijote. *Paremia, 14*. pp.117-128.

VENUTI, L. (1998). *The Scandals of Translation: Towards an Ethics of Difference*. London/New York: Routledge.

VENUTI, L. (2004). *The Translator's Invisibility: A History of Translation*. London/New York: Routledge.

YANG, WENFEN (2010). Brief Study on Domestication and Foreignization in Translation. *Journal of Language Teaching and Research* 1 (January): 77–80. doi:10.4304/jltr.1.1.77-80.

Traducciones:

楊絳(Yang Jiang) 译,1992.《*堂吉軻德*》台北: 聯經.

孙家孟(Sun Jianmeng) 译,2001,《*奇想联翩的绅士堂吉诃德·德·拉曼恰*》北京: 北京十月文艺出版社.

董燕生(Dong Yansheng) 译,2006.《*奇思异想的绅士堂吉诃德·德·拉曼却*》武汉: 长江文艺出版社.

崔维本译(Chui Weiben) 译,2007.《*堂吉诃德*》北京: 中国少年儿童.

LINGUEE COMO HERRAMIENTA DE ENSEÑANZA-APRENDIZAJE DE LAS UNIDADES FRASEOLÓGICAS

Maria Eugênia Olímpio de Oliveira Silva

Universidad de Alcalá

eugenia.olimpio@uah.es

Inmaculada Penadés Martínez

Universidad de Alcalá

inmaculada.penades@uah.es

La aplicación de la lingüística de corpus a la enseñanza del español como L2 es relativamente reciente y está poco desarrollada en comparación con el inglés en lo que se refiere a corpus producidos por hablantes nativos y no nativos. Los primeros son aptos para su uso en el aula por el profesor y los estudiantes, y para la realización de tareas en casa por estos últimos. Los segundos son especialmente útiles para un análisis de errores que permita diseñar contenidos de los cursos de la L2 más ajustados a las necesidades reales de los aprendices y para revisar los materiales pedagógicos en que se basa la enseñanza. Junto a ello, todavía es menor la utilización por el profesor de ELE de las técnicas que proporciona el procesamiento del lenguaje natural para analizar unidades fraseológicas, aunque la fraseología computacional ya permite su detección, extracción, análisis y representación (Corpas Pastor, 2013).

Desde este punto de partida, en la comunicación se presenta la aplicación a la enseñanza de ELE de Linguee, una herramienta considerada, por algunos, un corpus público de fácil acceso y, asimismo, un corpus paralelo (Alonso Jiménez, 2013), si bien sus creadores lo consideran un diccionario inteligente.

Este recurso ha recibido la atención de diferentes estudios, que han examinado su potencial como instrumento para el traductor o para la enseñanza de la traducción (Durán Muñoz, 2011; Patin, 2014). En menor medida, ha sido examinada su aplicabilidad al proceso de enseñanza-aprendizaje de lenguas extranjeras (Buyse y Verlinde, 2013; Volk *et al*. (2014).

Así pues, dado un conjunto de unidades fraseológicas del español, relativas a un mismo campo conceptual y vinculadas por un mismo proceso cognitivo de formación, se expondrán los resultados que se podrían obtener del examen, por parte de aprendientes de español, de sus equivalentes de traducción al portugués al establecer, a partir de su funcionamiento en el discurso, las analogías y diferencias morfológicas, sintácticas, semánticas y pragmáticas existentes entre las unidades de ambas lengua. Independientemente de la competencia que sobre ellas adquiera el estudiante, la reflexión metalingüística sobre la traducción facilita la confección de un glosario que puede servir de punto de referencia para analizar el tratamiento lexicográfico de las unidades fraseológicas en diccionarios bilingües de estas lenguas. Se trata de una propuesta de enseñanza-aprendizaje acorde con un enfoque constructivista, pues busca incidir positivamente en el desarrollo del trabajo reflexivo y autónomo del estudiante. Asimismo, se propone, de hecho, un empleo pedagógico de un recurso tecnológico, porque Linguee se usa como instrumento cognitivo, al servicio del aprendizaje de la fraseología (Coll, 1994; Solano Rodríguez, 2012).

Como la práctica expuesta responde a la cuestión ¿cómo trabajar las unidades fraseológicas?, en la comunicación se dará cuenta, asimismo, de una muestra de unidades con las que practicar a partir de esta propuesta y del nivel de enseñanza para el que la actividad es apropiada, teniendo en cuenta las distinciones en niveles establecidas por el *Marco común europeo de referencia para las lenguas* (Consejo de Europa, 2002).

References

ALONSO JIMÉNEZ, E. (2013). Linguee y las nuevas formas de traducir. *Skopos*, 2. pp. 5-28.

BUYSE, K. AND VERLINDE, S. (2013). Possible Effects of Free on Line Data Driven Lexicographic Instruments on Foreign Language Learning: The Case of Linguee and the Interactive Language Toolbox. *Procedia - Social and Behavioral Sciences*, 95. pp. 507-512. Available through:

<http://www.sciencedirect.com/science/article/pii/S1877042813041955> [Accessed 09 March 2015].

COLL, C. (2004). Psicología de la educación y prácticas educativas mediadas por las tecnologías de la información y la comunicación: una mirada constructivista. *Sinéctica*, 25. pp. 1-24. Available at: <http://www.virtualeduca.org/ifd/pdf/cesar-coll-separata.pdf> [Accessed 09 March 2015).

CONSEJO DE EUROPA (2002). *Marco común europeo de referencia para las lenguas*: *aprendizaje, enseñanza*, *evaluación.* Madrid: MECD / Anaya.

CORPAS PASTOR, G. (2013). Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In: I. Olza and E. Manero Richard, eds. *Fraseopragmática*. Berlin: Frank & Timme. pp. 335-373.

DURÁN MUÑOZ, I. (2011). Recursos electrónicos para la búsqueda terminológica en traducción: clasificación y ejemplos. *Revista Tradumàtica*: *Tecnologies de la Traducció i de la Informació i la Comunicació*, 9. [online] Available at: <http://revistes.uab.cat/tradumatica/issue/view/9> [Accessed 09 March 2015).

PATIN, S. (2014). Del uso de los córpora paralelos en la enseñanza de la traducción: el caso del Europarl. In: F. Olmo Cazevieille and J.-M. Mangiant, eds. *II Coloquio franco-español de análisis del discurso y enseñanza de lenguas para fines específicos*. *Lenguas*, *comunicación y tecnología digitales*. Valencia: Universidad Politècnica de València. pp. 159-174.

SOLANO RODRÍGUEZ, M.ª Á. (2012). Fraseodidáctica basada en tecnologías digitales. In: M.ª I. González Rey, ed. *Unidades fraseológicas y TIC* (Biblioteca Fraseológica y Paremiológica, Serie «Monografías», nº 2). Madrid: Instituto Cervantes / Centro Virtual Cervantes. pp. 167-186.

VOLK, M., GRAËN, J. CALLEGARO, E. (2014). Innovations in Parallel Corpus Search Tools. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (*LREC-2014*). Available at: <http://www.lrec-conf.org/proceedings/lrec2014/summaries/504.html> [Accessed 09 March 2015).

LA FRASEOLOGÍA COMO RECURSO HUMORÍSTICO EN NIÑOS DE EDUCACIÓN PRIMARIA

Larissa Timofeeva Timofeev

Universidad de Alicante

timofeeva@ua.es

En este trabajo –que se integra dentro del proyecto de investigación *Innovaciones lingüísticas del humor: géneros textuales, identidad y enseñanza del español* (FFI2012-30941)– pretendemos analizar el uso de diversos mecanismos fraseológicos por niños de 9-10 años que participaron en un estudio sobre el humor verbal infantil.

El análisis se apoya en los siguientes pilares teóricos. Por un lado, el desarrollo de la competencia metapragmática centra nuestra atención, pues los niños de la edad estudiada se encuentran en un momento clave marcado por el paso desde una comunicación epipragmática hacia una nueva fase caracterizada por la capacidad de, no solo ajustar su mensaje al contexto sino también de explicar y explicitar conscientemente sus decisiones pragmáticas (cfr. Gombert 1992; Eccles 1999; Stude 2007; Crespo & Alfaro 2010; Collins *et al.* 2014; Verschueren 2000). En este contexto, los estudios procedentes de la psicología evolutiva que indagan sobre el desarrollo de la competencia fraseológica en niños constituyen otro foco de interés (cfr. Laval 2003; Bernicot *et al.* 2007; Crespo *et al.* 2007). Finalmente, dentro de la habilidad metapragmática, la competencia humorística experimenta un cambio sustancial en la franja vital estudiada en que el lenguaje empieza a adquirir un papel

fundamental (Socha & Kelly 1994; Socha 2012; Martin 2007; McGhee 1979, 2002; Hoicka 2014).

Las nociones teóricas –conciencia metapragmática, competencia fraseológica, competencia humorística– obtienen su concreción en el estudio de marcas e indicadores verbales que utiliza el hablante para construir su discurso humorístico (Ruiz Gurillo 2012; Timofeeva 2014). En este caso, analizamos un corpus de 149 narraciones humorísticas escritas por niños y niñas de 4º curso de Educación Primaria procedentes de diversos colegios de la provincia de Alicante con el fin de observar cómo usan la fraseología y cómo estos usos se relacionan con el desarrollo de su competencia metapragmática del humor. Otras variables –como la perspectiva de género, el tipo de centro (público o privado) o el programa lingüístico (monolingüe, bilingüe o multilingüe) que siguen los alumnos– también se tienen en cuenta.

Así, tres grandes grupos de fenómenos ligados a la fraseología serán analizados: la composición sintagmática, el uso de locuciones y la desautomatización fraseológica. Según nuestra hipótesis, su incidencia en las narraciones escolares nos da valiosas pistas sobre el grado de madurez metapragmática y de la evolución de la competencia humorística. El objetivo, por tanto, es determinar en qué términos, cuantitativos y cualitativos, están presentes los fenómenos fraseológicos enunciados en las muestras escritas de nuestro corpus.

References

BERNICOT, J. *et al*. (2007). “Nonliteral language forms in children: In what order are they acquired in pragmatics and metapragmatics?”, *Journal of Pragmatics* 39 (2007), pp. 2115-2132.

COLLINS, A. *et al. (*2014). “Metapragmatic explicitation ability in children with typical language development: Development and validation of a novel clinical assessment”. *Journal of Communication Disorders*, 52, pp. 31-43.

CRESPO, N., BENÍTEZ, R. & CÁCERES, P. (2007). “La comprensión oral del lenguaje no literal y su relación con la producción escrita en escolares”. *Revista Signos*, 40 (63), pp. 31-50.

CRESPO, N. & ALFARO, P. (2010). “Desarrollo tardío del lenguaje: la conciencia metapragmática en la edad escolar”. *Universitas Psychologica*, 9:1, pp. 229-240.

ECCLES, J. S. (1999). “The Development of Children Ages 6 to 14”, The Future of Children. When school is out, 9:2, pp. 30-44. Available at: http://www.princeton.edu/futureofchildren/publications/journals/article/inde

x.xml?journalid=48&articleid=232&sectionid=1517. Accessed 20 December 2014.

GOMBERT, J. (1992). *Metalinguistic development*. New York: Wheatsheaf.

HOICKA, E. (2014). "The Pragmatic Development of Humor". In *Pragmatic Development in First Language Acquisition*, Matthews, Danielle (ed.), pp. 219–238.

LAVAL, V. "Idiom comprehension and metapragmatic knowledge in French children", *Journal of Pragmatics* 35 (2003), pp. 723-739.

MARTIN, R. A. (2007). *The Psychology of Humor: An Integrative Approach*. Burlington: Elsevier Academic Press.

MCGHEE, P. E. (1979). *Humor: Its Origin and Development*. San Francisco: W. H. Freeman.

MCGHEE, P. E. (2002). *Understanding and Promoting the Development of Children's Humor*. Dubuque: Kendall Hunt Publishing.

RUIZ GURILLO, L. (2012). *La lingüística del humor en español*. Madrid: Arco/Libros.

SOCHA, T. J. & KELLY, B. (1994). "Children making 'fun': humorous communication, impression management, and moral development". *Child Study Journal*, 24:3, pp. 237-252.

SOCHA, T. J. (2012). "Children's humor: Foundations of laughter across the lifespan". In R. DICIOCCIO (ed.), Humor: *Theory, Impact, and Outcomes* (Chapter 9). Dubuque, IA: Kendal Hunt.

STUDE, J. (2007). "The acquisition of metapragmatic abilities in preschool children", en BUBLITZ, W. & HÜBLER, A. (eds.), *Metapragmatics in use*. Amsterdam, John Benjamins, pp. 199-220.

TIMOFEEVA, L. (2014). "El humor verbal en niños de educación primaria: presentación de un estudio". *Femenismo/s,* 24.

VERSCHUEREN, J. (2000). "Notes on the role of metapragmatic awareness in language use", *Pragmatics*, 10:4, pp. 439-456.

Phraseology in E-Lexicography and E-Terminography La información fraseológica en la lexicografía y la terminología electrónicas

PHRASEOLOGY - CULTURAL CODE OF ETHNICITY (ON THE MATERIAL OF FRENCH, ENGLISH, AND GEORGIAN LANGUAGES)

Tsiuri Akhvlediani

Tbilisi State University, Georgia

tsiuriakhvlediani@yahoo.com

George Kuparadze

Tbilisi State University, Georgia

gkuparadze@yahoo.com

The language is considered to be a cultural code of ethnicity; an informative intermediary with a national marker contemplating feelings, perception and presentation of the universe.

It always identifies the nation's characteristic features and culture. National-cultural signs of phraseology are formed by: 1. Peculiarities of linguo-creative thinking; 2. Ethno-lingual specific interpretation of the universe; 3. Secondary conceptualisation and categorisation peculiarities of images reflected in the human consciousness based on the status of objects so important to the given ethnicity.

It is known that a language, especially its lexicon, influences the speakers' cultural patterns of thought and perception in various ways.

The aim of the presented paper is to explore the ethno-cultural dimension of a wide range of pre-constructed or semi-pre-constructed word combinations focusing on the scope of their diachronic evolution in French, English and Georgian languages.

The corpora for investigation includes multiword units of the kick-the-bucket type, collocations, irreversible binominals, phrasal verbs, compounds, metaphorical expressions, similes, proverbs, familiar quotations, etc – all of which have been subsumed under phraseology.

On the basis of analysis of conceptual content space of national character, the following features verbalized by phraseological units have been singled out: 1) Basic, 2) Ethic, 3) Aesthetic, 4) Abnormal.

1. Basic or Natural Features, that are genetic for the character, express:

- Features of Temperament types; they are subdivided into 4 groups:
    a) phlegmatic – characterised by indifference;
    b) melancholic - characterised by pessimism;
    c) sanguineous – boasting a fast response;
    d) d) choleric - characterised by nervousness and hot-temper;
- Level of Intellect such as the highest mental abilities; limited mental abilities;
- Arbitrary properties, which, at the linguistic level are associated with firmness, determination, targeting of attack;

2. Ethical Features of an individual in ethno-culture are formed in the process of socialisation. Person's status depends on his/her relationship to labour, social environment; they are defined by moral-ethical standards. Negative ethical assessments have such features as: cunning, hypocrisy, flattery, lies, deception, ruddiness, harshness, talkativeness;

3. Aesthetic signs characterize an individual in accordance with his/her attitude to his/her self-appearance. They (the signs) denounce careless, negligent people; they encourage elegance, refinement, sophistication;

4. Abnormal signs are ascribed to those characteristics that constitute a deviation from the norm; their basis are either the ideas of individual rationalism or, on the contrary, an absolute indifference to everything, including self-respect as well. Abnormal signs reflect a narrow practitzism, without any tangible benefit to the aspirations or, on the contrary, indifference to everything:

- egocentrism, arrogance, haughtiness, uncontrollability, pamper, excessive
- pride;
- involuntary, ungracious.

- Absence of any positive qualities of character; such a person is assessed as “supernagative”.

Thus, the signs of character are the same for people of any nationality but they are manifested differently according to various traditions, culture, national temperaments and mentalities. The importance of vital fragments characteristic of French, English and Georgian Ethnic cultures is, by no means, defined by those of the values that are so closely interwoven in them. The scrupulous observation has also revealed that a proper understanding of language accepts cultural, social and cognitive perspectives to develop a better understanding of what we do when we talk, listen, read, write and are engaged in thinking that activates the internalized linguistic system at any level.

References

SAKHOKIA T. (2009). Phraséologismes géorgiens (en géorgien).
BOLLY C. (2008). Les unités phraséologiques. Louvain-la-Neuve.
BALLY CH. (1962). Traité de Stylistique française. Genève.
10 Heart Idioms Explained to English as a Second Language Learners. (2011) available at Kerlynb.hubpages.com

PHRASEOLOGICAL CORRESPONDENCE IN ENGLISH AND SPANISH SPECIALIZED TEXTS

Miriam Buendía Castro

Universidad de Málaga

mbuendia@uma.es

Pamela Faber

Universidad de Granada

pfaber@ugr.es

The selection of suitable interlinguistic correspondences at the word, phrase, and even text level represents one of the most problematic aspects of translation. Although a great deal has been written about translation equivalence, little of it is very useful. Apart from the traditional opposition of *faithful/free*, other term pairs such as *semantic/communicative* (Newmark 1981) and *formal/dynamic* (Nida and Taber 1969) have also been proposed as descriptions for the degree of perceived similarity at the level of form and/or function between a source language text and a target language text. However, these changes of label have not been accompanied by significant new insights into the nature of interlinguistic or intertextual equivalence. Whatever the terms used, judgments of equivalence often tend to be depressingly based on linguistic form instead of any sort of shared conceptual reference or meaning representation (Faber and Ureña 2012) despite the fact that this type of mirror-reflection equivalence is a chimera.

In terminology and specialized translation, correspondences between specialized knowledge units (SKUs) in different languages are generally established at the conceptual level. In other words, two SKUs are regarded as equivalents if their conceptual properties coincide. Syntactic correspondence is

only relevant to the extent that syntax can be regarded as a reflection of semantic properties. An SKU can be a single term or a polylexical unit, which includes verbs and verb phrases. This paper describes a method of selecting correspondences between verbal phraseological units in English to Spanish within the specialized domain of the environment.

The underlying idea is that verbs in specialized texts and their argument structure can be classified and organized in a set of conceptual-semantic categories typical of a given specialized domain (Buendía 2013). In this context, when semantic roles and macroroles from the Role and Reference Grammar (Van Valin and LaPolla 1997) are specified as well as the resulting phrase structure, it is then possible to establish templates that represent this meaning for entire conceptual frames. Accordingly, the range of verbs generally associated with a certain category can be predicted within the frame of a specialized event (Faber 2012). In this regard, verbs belonging to the same frame normally have the same number and type of arguments. These arguments have similar semantic characteristics and constrain verb meaning within specialized texts. This makes it possible to establish frame-based correspondences between different languages.

In our study, we performed an in-depth contrastive analysis of the verb phrases in our Spanish-English corpus of specialized texts. Our results showed that verbs within the same semantic category often varied in regards to register. At other times, even though the Spanish and English verb had the essentially same meaning, it was constrained and modulated differently in each language by the semantic characteristics of the specialized arguments. There were even cases where frames had no verbal lexicalizations in the other language, given the differences in the way a certain lexical domain or frame was instantiated.

Verb analysis was first performed in English. The same methodology was subsequently used in Spanish analysis. However, once the semantic categories in each Spanish concordance line were identified, they were directly associated with the frames in English which had the same category labels. This methodology is in consonance with Pimentel (2012) in her assignment of specialized verb equivalents within the legal domain.

References

BUENDÍA CASTRO, M. (2013). *Phraseology in Specialized Language and its Representation in Environmental Knowledge Resources*. PhD Thesis presented at the University of Granada, Spain.

FABER, P. ed. (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin. Boston: Mouton de Gruyter.

FABER, P., and UREÑA, J.M. (2012). Specialized Language Translation. In P. Faber, ed. 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin, Boston: De Gruyter Mouton, pp. 73–92.

NEWMARK, P. (1981). *Approaches to Translation*. Oxford: Pergamon Press.

NIDA, E.A. and TABER, C.R., 1969/1982. *The Theory and Practice of Translation.* Leiden: E. J. Brill.

PIMENTEL, J. (2012). *Criteria for the Validation of Specialized Verb Equivalents: Application in Bilingual Terminography*. Ph.D. Thesis presented at the University of Montreal, Canada. [online] Available at: <http://olst.ling.umontreal.ca/pdf/Pimentel_J_thesis_2012.pdf> [Accessed 22 September 2013].

VAN VALIN, R.D.JR., AND LAPOLLA, R. (1997). *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.

LE TRAITEMENT AUTOMATIQUE DES COLLOCATIONS VERBALES DES NOMS PREDICATIFS DES <AIDES MATERIELLES>

Elena Diego Hernández

Universidad de Salamanca

elenadiegohernandez@usal.es

Nous menons des recherches dans l'élaboration d'un dictionnaire électronique bilingue (français - espagnol) des noms prédicatifs des <aides/ayudas> (*bourse*, *amnistie*, *dispense*, *réconfort*), qui permette le traitement automatique de ces mots. Notre principal propos est l'inscription du mot dans la phrase, pour pouvoir prédire de manière automatique les collocations verbales.

Nous avons considéré que la théorie des classes d'objets (Gaston Gross et le laboratoire *Lexiques, Dictionnaires, Informatique*, CNRS-Université Paris 13) s'adapte bien aux spécificités des automates. Leur notion d'*emploi* nous permet d'analyser le vocabulaire sélectionné (le lexique) en l'inscrivant dans la syntaxe de la phrase (la combinatoire) et en tenant compte aussi de la sémantique (le résultat des éléments lexicaux organisés d'une façon déterminée dans le cadre de la phrase). Ainsi, notre unité minimale d'analyse est la phrase.

Les entrées de notre dictionnaire sont décrites de manière systématique grâce à l'application d'une grille n'analyse. Nous avons retenu pour cette grille deux types de propriétés : des propriétés configurationnelles (le nombre d'arguments sélectionnés par chaque nom prédicatif, leur mode de structuration

dans la phrase et leur nature morphologique et sémantique) et les propriétés combinatoires (les verbes supports, les verbes aspectuels et les verbes prédicatifs appropriés sélectionnés par chaque nom prédicatif).

Cette méthode d'analyse strictement linguistique nous a permis de résoudre un problème auquel nous avons été confronté dès le début : déterminer s'il existe une relation de dépendance hiérarchique des <aides financières> sur les <aides matérielles> (les <aides financières> étant une sous-classe des <aides matérielles>) ou s'il s'agit, au contraire, de deux classes d'objets qui seraient au même niveau d'une arborescence de la classe <aide>.

D'une part, d'un point de vue configurationnel, nous avons observé que la nature des arguments est différente (*N2:de<montant>* pour les <aides financières> ; *N2:de<inc>* pur les <aides matérielles>)[7], ce qui entraine, d'autre part, la seconde différence : les verbes actualisateurs sont, en partie, différents. En effet, le montant d'un *héritage* (<aide financière>) peut être *versé* sur le compte du bénéficiaire, mais les biens matériaux de cet *héritage* (<aide matérielle>) ne peuvent pas être *versés*, ils sont *transmis*. Nous sommes donc face à deux emplois du nom *héritage*, qui constituent deux entrées différentes dans notre dictionnaire.

Nous avons obtenu 5 sous-classes des <aides matérielles> :

CLASSE D'OBJETS	**SCHÉMA D'ARGUMENTS**	**VERBES APPROPRIÉS**	**EXEMPLE**
<aide matérielle : général>	N0:hum, coll_hum/N1:à<hum, coll_hum>/N2:de<inc>	- porter, demander ; - brindar, prestar	soutien/apoyo
<aide matérielle : gouvernementale>	N0:coll_hum/N1:à <coll_hum>/ N2:de<inc>	- fournir, livrer, renforcer ; -enviar, distribuir, reforzar	aide militaire/ayuda militar
<aide matérielle : donation>	N0:hum, coll_hum/N1:à<hum, coll_hum>/N2:de<inc>	- accorder, collecter, faire, octroyer ; - colectar, hacer, realizar	donation/donativo
<aide matérielle : cadeau>	N0:hum, coll_hum/N1:à<hum, coll_hum>/N2:de<inc>/N3:pour<évén>	- apporter, donner, faire, offrir ; - dar, hacer, traer	présent/presente
<aide	N0:hum /N1:à<hum,	- diviser, laisser (en), transmettre (en) ;	héritage /

[7] Le deuxième argument du nom prédicatif étant un montant d'argent ou un inanimé concret.

matérielle : héritage>	coll_hum>/N2:de<inc>	- dejar de/como/en, dividir, legar, transmitir (en)	herencia

Ce dictionnaire permettrait à chaque collocation d'être reconnue en langue source et d'être générée en langue cible par le biais du nom prédicatif qui la sélectionne.

References

BLANCO X. (2001). Dictionnaires électroniques et traduction automatique espagnol-français. *Langages*, 143, p. 49.

BOSQUE. I. y MALDONADO C. (2004). *REDES. Diccionario combinatorio del español contemporáneo*. Madrid: SM.

BUVET P-A. (2009). Quelles procédures d'étiquetage pour la gestion de l'information textuelle électronique ? *L'information grammaticale*, 122, p. 40.

GREZKA A. et BUVET P.-A. (2007). Élaboration d'outils méthodologiques pour décrire les prédicats du français. *Linguisticae Investigationes*, 30, p. 217.

GROSS G. (2010). La notion d'*emploi* dans une grammaire de prédicats. *Cahiers de lexicologie*, 96, p. 97.

GROSS G. (2008). Les classes d'objets. *LALIES*, 28, p. 111.

GROSS G. (1992). Forme d'un dictionnaire électronique. *L'environnement traductionnel*.

LE FUR D. (2007). *Dictionnaire des combinaisons des mots: les synonymes en contexte*. Paris: Dictionnaires Le Robert.

MASSOUSSI T. et SFAR I. (2009). Description des prédicats nominaux : de la langue générale aux langues spécialisées. *Neophilologica*, 21, p. 62.

MEJRI S. (1997). *Le figement lexical : descriptions linguistiques et structuration sémantique*, Tunis: Publications de la faculté des Lettres de la Manouba.

MEL'ČUK I. (1984, 1988, 1992, 1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherche lexico-sémantique* I, II, III et IV. Montréal: Les Presses Universitaires de Montréal.

VIVÈS R. (2001). Quelques remarques sur les classes d'objets, bilingues. *Langages*, 143, p. 120.

LA WEB COMO CORPUS Y BASE DE INVESTIGACIÓN CIENTÍFICA

Heloisa Fonseca
Universidade Estadual Paulista
heloisafonseca25@gmail.com

Esta comunicación pretende demostrar el potencial productivo de las investigaciones fraseológicas que se basan en la *World Wide Web*, con objeto de analizar las ventajas y desventajas que presenta el uso de este medio de investigación; en concreto, este trabajo se centrará en el uso de la *web* como *corpus* lexicográfico y como fuente de búsqueda avanzada, de contextos y de indicadores de uso. Partimos de la confección del BD-FraZoo (Banco de Dados de Fraseologismos Zoônimos, Universidade Estadual de São Paulo) y de algunas investigaciones recientes desarrolladas en la Universidad de Tilburg, de los Países Bajos, que trabajan con el Procesamiento de Lenguaje Natural (PLN) y usan recursos tecnológicos para sus investigaciones lingüísticas. Además, observaremos el uso de *Google* como instrumento para las investigaciones fraseo-lexicográficas, especialmente para confirmar los equivalentes. Para ello, hemos analizado algunas expresiones idiomáticas y proverbios brasileños y franceses (*quem não tem cão caça com gato / faute de griveson mange des merles; matar dois coelhos com uma cajadada só / courir deuxlièvres à la fois; dar bode / faire long feu; estar com a macaca / bouffer du lion*). Los resultados obtenidos en este trabajo revelan algunos aspectos desconocidos, sobre todo culturales, ya que es posible encontrar en Internet un lenguaje sin filtros y más cercano al uso cotidiano. En este sentido, nuestro objetivo consiste en comprobar si Internet, más específicamente los buscadores automáticos, puede ayudar a establecer y confirmar equivalencias

mediante contextos y, sobre todo, mediante la cantidad aproximada de resultados de las búsquedas. Por otra parte, los contextos encontrados se analizan basándonos en la metodología de la lingüística de corpus, Berber Sardinha (2000, 2004), Baker (1995, 1999), Tagnin (2001, 2010), y en la herramienta *Concord* del programa *WordSmith Tools*, que es un conjunto integrado de programas para observar cómo las palabras o expresiones se comportan en los textos. Creemos que es posible desarrollar una investigación científica tomando como base fundamentalmente Internet y aplicativos o demos gratuitos. A partir de los resultados obtenidos en este trabajo, podemos adelantar que el uso tanto de los presupuestos del PLN como de la lingüística de *corpus* ayuda a la elaboración de proyectos lexicográficos al proporcionar usos auténticos de la lengua.

References

BAKER, M. (1995). *Corpora in Translation Studies*: an overview and some suggestions for the future research. *Target*, 7:2, pp. 223-243.

_____. (1999). Lingüística e estudos culturais: paradigmas complementares ou antagônicos nos estudos da tradução? In: MARTINS, M. A. P. (Org.), 1999. *Tradução e multidisciplinaridade*. Rio de Janeiro: Lucerna, pp. 15-34.

BERBER SARDINHA, A. (2000). *Linguística de corpus*: histórico e problemática. *D.E.L.T.A. (*16 (2), pp. 323-367.

_____. (2003). Uso de corpora na formação de tradutores. *D.E.L.T.A. (*19: Especial, pp. 43-70.

_____. (2004). *Linguística de Corpus*. São Paulo: Monole.

TAGNIN, S. E. O. (2001. *COMET: um Corpus Multilíngüe para Ensino e Tradução*. São Paulo: USP.

_____. (2002). Os Corpora: instrumentos de auto-ajuda para o tradutor. In *Cadernos de Tradução* IX. Florianópolis. [Disponible en: <http://www.periodicos.ufsc.br/index.php/traducao/article/view/5986/5690>].

_____. (2004). *Corpora: o que são e para quê servem*. [Disponible en: <http://www.fflch.usp.br/dlm/comet/Novo/Lexicografia.pdf>].

TEIXEIRA, E. D. (2006). *Como usar o Wordsmith Tools*. V.3. São Paulo: Universidade de São Paulo.

XATARA, M. C. (2008). A Web para um levantamento de freqüência. UNESP, São José do Rio Preto. [Disponible en: http://www.filologia.org.br/ileel/artigos/artigo_398.pdf].

SYNTACTIC ENCODING OF VERBAL PHRASEMES IN A LARGE-SCALE VALENCE DICTIONARY OF POLISH

Elżbieta Hajnicz
Polish Academy of Sciences
hajnicz@ipipan.waw.pl

Agnieszka Patejuk
Polish Academy of Sciences
aep@ipipan.waw.pl

Adam Przepiórkowski
Polish Academy of Science
adamp@ipipan.waw.pl

Marcin Woliński
Polish Academy of Sciences
wolinski@ipipan.waw.pl

The first aim of this paper is to present a comprehensive formalism for encoding the syntax of phraseological expressions headed by verbs. The proposed formalism is an extension of the rich notation used to describe valence properties of Polish predicates in the Walenty dictionary (http://zil.ipipan.waw.pl/Walenty; Przepiórkowski *et al.* 2014).

An example of a phraseological valence schema for WITAĆ 'welcome' is given below:

(1) subj{np(str)} + obj{np(str)} +
{lex(prepnp(z,inst),pl,XOR('ramię','ręka'),
ratr1({lex(adjp(agr),agr,agr,'otwarty',
atr1({lex(advp(mod),'szeroko',natr)}))})}

There are three arguments in (1), enclosed in curly brackets and separated by +: the structurally-cased (usually nominative) NP subject (subj{np(str)}), the structurally-cased (normally accusative) object (obj{np(str)}) and a lexicalised argument (lex). The lex metacategory takes any base category used in Walenty

(e.g., np for a nominal phrase or prepnp for a prepositional phrase) as the first parameter, followed by parameters imposing constraints appropriate for the relevant base category and finally the displayed modification pattern. Here, the lexical argument is a prepositional phrase (prepnp) headed by the preposition Z 'with' which takes an instrumental (inst) NP in the plural (pl). This NP must be headed by either the plural form of RAMIĘ 'arm' or RĘKA 'hand'. The final parameter of this lex specification expresses the complex information that this head noun *must* be modified (ratr1) by an adjectival phrase agreeing with the noun in case (adjp(agr)), as well as in number and gender (the other two occurrences of agr), and headed by an appropriate form of the adjective OTWARTY 'open'. This adjective *may* in turn be modified (atr1) by an adverbial phrase headed by SZEROKO 'widely', which may not be modified any further (natr). In summary, the valence schema in (1) describes the Polish versions of the phraseological expression *somebody welcomes somebody with arms wide open* in a rather detailed way.

The second aim is to investigate whether the considerable expressive power of this formalism is necessary and sufficient to represent such phrasemes in general, cross-linguistically; after all, other lexica combining valence and phraseological information often make do with simpler dependency formalisms (Mel'čuk and Zholkovsky, 1984; Mel'čuk, 2006; Urešová, 2009). We show that all this expressive power is not only necessary, but in fact it is still insufficient to describe the variety verbal phraseological expressions. One area of possible improvement concerns word order: even in a relatively free word order language such as Polish, where phrasemes do not in general have a fixed word order, some expressions require positing linearisation constraints. The second area concerns the flexibility of apparently fixed to clichés such as *The fat is in the fire* 'trouble is about to start', which cannot be encoded as immutable and syntax-less strings, given that a raising verb may enter their syntactic (and semantic) structures, as in *The fat seems to be in the fire*. Finally, we give Polish examples of constructions which violate the continuity constraint of O'Grady (1998, 284) and which are problematic for all formalisms we are aware of.

References

Mel'čuk, I. (2006). Explanatory combinatorial dictionary. In G. Sica, editor, *Open Problems in Linguistic and Lexicography*, pages 225–355. Polimetrica, Monza.

Mel'čuk, I. and Zholkovsky, A. (1984). *Explanatory Combinatorial Dictionary of Modern Russian. Semantico-syntactic Studies of Russian Vocabulary*. Wiener Slawistischer Almanach, Vienna.

O'Grady, W. (1998). The syntax of idioms. *Natural Language and Linguistic Theory*, **16**, 279–312.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F., and Świdziński, M. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 2785–2792, Reykjavík, Iceland. ELRA.

Urešová, Z. (2009). Building the PDT-VALLEX valency lexicon. In *On-line Proceedings of the fifth Corpus Linguistics Conference*. University of Liverpool.

DO WE NEED EQUIVALENCE-BASED E-TOOLS?

Maciej Jaskot

University of Social Sciences and Humanities, Poland

mjaskot@swps.edu.pl

Recently a lot of effort has been put into creation of lexicographic e-tools such as bi- or multilingual dictionaries, phraseological lexicons and parallel corpora. In order to make them efficient it is necessary to address such theoretical issues as the formulation of linguistically grounded principles of selection of phraseological units, their translation, semantic interpretation based on differentiation of the linguistic image of the world. We cannot forget about linguistic ambiguities and, last but surely not least, the equivalence issue (Luchyk, A., & Antonova, O., 2012; Ivanov, A. O., 2006). Defining the notion of equivalence, or taking into consideration one of the possible interpretations of this concept, is the first step to prepare a base of equivalent units. The second step would be to reconsider the notion of the phraseological unit (do we have equivalent praseological units?) as equivalence seems to be possible at many levels, and the linguistic one does not solve the problem of translating (transposing) a more complex word unit (as a phraseological unit, for example). What we see while comparing a translated form is an oncoming meaning, but often the cultural background is missed. That is why rethinking equivalence through the notion of cultureme (Bukhonkina, A., 2002; Gak, V. G., 1998; Vorob´ëv, V. V., 1997) would be useful and helpful to answer the question what shall we demand from modern e-lexical tools.

References:

Bukhonkina, A. (2002) *Tipy asymetrii kul´turem: Ha materiale frantsuzkogo i russkogo iazykov.* Волгоград.
Gak, V. G. (1998) *IAzykovye prebrazovaniia*. Moskva: IAzyki russkoĭ kul´tury.
Ivanov, A. O. (2006) *Bezėkvivalentnaia leksika.* Sankt Peterburg: Izdatel´stvo SpbGU.
Luchyk, A., & Antonova, O. (2012). *Pol'ʼko-ukraïns'kyĭ slovnyk ekvivalentiv slova.* (A. Kisiel & V. Koseska-Toszewa, Eds.). Kyïv: Ukraïnsʼkyĭ movno-informatsiĭnyĭ fond NAN Ukraïny, Natsional'nyĭ Universytet «Kyievo-mohylians'ka Akademiia», Instytut Slavistyky Pol's'koï Akademiï Nauk
Vorob´ëv, V. V. (1997) *Lingvokul´turologiia (teoriia i metody).* Moskva: Izd-vo MGU.

FRASEOGRAFÍA Y LINGÜÍSTICA DE CORPUS: SOBRE EL TRATAMIENTO DE LOCUCIONES VERBALES EN LA NUEVA EDICIÓN DEL *DICCIONARIO DE LA LENGUA ESPAÑOLA*

Jorge Leiva Rojo

Universidad de Málaga

leiva@uma.es

En los últimos años han aparecido multitud de trabajos en los que se aborda el estudio de la fraseología tanto en repertorios bilingües —caso de Corpas Pastor (1996) o Santamaría Pérez (1998, 2001)— como monolingües —sirvan de ejemplo las contribuciones de Ettinger (1982), Carneado Moré (1985), Montoro del Arco (2004) o de Oliveira Silva (2007)—. En todos ellos se pone de manifiesto con frecuencia la necesidad imperiosa de contar con el apoyo de la lingüística de corpus para crear diccionarios en los que el tratamiento fraseológico se base en evidencias reales y no en suposiciones o en ejemplos creados por el lexicógrafo de turno.

Las unidades fraseológicas, que según algunos estudiosos es el nivel más alto de conocimiento posible de toda lengua, han recibido tradicionalmente un tratamiento dispar en los repertorios lexicográficos, bien por no incluir lemas en los que se refleje la variación fraseológica, bien por no acotar con precisión su uso —por ejemplo, si han caído en desuso o si sin propias de usos coloquiales o incluso vulgares de la lengua — o bien por ofrecer significados que no se ajustan a la realidad o que la acotan de forma parcial.

El objetivo de nuestro trabajo es doble: de una parte, pretendemos analizar la forma en que ha evolucionado el tratamiento lexicográfico que se da en la

nueva versión del *Diccionario de la lengua española*, de la Real Academia Española, en comparación con la versión anterior, la vigésimo segunda. Para ello, seleccionaremos algunas de las locuciones verbales más significativas, cuyo tratamiento se compararán a su vez con el que se les da en otros repertorios lexicográficos, tanto bilingües como monolingües. De otra parte, mediante el uso los de corpus de textos electrónicos de libre acceso —como, por ejemplo, el *Corpus del español del siglo XXI*, el *Corpus de referencia del español actual* o el *Corpus del español*— se intentará establecer cuál es el uso real que tienen tales locuciones verbales en la lengua que recogen los corpus analizados. Se pretende, en definitiva, comprobar si los corpus de textos *sancionan* en la práctica el uso teórico que se les presupone en los diccionarios a las locuciones verbales en cuestión. Finalmente, se propondrán soluciones de mejora para aquellos lemas analizados, tanto en lo referente a la variación fraseológica como a las acepciones que se ofrezcan de ella y a los ejemplos que se proporcionen, con objeto de sacar el máximo partido a las situaciones reales de la lengua que contienen los corpus de textos empleados como soporte.

References

ETTINGER, Stefan (1982). Formación de palabras y fraseología en la lexicografía. En: G. Haensch *et al.*, eds. *La lexicografía. De la lingüística teórica a la lexicografía práctica*. Madrid: Gredos, p. 233-258.

SANTAMARÍA PÉREZ, María Isabel (1998). El tratamiento de las unidades fraseológicas en la lexicografía bilingüe. *ELUA. Estudios de Lingüística* (12), p. 299-318.

CORPAS PASTOR, Gloria (1996). La fraseología en los diccionarios bilingües. In: Manuel Alvar Ezquerra, ed. *Estudios de Historia de la Lexicografía del Español*. Málaga: Servicio de Publicaciones de la Universidad, p. 167-182.

MONTORO DEL ARCO, Esteban Tomás (2004). La variación fraseológica y el diccionario. *De Lexicografía*, p. 1000-1014.

OLÍMPIO DE OLIVEIRA SILVA, Maria Eugênia. (2007). *Fraseografía teórica y práctica*. Fráncfort del Meno: Peter Lang.

CARNEADO MORÉ, Zoila V. (1985). *La fraseología en los diccionarios cubanos*. La Habana: Editorial de Ciencias Sociales.

SANTAMARÍA PÉREZ, María Isabel. (2001). *Tratamiento de las unidades fraseológicas en la lexicografía bilingüe español-catalán*. Alicante: Biblioteca virtual Miguel de Cervantes.

FRASEOLOGÍA, DICCIONARIOS ELECTRÓNICOS Y CORPUS: A LA BÚSQUEDA DE LA EQUIVALENCIA

Pedro Mogorrón Huerta

Universidad de Alicante

pedro.mogorron@ua.es

La fraseología española conoce sin lugar a dudas una época dorada como lo atestiguan los numerosos libros, artículos, jornadas, congresos, tesis doctorales e incluso revistas que en los últimos años están floreciendo debido al interés cada vez más grande que despiertan todos y cada uno de los diferentes tipos de Unidades fraseológicas (UF). Sin embargo se ha observado que el contenido fraseológico de numerosos diccionarios referenciales analizados es muy deficiente pues incluye un número limitado de UF sin especificar las razones de esa selección (expresiones importantes por su contenido cultural, por su vigencia o frecuencia de uso, etc.). Por otro lado y también recientemente, ha hecho su aparición en el mundo de la lingüística otro tipo de diccionario cuya elaboración está relacionada con diferentes aplicaciones al tratamiento automático de textos en el campo de la lingüística aplicada. Se trata de los diccionarios electrónicos. Si bien en algunos casos no está tan claro si realmente, son un avance respecto del diccionario en papel, superan con creces la capacidad informativa y de espacio del diccionario en soporte tradicional y, de los diccionarios digitales o de los diccionarios que van incluidos en los procesadores de texto. Parte de la información de estos diccionarios se puede completar con la información que contienen los corpus

textuales: frecuencia de uso, posibles variantes, contextos de uso, etc.). Toda esta información es más que importante de cara a la búsqueda de la equivalencia fraseológica que ya no puede tratarse siguiendo el tratamiento de equivalencias ofrecido por los diccionarios bilingües que se limitan a ofrecer cuando lo hacen un equivalente fraseológico, cuando muchas veces para contenidos muy usuales existen numerosas UF para sinónimas, y sin tener en cuenta criterios tan importantes como la frecuencia de uso y las creaciones diatópicas presentes en alguno de los países que usan una misma lengua oficial como puede ser el caso con lenguas como el español, el portugués, el francés, el inglés. Deseamos presentar los trabajos de elaboración de un diccionario electrónico de expresiones fijas del español peninsular e hispanoamericano con la ayuda de corpus textuales que ofrece equivalentes traductológicos en catalán, francés, inglés, árabe, italiano, etc.

NUTZERGENERIERTE INTERNETWÖRTERBÜCHER UND IHRE ANWENDUNG IN DER LEXIKOGRAPHIE

Elena Krotova

Lomonosov Moscow State University

elena_krotova@inbox.ru

Im Vortrag werden mehrere Internetwörterbücher analysiert, die nicht nur von den professionellen Lexikographen, sondern auch von Nutzern selbst generiert bzw. erweitert werden. Dabei wird auf folgende Fragen eingegangen:

- Welche Informationen über Idiome sind in solchen Wörterbüchern zu finden?
- Wie können Nutzer zur Entwicklung der Internetressource beitragen?
- Können solche nutzergenerierte Wörterbücher für Lexikographen von Nutzen sein (diese Frage wird am Beispiel der deutsch-russischen Phraseographie erläutert)?
- Sind solche Ressourcen Deutschlernenden zu empfehlen?

Analysiert werden (1) Ressourcen für deutsche Phraseologie (darunter *Redensarten-Index*, *Redensarten*, *Deutsche Redewendungen* auf der Web-Seite *Wiktionary*) und (2) universelle Internetwörterbücher (*dict.cc* und *Glosbe*). Bei jedem der genannten Nachschlagewerke werden Informationen zum Ziel des Projekts und dem Beitrag des Nutzers sowie zum Aufbau des Artikels gegeben.

Bei allen diesen Internetwörterbüchern finden sich viele Kritikpunkte. Vgl. dazu u.a.:

- meist fehlt die Lemmatisierung und die Suche ist nur nach einer strikten Folge von Zeichen möglich, was sich negativ auf die Suchergebnisse auswirkt. So finden sich nur wenige Belege, auch wenn die Ressource über größere Parallelkorpora verfügt (wie bei *Glosbe*),
- Wörterbuchartikel sind oft schlecht aufgebaut und bieten zu viele Übersetzungsvarianten bzw. Paraphrasen, ohne sie zu strukturieren (z.B. neun aufeinanderfolgende Paraphrasen für *jmdm. aufs Dach steigen* bei *Redensarten*, wobei es schlicht mit *jmdn. zurechtweisen, in die Schranken weisen* wie bei *Duden* erklärt werden kann),
- Idiom-Varianten bleiben meist unbeachtet. Im besten Fall werden sie als verschiedene selbstständige Wörterbuchartikel behandelt. Dasselbe betrifft manchmal sogar die Valenz der Idiome (z.B. zwei Artikel bei Redensarten-Index *etwas in Fahrt bringen* und *jemanden in Fahrt bringen*).

Solche Wörterbücher können aber in mehreren Hinsichten nützlich sein. So finden sich bei *Redensarten-index* einige Idiome, die in professionellen Wörterbüchern (z.B. in *Duden-Redewendungen*) nicht verzeichnet sind, obwohl sie gebräuchlich zu sein scheinen (z.B. *fein heraus putzen* (*jmdn., etw. A, sich*), 97 Belege im Deutschen Referenzkorpus) oder wird die Semantik mancher Idiome besser beschrieben. Am wichtigsten ist aber, dass nutzergenerierte Wörterbücher dem Lexikographen zeigen können, was sich Nutzer eigentlich von einem Internetwörterbuch wünschen und was ihnen möglicherweise bei den vorhandenen professionellen Online-Nachschlagewerken noch fehlt (z.B. bequemere Suche, mehr Beispiele bzw. Belege, Vertonung von Lemmata usw.). Aussichtsreich scheint die Idee ein Internetwörterbuch mit Parallelkorpora zu kombinieren, wie es bei *Glosbe* gemacht wurde. Dabei ist die Qualität der Korpora bei *Glosbe* eher niedrig im Gegensatz zum ähnlichen Projekt *Linguee*, wo man mithilfe von Machine Learning eine viel höhere Qualität der automatisch erstellten Parallelkorpora erreichte. Zusammenfassend lässt sich sagen, dass auch wenn nutzergenerierte Wörterbücher aus der Sicht von der professionellen Lexikographie weit unter dem vertretbaren Niveau

liegen, geben sie Aufschluss darüber, was von einem Internetwörterbuch von Nutzern erwartet wird und was in diesem Bereich noch gemacht werden kann.

References

Redensarten-index. [online] Available at: <http://www.redensarten-index.de/suche.php> [Accessed 31 March 2015].

Redensarten. [online] Available at: < http://www.redensarten.net/> [Accessed 31 March 2015].

Deutsche Redewendungen. [online] Available at: < http://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Redewendungen> [Accessed 31 March 2015].

Wörterbuch für Russisch-Deutsch und andere Sprachen. [online] Available at: < http://deru.dict.cc/> [Accessed 31 March 2015].

Glosbe - das mehrsprachige Online-Wörterbuch. [online] Available at: < https://de.glosbe.com/> [Accessed 31 March 2015].

Linguee. Wörterbuch Englisch-Deutsch und Suche in einer Milliarde Übersetzungen. [online] Available at: < http://www.linguee.de/> [Accessed 31 March 2015].

THE COMPILATION OF AN ONLINE CORPUS-BASED BILINGUAL COLLOCATIONS DICTIONARY

Adriane Orenha-Ottaiano
Universidade Estadual Paulista Júlio de Mesquita Filho
adriane@ibilce.unesp.br

The contribution of corpora to lexicology/lexicography, as well as to phraseology/ phraseography, has already been pointed out by many researchers (Altenberg & Granger 2002; Burger et al. 2007; Halliday et al. 2004; Meunier & Granger 2008; Moon 2008; Orenha-Ottaiano 2013; Sinclair 2007; Teubert 2004, 2007, among others). Moon (2008), for instance, stated that corpora is "the great facilitator for the description of phraseology, and its use in dictionary-making has heavily influenced the ways in which lexicographical attitudes towards phraseology have developed over the last twenty-five years". Moreover, it is also true that the quality of monolingual and bilingual dictionaries has also improved, due to the methodology provided by Corpus Linguistics. The use of corpora has enabled us to identify and extract phraseological units more easily and quickly, especially collocations. Taking that into account, and considering the fact that collocations pose a serious problem to foreign language learners and trainee translators with regard to production (oral or written), our aim is to compile an online bilingual collocations dictionary, in the English-Portuguese and Portuguese-English directions, based on learner, parallel and reference corpora, focusing on all types of collocations (verbal, noun, adjectival and adverbial collocations), and designed for teachers and

learners of English as a second language as well as learner and professional translators, among others, in order to help the referred audience use them more accurately and productively. The methodology first relies on the extraction and analysis of collocations from a *Translation Learner Corpus* made up of newspaper articles taken from well-known Brazilian newspapers and magazines. The typology of the texts is related to current world news such as *Financial crises in Europe*; *Unemployment*; *Elections in the US*; *Bullying*; *Marijuana Legalization* etc. The referred corpus was compiled at *Universidade Estadual Paulista* (UNESP), in Brazil, and the collocations were extracted with the help of *WordSmith Tools* (Scott 2008). In a second stage, based on the keywords and collocations analyzed in the learner corpora, more collocational patterns are extracted with the help of *The Corpus of Contemporary American English* (Davies 2008-2012) – COCA. In order to include more collocations as well as to ensure dictionary users will have access to more frequent, recurrent and sophisticated collocations, we also use the frequency list from COCA to extract more patterns. The idea of having the proposed dictionary in online format will allow us to incorporate collocational information more qualitatively and quantitatively. Besides that, more examples can be included, different from conventional collocations dictionaries. Being the first bilingual collocations dictionary, we hope to achieve the challenge of meeting learners' collocational needs as the collocations will be selected according to Brazilian learners' difficulties regarding the use of collocations.

References

ALTENBERG, B. & GRANGER, S. (2002). *Lexis in Contrast*. Amsterdam/ Philadelphia: Benjamins.

BURGER, H., D. DOBROVOLSKIJ, P. KÜHN AND N. R. NORRICK (eds). (2007). *Phraseologie/Phraseology*. Ein internationales Handbuch zeitgenössischer Forschung/An International Handbook of Contemporary Research. 2 Halbbände (= HSK 28.1/2). Berlin/New York: de Gruyter.

DAVIES, M. (2008-2012). *The Corpus of Contemporary American English:* 425 million words, 1990-present. Available: <http://corpus.byu.edu/coca/.>. Accessed: April 20th, 2012.

HALLIDAY, M.A.K., W.TEUBERT, C. YALLOP & A. ČERMÁKOVÁ (2004). *Lexicology and Corpus Linguistics*: an introduction. London: Continuum.

MEUNIER, F., & S. GRANGER. (2008). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins.

MOON. R. (2008). Dictionaries and collocations. In: Meunier, F., & S. Granger (eds.) (2008). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins. 247-252.

ORENHA-OTTAIANO, A. (2013). The proposal of an electronic bilingual dictionary based on corpora. In *X International School on Lexicography*, Florença, Itália. Life Beyond Dictionaries. Ivanovo: University of Ivanovo. 1: 405-408.

SINCLAIR, J. MCH. (2007). Data-derived multilingual lexicons. In: Teubert, W. (Ed.). 2007. *Text corpora and multilingual lexicography*. Amsterdam/Philadelphia: John Benjamins, 69-82.

SCOTT, M. (2008). *WordSmith Tools*, version 5.0. Liverpool: Lexical Analysis Software Ltd.

TEUBERT, W. (2004). Language and Corpus Linguistics. In Halliday, M.A.K., W.Teubert, C. Yallop & A. Čermáková (2004). *Lexicology and Corpus Linguistics*: an introduction. London: Continuum.

TEUBERT, W. (Ed.). (2007). *Text corpora and multilingual lexicography*. Amsterdam/Philadelphia: John Benjamins.

POUR UN CONTINUUM DES PHRASÈMES NON-COMPOSITIONNELS

Marie-Sophie Pausé

ATILF-CNRS & Université de Lorraine

marie-sophie.pause@orange.fr

Notre contribution sera consacrée au statut lexical des phrasèmes non-compositionnels. Bien qu'il soit aujourd'hui largement admis qu'une unité phraséologique comme BRISER LA GLACE ' mettre fin à une situation gênante' est une locution, tandis qu'une unité comme LANCE-FLAMMES ' arme destinée à lancer du feu' est un mot composé, de nombreux cas (comme POMME DE TERRE) restent moins, voire pas du tout consensuels, classés tantôt parmi les locutions, tantôt parmi les mots composés.

BRISER LA GLACE et POMME DE TERRE sont toutes deux formées d'une manière analogue à des combinaisons libres, respectivement comme les collocations casser le vase et livre de français. Il s'agit donc de syntagmes (Kahane 2008 : 2539) non-compositionnels (Anscombre 2011, Legallois et Tutin 2013). Néanmoins, toutes deux n'admettent pas le même nombre de variations syntagmatiques et paradigmatiques. BRISER LA GLACE peut être passivée (1) et subir l'insertion de modifieurs (2), tandis que POMME DE TERRE n'admettra pas ou peu de variations, se comportant ainsi comme un nom simple.

(1) Dans le fond, Marie-Thérèse (Charlotte pour les intimes) est d'une grande sensibilité et timidité qu'elle compense en se montrant un peu fière avec les gens qu'elle ne connaît pas. Mais une fois que la glace est brisée, elle se révèle réellement adorable.

(FrWac: http://pilleul.sylvain.club.fr/les_enfants_de_france.html)

(2) Cassie tourna la tête, et vit Peyton… une jeune fille qu'elle connaissait de loin. Elle lui fit tout de même un sourire, histoire de briser un peu la glace.

(FrWac: http://do-your-life.jeun.fr/plage-f47/seule-au-crepuscule-liiibre--t303.htm)

Une approche privilégiant le comportement de l'unité dans la phrase considérera POMME DE TERRE comme un mot composé (Gross 1996 : 25-59, Corbin 1997) ; tandis qu'une approche contraire, considérera POMME DE TERRE comme une locution (Polguère 2008 : 51-58). Cette dernière approche applique rigoureusement la définition du mot composé prototypique, qui consiste en un « lexème construit à partir de lexèmes selon un mode d'organisation qui n'est pas syntaxique » (Villoing 2003 p. 185). Par opposition, la locution est un syntagme lexicalisé.

Nous développons actuellement, suivant suivant les principes de la Lexicologie Explicative et Combinatoire (Clas et coll. 1995), un modèle de description lexico-syntaxique des locutions françaises au sein du Réseau Lexical du Français (Lux-Pogodalla & Polguère 2011), qui nous permettra à terme de prévoir les variations des locutions en discours. Il nous est pour cela indispensable de distinguer les locutions dont l'origine syntagmatique est encore fortement présente et laisse des marques sur leurs emplois, de celles qui se comportent comme des unités simples. Nous proposons à ce titre, en nous appuyant sur la classification des phrasèmes opérée par I. Mel'čuk (2013), un continuum des phrasèmes non-compositionnels, basé sur la nature des liens entre les unités constituantes : liens morphologiques véritables (TELEPHONE), liens morphologiques potentiellement d'origine syntaxique (OUVRE-BOITE), liens syntaxiques morphologisés (QU'EN DIRA-T-ON), liens syntaxiques non opérants (POMME DE TERRE), et liens véritablement syntaxiques (BRISER LA GLACE), en passant par des cas intermédiaires comme CHEZ SOI ou A QUI MIEUX MIEUX.

References

ANSCOMBRE, J.-C. (2011). Figement, idiomaticité et matrices lexicales. In: *Le figement linguistique: la parole entravée*, Paris: Honoré Champion. pp.17-40.

CLAS, A. MEL'CUK, I. AND POLGUERE, A. 1995. *Introduction à la lexicologie explicative et combinatoire*, Bruxelles: Duculot.

Corpus internet *FrWac*. Available at: <http://nl.ijs.si/noske/wacs.cgi/corp_info?corpname=frwac> [Accessed 26 mars 2015].

GROSS, G. (1996). *Les expressions figées en français. Noms composés et autres locutions.* Paris: Ophrys.

KAHANE, S. (2008). Les unités minimales de la syntaxe et de la sémantique: le cas du français. *Congrès Mondial de Linguistique Française – CMLF'08,* Paris: Institut de Linguistique Française, pp.2531-2550.

LEGALLOIS, Dominique, TUTIN, Agnès (2013). Vers une extension du domaine de la phraséologie (présentation). *Langages*, 189, pp.3-25.

LUX-POGODALLA, V. AND POLGUERE, A. (2011). Construction of a French Lexical Network: Methodological issues. *Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011, An ESSLLI 2011 Workshop*, pp.54-61.

MEL'CUK, Igor (2013). Tout ce que nous voulions savoir sur les phrasèmes, mais…, *Cahiers de lexicologie*, 102, pp. 129-149.

POLGUERE, A. (2008). *Lexicologie et sémantique lexicale. Notions fondamentales*. Montréal: Presses universitaires de Montéral.

VILLOING, Florence (2003). Les mots composés VN du français: arguments en faveur d'une construction morphologique, *Cahiers de Grammaire*, 28, pp. 183-196.

THE PROBLEM OF LEMMATISATION IN THE POLISH INFLECTIONAL DICTIONARY OF VERBAL MWES[8]

Sebastian Przybyszewski
University of Warmia and Mazury in Olsztyn
sebastian.przybyszewski@wp.pl

Monika Czerepowicka
University of Warmia and Mazury in Olsztyn
czerepowicka@gmail.com

Iwona Kosek
University of Warmia and Mazury in Olsztyn
i.kosek@uwm.edu.pl

Introduction: Pointing out the problems referring to the choice of the headword for a verbal group of MWEs in the emerging Polish electronic inflectional dictionary and discussion of the possible solutions are the aims of the presentation. The headword ought to be not only a convenient technical solution but also it should provide the user with the most extensive information. The shape of the headword influences also an amount of graphs by which MWEs are coded in the dictionary (Toposlaw lexicographical application uses Multiflex formalism and Unitex graph editor).

Infinitive as a headword

In the modern lexicography, it is widely accepted to use infinitive as the standard form of lemmatisation. Yet the infinitive is not always an apposite choice in inflectional languages because it is not a representative form of a verb and, consequently, of a verbal phraseological unit. The following problems appear:

a) invariability of the infinitive

[8] The paper is connected with the realization of a project funded on the basis of the decision No. DEC-2013/09/B/HS2/01222, by the National Science Centre in Poland.

The infinitive form does not provide any information if the MWE is an inflective phraseological unit, e.g. one cannot differentiate between *palce lizać* (‘expression used to express that sth is scrumptious’; lit. *to lick one’s fingers*), *zbijać bąki* (lit. *to shoot down bitterns*;‘to fool around’). Although both of them have the same inner syntax (V + N_{acc}), the first one is non-inflected and the second one has a full paradigm (*zbijam bąki, zbijasz bąki...*). They are in fact two different kinds of phraseological units.

b) ambiguity

The headword given as an infinitive does not enable to notice the differences in meaning and the differences in number of slots between MWEs’s e.g. koń/pojazd rusza z kopyta (a horse/a vehicle gets off vigorously; lit. sth gets off from the hoof) vs. coś rusza z kopyta (sth begins to happen quickly).

c) the lack of infinitive

Some expressions do not have infinitive due to the properties of grammatical components or the meaning of a whole MWE, e.g. gdzieś siekierę można powiesić (you can cut the air with a knife; lit. one can hang the axe somewhere).

3rd person of singular as a headword

The 3rd person acts much better as a headword of MWEs in Polish for it provides a user with more information:

- it indicates slots in MWEs and the way they should be filled e.g. *ktoś nie ma życia gdzieś* (‘sb is not accepted’; lit. sb does not have life somewhere);
- thanks to the use of a form of tense and mood of a verbal component it is possible to indicate the most typical form e.g. *ktoś robi komuś wodę z mózgu* (‘sb misleads sb’; lit. sb makes water of sb’s else brain; typically in the present tense).

Noting verbal MWEs in the 3rd form causes also several complications. For instance, *ktoś* (sb) and *coś* (sth) are in agreement with singular, hence there occurs a problem when the subject place in a verbal MWE must be filled with plural. However, there are some solutions to cope with that.

References

BIEŃ, J. S. (1991). Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji, vol. 383, *Rozprawy Uniwersytetu*

Warszawskiego, Warszawa: Wydawnictwa UW. Available at: <http://bc.klf.uw.edu.pl/12/> [Accessed 15 November 2013].

CZEREPOWICKA, M., KOSEK, I., PRZYBYSZEWSKI, S. (2014). O projekcie elektronicznego słownika odmiany frazeologizmów czasownikowych. *Polonica*, XXXIV, pp. 115-123.

COWIE, A. P. ed. (1998). *Phraseology: theory, analysis, and applications*. Oxford: Clarendon Press.

KOSEK, I. (2013). Paradygmaty zwrotów frazeologicznych – problemy opisu leksykograficznego. In: G. Dziamska-Lenart, J. Liberek, eds. *Perspektywy współczesnej frazeologii polskiej. Między teorią a praktyką leksykograficzną*. Poznań: Wydawnictwo UAM, pp. 51-61.

LEWICKI, A. M. (1986). Składnia związków frazeologicznych. *Biuletyn PTJ*, XL, pp. 75-82.

MEL'CUK, I. (1995). Phrasemes in language and phraseology in llinguistics. In: M. Everaert, E-J. van der Linden, A. Schenk and R. Schreuder, eds., *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum Associates, pp. 167-232.

SALONI Z., GRUSZCZYŃSKI W., WOLIŃSKI M., WOŁOSZ R., SKOWROŃSKA D. (2014). *Słownik gramatyczny języka polskiego*. 2nd ed. Warszawa.

SALONI, Z. ed. (1988). *Studia z polskiej leksykografii współczesnej*, Wrocław – Warszawa – Kraków – Gdańsk: Ossolineum.

SALONI Z. (2000). *Wstęp do koniugacji polskiej*, Olsztyn: Wydawnictwo UWM.

SALONI, Z. (2001). *Czasownik polski*, Warszawa: Wiedza Powszechna.

SAVARY, A. (2009). Multiflex: a Multilingual Finite-State Tool for Multi-Word Units. In: S. Maneth (ed.), *CIAA 2009*. Berlin Heidelberg: Springer-Verlag, pp.237-240.

SAVARY A. et al. (2010). Computational Lexicography of Multi-Word Units: How Efficient Can It Be? Accepted to: Proceeding of Multi-Word Units: from Theory to Applications (MWE'10), Workshop at the International Conference on Computational Linguistics (COLING'10), Beijing, China (presented on 28.08.2010; manuscript).

The Homepage of the project VERBEL. Available at: <http://uwm.edu.pl/verbel/> [Accessed 20 March 2015].

TOKARSKI J. (2000). *Fleksja polska*, Warszawa: PWN.

WĘGRZYNEK, K. et al. (2012). Opis jednostek nieciągłych w Wielkim słowniku języka polskiego PAN. *Język Polski*, XCII (5), pp. 353-367.

EIN NEUES PHRASEOLOGISCHES ONLINE-WÖRTERBUCH FÜR SPANISCH ALS FREMDSPRACHE

Stefan Ruhstaller

Universidad Pablo de Olavide

sruhkuh@upo.es

Im Vergleich mit der auf das Englische bezogenen phraseologischen Lexikographie stehen für das Spanische zur Zeit noch wenige Werke zur Verfügung. Ein Desideratum stellt insbesondere ein frei zugängliches elektronisches Wörterbuch für das Erlernen des Spanischen als Fremdsprache dar. Um diese Lücke zu schließen wird derzeit das in der vorliegenden Studie besprochene spezifische phraseologische Online-Wörterbuch erarbeitet, dessen Ziel in erster Linie nicht die möglichst zahlreiche Erfassung der im Spanischen gebräuchlichen phraseologischen Einheiten, sondern vielmehr die präzise und didaktisch überzeugende Beschreibung und Charakterisierung einer nach objektiven Kriterien getroffenen Auswahl solcher Elemente ist. Konsequenterweise enthält deshalb die den einzelnen Artikeln zu Grunde liegende Mikrostruktur eine reiche Auswahl von Daten zu den lemmatisierten pluriverbalen Einheiten: Der Benutzer kann ihr Information sowohl zur diatopischen als auch zur diaphasischen und diastratischen Verbreitung der phraseologischen Elemente entnehmen; des Weiteren kann er in ihr pragmatische und kulturelle, für die Verwendung in der Rede äußerst nützliche Anweisungen finden; eine Fülle von in diskursiven Kontext eingebetteten Satzbeispielen runden diese Information ab. Besonders innovativer Natur sind schließlich die Definitionen, welche, im Unterschied zum traditionellen Muster,

das Lemma in einer syntaktischen Sequenz eingefügt präsentieren und so weitere für den kreativen Gebrauch wertvolle Information liefern. Anhand zahlreicher Beispiele wird schließlich aufgezeigt, wie das Wörterbuch nach seiner Vollendung praktisch zugänglich und nutzbar gemacht werden soll.

References

ALEXANDER, R. J. (1992). Fixed expressions, idioms and phraseology in recent English learner's dictionaries. In: Tommola, H / Varanto, K. / Salmi-Tolonen, T. /y Schopp, J. (eds.), *Euralex '92 Proceedings. Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere, Finland, 4-9 August 1992*, Tampere: Department of Translation Studies - University of Tampere, 1992, pp. 35-42.

BURGER, H. (1989). Phraseologismen im allgemeinen einsprachigen Wörterbuch. In: Hausmann, Franz Josef; Oskar Reichmann; Ernst Wiegand y Ladislav Zgusta (eds.), *Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie*,1, De Gruyter, Berlin-New York: De Gruyter, pp. 593-599.

MCALPINE, J. / JOHANNE M. (2003). Capturing phraseology in an online dictionary for advanced users of English as a second language: a response to user needs, *System*, 31, 1, pp. 71-84.

LA PRESENCIA DE COLOCACIONES ESPECIALIZADAS EN LAS BASES DE DATOS Y LOS DICCIONARIOS ELECTRÓNICOS

María Isabel Santamaría Pérez

Universidad de Alicante

mi.santamaria@ua.es

Este trabajo es parte de un proyecto[9] de investigación centrado en el estudio y análisis de las colocaciones en el discurso especializado con el fin de determinar si existen colocaciones propias del discurso de especialidad y contribuir a su descripción y clasificación, lo que nos permitirá proponer estrategias para su recuperación (semi)automática en corpus textuales. Como tarea previa nos interesa observar cómo se tratan estas combinaciones en los diccionarios especializados y en las bases de datos terminológicas de tipo electrónico. Los resultados nos permitirán diseñar un prototipo de diccionario electrónico de colocaciones especializadas.

Partimos de un concepto de colocación especializada que considera que:

a) La colocación especializada configura una clase específica dentro del conjunto de combinatorias léxicas que conviven en un discurso de especialidad.

[9] This work is part of the research project RICOTERM-4: *Processing of specialized corpora for extracting terminologically relevant multiword expressions (FFI2010-21365-C03-01)*, funded by the Ministry of Economy Affairs.

b) La base de una colocación especializada siempre es una unidad terminológica.

c) Las estructuras de las colocaciones especializadas responden a un conjunto reducido o preferente de patrones, si lo comparamos con la diversidad estructural de las colocaciones de contenido general o no especializado.

Nuestro objetivo es doble. Por un lado, nos proponemos contribuir parcialmente al análisis de la inclusión y de la representación de las colocaciones especializadas en recursos lexicográficos y terminológicos; y por otro lado, nos aventuramos a sugerir algunas innovaciones que mejoren la representación de estas combinaciones en los recursos accesibles sobre todo a traductores y mediadores lingüísticos en general. Hay que tener en cuenta que los expertos de cualquier ámbito de especialidad adquieren la terminología propia de su sector progresivamente y siempre en contexto, de modo que integran también las combinatorias preferentes y restrictivas de manera natural cuando adquieren el conocimiento especializado. En cambio, para los mediadores lingüísticos y los docentes de lenguas con fines específicos se hace necesario disponer de información específica sobre la combinatoria de las unidades terminológicas en los recursos terminológicos accesibles (bancos de datos, diccionarios y vocabularios, ontologías y taxonomías, sistemas de traducción asistida).

Para cubrir este objetivo, analizamos un conjunto de diccionarios monolingües y multilingües de ámbitos especializados distintos (derecho, economía, medicina, ciencia y tecnología, industria, turismo) y algunos bancos de datos terminológicos multitemáticos, que son de uso regular entre traductores y docentes. Complementariamente exploramos la existencia de diccionarios de colocaciones de temática específica o alternativamente la presencia de colocaciones especializadas en diccionarios combinatorios de carácter general.

References

ALMELA, M. (2002). Convergencias de las descripciones de la colocación en la lingüística actual. *Revista de Investigación Lingüística*, 1 (5), pp. 31-62.

ALONSO, M. (1994-1995). Hacia una definición del concepto de colocación: de J. R. Firth a I. A. Mel'čuk. *Revista de Lexicografía*, 1, pp. 9-28.

BOSQUE, I. (2001). Sobre el concepto de “colocación” y sus límites. *Lingüística Española Actual*, 23 (1), pp. 9-40.

— (dir.) (2004). *Redes. Diccionario combinatorio del español contemporáneo*. Madrid: SM.

CABRÉ, M. T. (1999). Terminología: representación y comunicación. Una teoría de base comunicativa y otros artículos. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

CARUSO, V. (2011). Online specialised dictionaries: a critical survey. In: Kosem, Iztok; Kosem, Karmen (eds.). *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 66-75.

CORPAS, G. (1996). Manual de fraseología española. Madrid: Gredos.

FERRANDO, V. (2013). El tratamiento de las colocaciones en la lexicografia española y alemana: estudio contrastivo. *Revista Internacional de Lenguas Extranjeras*, 2, pp. 31-53.

KOIKE, K. (2002). Comportamientos semánticos en las colocaciones léxicas. LEA, XXIV (1), pp. 5-23

L’HOMME, M.C. (2000). Understanding specialized lexical combinations. *Terminology*, 6 (1), pp. 89-110.

L’HOMME, M.C. (2009). A methodology for describing collocations in a specialised dictionary. In: NIELSEN, Sandro; TARP, Sven (eds.). *Lexicography in the 21st Century: In honour of Henning Bergenholtz.* Amsterdam/Philadelphia: John Benjamins, pp. 237-256.

LORENTE, M. (2006-2007). Colocaciones con verbos de soporte en el discurso especializado. *Filología*, 38-39, pp. 99-137.

LORENTE, M., MARTÍNEZ-SALOM, A., SANTAMARÍA PÉREZ, I. AND VARGAS SIERRA, CH. (2014). La información sobre colocaciones en terminología. *XIV Simposio Iberoamericano de terminología RITerm 2014,* Terminología, innovación e impacto social: a 25 años de la fundación de RITerm. 1-4 de diciembre de 2014. Facultad de Letras. Pontificia Universidad Católica de Chile. Santiago de Chile.

LORENTE, M., MARTÍNEZ-SALOM, A., SANTAMARÍA PÉREZ, I. AND VARGAS SIERRA, CH. (2015). Specialized collocations in specialized dictionaries. In TORNEL, S. AND BERNAL, E. (eds.). Collocations and other lexical combinations in Spanish. Theoretical and Applied approaches. Theoretical Developments in Hispanic Linguistics (Ed. Javier Gutiérrez-Rexach) (en prensa).

MEL’ČUK, I *et al.* (1984-1999). Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexicó-sémantiques. Vols. 1-4. Montréal: Les Presses de l’Université de Montréal.

VARGAS-SIERRA, Ch. (2010). Combinatoria terminológica y diccionarios especializados para traductores. In: Ibáñez Rodríguez, Miguel (ed.). *Lenguas de especialidad y terminología.* Granada: Comares, pp. 17-46.

PRAGMATIC INFORMATION AND UNPREDICTABILITY IN LEARNER'S DICTIONARIES

Irena Srdanovic

University of Ljubljana

irena.srdanovic@ff.uni-lj.si

Linguistic and lexicographic tradition has long been neglecting the pragmatic aspects of language. In lexicography, dictionaries designed for foreigners pay attention on pragmatic information, especially the dictionaries that used language corpora as material, starting with the Cobuild dictionary (Sinclair 1987). In addition to the selection and method of presentation of this type of information, the dictionaries also differ in defining the pragmatic information. They are talking about pragmatic, encyclopedic and cultural information on the level of description of nonlinguistic elements. On more narrow level of language use, some dictionaries consider examples as pragmatic information, while others descriptions of the use of certain lexical items, or classification into categories such as free combinations, collocations and idioms (cf. Sharpe 1989, Nuccorini 1993, Inoue 1998). This research analysis the treatment of such information in a variety of monolingual and bilingual dictionaries of Japanese language and aims to provide a typology of pragmatic information.

Nation (2001) believes that a collocation is often grammatically and lexically unpredictable, so students cannot easily produce native-like expressions. The problem of unpredictability of collocations is closely related to differences between the learner's mother tongue and the foreign languages in the process of acquiring by the learner. For example, the English expression 'to make a tea',

could be literally translated into *ocha wo tsukuru* 'to make a tea" in Japanese. This sounds unnatural compared to the proper expression *ocha wo ireru* 'to insert a tea'. Similarly, the adjective 'cold', which is typically used as one of the two different words in Japanese, *tsumetai* or *samui* 'cold' can be as well regarded as unpredictable, since the collocation 'cold water' could easily be mistaken with *samui mizu* instead of the correct usage *tsumetai mizu* 'cold water'. Further, the research takes Japanese adjectives and nouns as an example and based on the corpus and dictionary analysis indicates the types of unpredictable collocations comparing them in Japanese and English. As shown in the examples, it is harder to predict 1) collocation that is morpho-syntactically different (*sei ga takai hito* "a tall person", lit. 'a person with a high back', which cannot be simple used as *takai hito*, lit. 'a tall person', 2) collocation, which in the native language do not have the same combination of words (*takai okane* "a lots of money", lit. 'high money '), and 3) collocation with equivalent combination of words in mother tongue, but partially or fully different meaning (*takai kyouiku* "a high-quality education / high education". The results indicate that dictionaries targeted at language learners need to emphasize information on combinations that are difficult to predict by language learners, which is one type of pragmatic information.

References

INOUE, N. (1998). Gakushuu eiwa jiten ni okeru gohou jouhou to koroke-shon jouhou – ko-pasu de nani ga dekiru ka. *Eigo kyouiku to eigo kenkyuu*, 15, pp. 71–86.

NATION, P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

NUCCORINI, S. (1993). Pragmatics in learners' dictionaries. *Journal of Pragmatics*, 19(3), pp. 215-237.

SINCLAIR J.M. ed. (1987). *Looking up: An Account of the COBUILD Project in Lexical Computing*. Collins, London.

SHARPE, P. A. (1989). Pragmatic Considerations for an English-Japanese Dictionary. *International Journal of Lexicography*, 2, pp. 315-323.

GENREKONSTITUTIVE KOLLOKATIONEN IN WISSENSCHAFTLICHEN TEXTEN

Mirjam Weder

Universität Basel

mirjam.weder@unibas.ch

Schon verschiedentlich wurde die Bedeutung von wiederkehrenden Formulierungsmustern, d.h. mehr oder minder idiomatischen Mehrworteinheiten wie Kollokationen, Bundles, Clusters, Multiword-Units oder Phraseologismen, für wissenschaftliche Textsorten hervorgehoben: Empirische Untersuchungen aus dem Englischen gibt es dazu etwa von Hyland (2012), Biber (2006), Römer & Wulff (2010) oder Simpson-Vlach & Ellis (2010). Für das wissenschaftlichen Schreiben auf Deutsch liegen erst wenige Untersuchungen vor, die das Phänomen oft auch aus leicht anderer Perspektive beleuchten, so etwa Gnutzmann & Lange (1990), die die Struktur von Einleitungen deutschsprachiger Forschungsartikel untersuchen, oder Pohl (2007) und Steinhoff (2007), die das wissenschaftliche Schreiben unter der Erwerbsperspektive betrachten.

Die Bedeutsamkeit von Kollokationen und ähnlichen Mehrworteinheiten liegt u.a. in ihren textkonstitutiven (Reder 2006: 41), aber auch genrekonstitutiven Eigenschaften (Hyland 2012; Biber 2006). Eine typische Eigenheit von wissenschaftlichen Texten besteht etwa darin, dass sie fortlaufend metadiskursiv ihre Entstehungsbedingungen explizit machen müssen, vgl. dazu (Feilke 2010: 6-7). Dies betrifft die Einordnung in den Forschungsdiskurs, die Forschungspraxis, auf denen die Texte beruhen (z. B. empirische

Untersuchung versus theoretische Problematisierung etc.), die Forschungsprozesse, die dem Artikel vorausgegangen sind, sowie die Darstellungsmittel, die gewählt werden. Zur Darstellung dieser Zusammenhänge kommen oft ähnliche Formulierungsmuster zum Einsatz, teils formelhafte Versatzstücke akademischen Schreibens Simpson-Vlach & Ellis (2010), die insofern als metadiskursive Strategien gelten dürften, als sie ein Textexemplar inhaltlich und strukturell mit dem Forschungsdiskurs koppeln, Vorstellungen des Autors über diesen Diskurs transportieren als auch Leser-Erwartungen steuern (Hyland 2005: 13).

Der Beitrag fokussiert auf das oben skizzierte text- und genrekonstitutive Potential von Mehrworteinheiten mit metadiskursiver Funktion in deutschen wissenschaftlichen Texten. Als Korpus dienen Einleitungen und Schlussworte aus 100 wissenschaftlichen Artikeln aus deutschsprachigen Fachzeitschriften aus den Fächern Geschichte, Philosophie, Medienwissenschaften/Publizistik und Germanistik (Literaturwissenschaft und Sprachwissenschaft). Methodisch werden datengeleitete automatisierte korpuslinguistische Methoden (Bondi 2010, Römer & Wulff 2010, Biber 2006) mit genre-analytischen Interpretationsverfahren kombiniert (Swales 1990, 2004). Es werden die häufigsten Mehrworteinheiten und ihre metadiskursive Funktion vorgestellt sowie das Verfahren der Extraktion und Identifikation dieser Einheiten problematisiert.

References

BIBER, D. (2006). *University language a corpus-based study of spoken and written registers*. Amsterdam, Philadelphia (Pa.): J. Benjamins.

BONDI, M. (2010). Metadiscursive Practices in Introductions: Phraseology and Semantic Sequences across Genres. *Nordic Journal of English Studies* 9, pp. 99–123.

FEILKE, H. (2010). „Aller guten Dinge sind drei“ – Überlegungen zu Textroutinen & literalen Prozeduren. In: BONS, I. and GLONING, T. (KALTWASSER, D. (ed. (FEST-PLATTE FÜR GERD FRITZ. Giessen. [online] Available at: http://www.festschrift-gerd-fritz.de/files/feilke_2010_literale-prozeduren-und-textroutinen.pdf [Accessed 1.3.2015]

GNUTZMANN, C. and LANGE, R. (1990). Kontrastive Textlinguistik und Fachsprachenforschung. In: Gnutzmann, C. (ed. *Kontrastive Linguistik*. Frankfurt am Main□; New York□: Duisburg: P. Lang□; Gesellschaft für Angewandte Linguistik. (= Forum Angewandte Linguistik Bd. 19), pp. 85–116.

HYLAND, K. (2005). *Metadiscourse: exploring interaction in writing*. London, New York: Continuum.

HYLAND, K. (2012). Bundles in Academic Discourse. *Annual Review of Applied Linguistics* 32, pp. 150–169.

POHL, T. (2007). Wissenschaftliches Einleiten – systematisch und ontogenetisch. In: Ursula DOLESCHAL, U. and Gruber, H. (ed. *Wissenschaftliches Schreiben abseits des englischen Mainstreams*. Frankfurt am Main, New York: Peter Lang, pp. 217–251.

REDER, A. (2006). *Kollokationen in der Wortschatzarbeit.* Wien: Praesens.

RÖMER, U. & WULFF, S. (2010): Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing Research* 2(2), pp. 99–127.

SIMPSON-VLACH, R. and ELLIS, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics* 31(4), pp. 487–512.

STEINHOFF, T. (2007). *Wissenschaftliche Textkompetenz. Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten*. Tübingen: Max Niemeyer.

SWALES, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

SWALES, J. M. (2004). *Research genres: explorations and applications*. Cambridge, New York: Cambridge University Press.

SEMANTIC STABILITY OF ENGLISH IDIOMS

Margarita Yagudaeva

Sussex University

yagudaevam@gmail.com

Idioms, fixed expressions whose overall meaning is not deducible from their constituents, have always been of great interest to linguists. Idiom categorization, fixedness, figurativeness, and other aspects have been studied to varying degree across disciplines and in different languages (cf. Gibbs with colleagues 1991, Glucksberg 2001, Knappe 2004, Kunin 1996, and Moon 1998, to name just a few). The majority of studies so far have examined the structural changes of idioms, that is, their flexibility and/or fixedness, idiom comprehension, and the motivation for idiom meaning, rather than their meaning variation, preserving their canonical form. In particular, the semantic stability of idiomatic expressions has rarely been questioned. The hypothesis of the current research is that English idioms are capable of meaning change over time, as is the case with other lexemes.

The purpose of the presentation is to briefly outline the existing views on idiomatic expressions, identifying the gap in the existing research. In addition, the methodology developed to trace meaning variation in English idioms will be discussed, including, information on the sample selection, the sources used, and finally the questionnaire design. I propose to start examining idioms that have multiple meanings, i.e. polysemous idioms. Polysemy, according to some studies, is regarded as an indicator of meaning change taking place within a word or lexeme; in other words, the multiplicity of meanings is synchronically

regarded as polysemy, and diachronically is regarded as a meaning change in process. Therefore, polysemous idioms can act as the most representative case for the meaning variation an idiom may undergo over time.

Within the methodology section of the presentation, I will discuss the difficulties involved in conducting current research, such as, the availability of modern and historical English language corpora with untagged idioms and idioms that vary in structure, online and printed databases, for idiom extraction and comparison, as well as corpus-based analysis for the attestation of meaning differences. The crucial step to identify the change of meaning in idiomatic expressions is to determine when a phrase has become a conventionalized idiom and has started to be used as such. Since the available English language corpora provide only concordances or word strings that occur together, it is likely that in some cases the phrase would be used in its literal meaning, rather then idiomatic. In contrast, I argue that an idiom, after having been established as one, is exposed to certain semantic shifts. Consequently, several methods in combination should be applied for searching the selected sample at different periods, which will be addressed during the presentation.

References

Gibbs, Jr., R.W., & Nayak, N.P. (1991). 'Why idioms mean what they do', *Journal of Experimental Psychology*, 120(1), pp. 93-95

Glucksberg, S. (2001). *Understanding Figurative Language: From Metaphor to Idioms*, Oxford: Oxford University Press

Knappe, G. (2004). *Idioms and Fixed Expressions in English Language Study before 1800*. Frankfurt am Main: Peter Lang GmbH

Kunin, A.V. (1996). *Kurs fraseologii sovremennogo angliyskogo yazika*, 2nd ed. Moskva: Visshaya Shkola

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus Based Approach*, Oxford: Clarendon

Machine-aided and corpus-based translation of phraseological units Traducción de unidades fraseológicas basada en corpus o asistida por ordenador

MULTILINGWIS – A MULTILINGUAL SEARCH TOOL FOR MULTI-WORD UNITS IN MULTIPARALLEL CORPORA

Simon Clematide

Institute of Computational Linguistics

simon.clematide@gmail.com

Large collections of multiparallel texts – i.e. multilingual documents with aligned paragraphs or sentences across all languages – are openly available, for instance debates from the European parliament (Koehn 2005), administrative and legislative texts (Steinberger et al. 2012), patents, as well as translated subtitles (see Tiedemann 2012). These corpora are highly useful and valuable for translators, terminologists, and contrastive corpus linguists if they can be exploited effectively.

The web-based search interface

There are a number of parallel corpus search systems[10], some of them include multiparallel searches (von Waldenfels 2011; Tiedemann 2012). Our goal is to provide a user-friendly web-based tool for ad hoc searches for translation variants of multi-word units – typically complex noun phrases – in multiparallel corpora. For instance, a Spanish term such as *el cotejo de datos del ADN* (DNA-matching). The search focuses on content words – represented by their lemmas –, ignoring any functional word, which includes prepositions

[10] For instance, http://www.linguee.de, http://www.tradooit.com, http://glosbe.com, http://www.tausdata.org, http://pub.cl.uzh.ch/purl/bilingwis, or the Sketch Engine http://www.sketchengine.co.uk.

(resulting in "*cotejo dato ADN"* for the example given above). While the order of search terms must match the order in the sentences of the search language by default, there is, of course, no order restriction in the corresponding parallel sentences.

A simple search interface allows the user to enter one or more lemmas in a text field. User-friendliness means automatic language identification, removal of function words and lemmatization of inflected word forms – if the user prefers to enter a phrase instead of content lemmas. An advanced search form supports part-of-speech filtering and more specific options.

The primary result presentation shows parallel sentences with the corresponding words for the search terms being highlighted – a useful feature which some systems lack completely or which does not work reliably for multi-word searches (see Volk et al. 2014). Each example sentence provides a hyperlink directed to a new search with the highlighted content words as search terms, therefore supporting quick explorations across cross-lingual and monolingual verbalizations of the same concepts. An alternative presentation groups the hits according to the most frequent translation patterns.

Data preparation and backend technology

Part-of-speech tagging and lemmatization was performed by the Treetagger (Schmid 1994) in our prototype. It bases on an improved version of Europarl (Graën et al. 2014) and covers German, English, French, Italian, and Spanish to date. Bilingual sentence alignments of all involved language pairs and a subsequent multiparallel harmonization of them lead to multiparallel sentences. We used GIZA++, a bilingual statistical word alignment software (Och & Ney 2003), to align the content lemmas (adjectives, nouns, verbs) for each language pair in each direction. These directed *1:n* alignments are symmetrized by an application-specific machine learning approach based on a multilingual gold standard for multi-word units (*reference suppressed*). A GDEX score (Rychlý et al. 2008) for good example sentences is computed offline and stored as an attribute for each sentence. We use a relational database management system to store and efficiently retrieve the corpus data, which includes the lemma alignments.

References

GRAËN, J., BATINIC, D. & VOLK, M. (2014). Cleaning the Europarl Corpus for Linguistic Applications. In Konvens 2014. Stiftung Universität Hildesheim. Available at: http://dx.doi.org/10.5167/uzh-99005.

KOEHN, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Machine Translation Summit. pp. 79–86.

OCH, F.J. & NEY, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational linguistics, 29(1), pp.19–51. Available at: http://dx.doi.org/10.1162/089120103321337421.

RYCHLÝ, P. ET AL. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In Proceedings of the XIII EURALEX International Congress. pp. 425–432.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference on New Methods in Language Processing. Studies in Computational Linguistics. pp. 44–49. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.1139.

STEINBERGER, R. ET AL. (2012). DGT-TM: A freely available Translation Memory in 22 languages. In Proc of LREC 2012. pp. 454–459. Available at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/814_Paper.pdf.

TIEDEMANN, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Proc of LREC 2012. pp. 2214–2218.

VOLK, M., GRAËN, J. & CALLEGARO, E. (2014). Innovations in Parallel Corpus Search Tools. In Proc LREC 2014. pp. 3172–3178.

VON WALDENFELS, R. (2011). Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. In M. Daniela & R. Garabik, eds. Natural Language Processing, Multilinguality. Proceedings of Slovko 2011. Bratislava, pp. 156–162.

COMPARATIVE STRUCTURES IN CROATIAN: MWU APPROACH

Kristina Kocijan
University of Zagreb
krkocijan@ffzg.hr

Sara Librenjak
University of Zagreb
sara.librenjak@gmail.com

Croatian language has a very rich phraseme structure, as described in Matešić (1982), Menac et.al. (2003; 2007) and Menac-Mihalić (2004), as well as many others. The authors have analyzed 2500 Croatian idiomatic expressions as defined by the Croatian Phraseme Dictionary (Menac et al.; 2003), and sorted them according to their syntactic properties. On that basis, five major syntactic groups of Croatian idioms were found, and NooJ grammars were constructed accordingly. We were able to recognize five syntactic types of idiomatic expressions: 1. fixed structure, 2. noun phrase with an attribute or apposition, 3. verbal phrase with a direct object, 4. verbal phrase with the optional indirect object that can disrupt the syntactic structure, and 5. comparative structure (verb or adjective as a noun). The last four types required building syntactic NooJ grammars in order to be recognized in all their varieties (split cases, inversions), while the first one was added directly to the NooJ dictionary. In this paper we will more closely describe those idiomatic expressions that belong to the Type 5 expressions.

The set of NooJ grammars for detecting idioms in Croatian is trained on digitized corpus made from Croatian literally text in which all of the processed idioms can be found. Subsequently, finished grammars were tested on web based Croatian corpus sample (Agić and Ljubešić, 2014), and compared with manually marked results. Except for detecting idioms and providing statistical data, this work can be helpful in machine (aided) translation of Croatian, since MWUs are typically harder to process in MT and require special attention. A few can be understood in direct translation to e.g. English, and are usually about

universal characteristics, not uncommonly animalistic (slobodan kao ptica – free as a bird, slijep kao šišmiš – blind as a bat, debeo kao svinja – as fat as a pig). Then, some are only similar, but not direct translation (spavati kao top –> lit. to sleep as a canon = to be fast asleep; puši kao Turčin -> lit. he smokes like a Turk= to smoke like a chimney). Finally, the comparative structures can be completely different, especially when culturally dependent (kao kec na jedanaest -> lit. as an ace on Jack = at the worst possible moment; provesti se kao Janko na Kosovu -> lit. had the time as Janko on Kosovo = had a very bad experience; kao pokisla kokoš -> lit. as a wet chick = feel very sad, or very blue.).

As the phrasemes are rooted in the tradition of the language and the society from which they hail from, they need a special treatment in computational linguistics. With this tool, Croatian idioms can be successfully detected and matched with their translation in corresponding language, which would eliminate awkward and completely wrong automated translations. Thus, this work seeks to aid not only the successful development of additional language resources for Croatian language, but a possible assistance in future work relating to machine assisted translation.

References

AGIĆ, Ž. AND LJUBEŠIĆ. N. (2014). The SETimes.HR Linguistically Annotated Corpus of Croatian. In Proceedings of the Ninth International Conference on Language Resources and Evaluation, Reykjavik, pp. 1724–27.

MACHONIS, P.A. (2012). Sorting NooJ out to take Multiword Expressions into account. In K. Vučković, B. Bekavac & M. Silberztein Eds. Formalising Natural Languages with NooJ : Selected Papers from the NooJ 2011 International Conference. Cambridge Scholars Publishing, Newcastle., UK. pp. 152-165.

MATEŠIĆ, J. (1982). Frazeološki rječnik hrvatskoga ili srpskog jezika. Zagreb: Školska knjiga.

MENAC, A., FINK-ARSOVSKI, Ž. AND VENTURIN, R. (2003). Hrvatski frazeološki rječnik. Zagreb: Naklada Ljevak.

MENAC-MIHALIĆ, M. (2007). Hrvatski Dijalektni Frazemi S Antroponimom Kao Sastavnicom. In Folia Onomastica Croatica, no. 12/13, 361–85.

IDENTIFICATION, CLASSIFICATION AND ANALYSIS OF PHRASEMES IN AN L2 LEARNER CORPUS OF ITALIAN

Christine Konecny
University of Innsbruck
Christine.Konecny@uibk.ac.at

Erica Autelli
University of Innsbruck
Erica.Autelli@uibk.ac.at

Andrea Abel
European Academy
Bolzano/Bozen
andrea.abel@eurac.edu

Lorenzo Zanasi
European Academy
Bolzano/Bozen
lorenzo.zanasi@eurac.edu

In spite of the increased application of corpus-based methods in phraseological research in the past years (cf. Heid 2005; Heid/Weller 2010; Steyer 2013), the initiating interest for phraseological aspects in learner corpora research (cf. Paquot/Granger 2012) and the constantly growing number of phraseodidactic studies (cf. Kühn 1987;1992; Lorenz-Bourjot/Lüger 2001; Gonzáles Rey 2013;2014; Konecny et al. 2013; Sułkowska 2013), suggestions of appropriate criteria for identifying, classifying and analyzing phrasemes in learner corpora seem to be still underrepresented. Studies of such kind could be useful not only for revealing the actual use of phrasemes at various CEFR levels (2001) and for detecting recurrent mistakes and error causes, but also for developing suitable didactic material in order to achieve a certain target level in the phraseological use at different CEFR levels. Within the LeKo project (www.leko-project.org), carried out in cooperation between the University of

Innsbruck and the European Academy of Bolzano/Bozen, we aim at describing the use of phrasemes by L2 learners of Italian for didactic purposes, by combining both quantitative and qualitative methodological approaches. For this purpose, we analyze a subset of the KOLIPSI corpus, which consists of written German and Italian L2 productions by South Tyrolean secondary school pupils that have already been assigned to CEFR levels in reliable way (cf. Abel et al. 2012); the LeKo subcorpus covers the levels A2-C1 and contains 288 Italian texts written by German L1 pupils.

Before being able to analyze the phrasemes present in our corpus, it was necessary to establish first our conception of phrasemes and the criteria that should be applied for identifying and classifying them. To this end, we adopted a combination of deductive and inductive methods, i.e. following current concepts present in pertinent studies as well as pre-analyzing selected KOLIPSI texts in order to get an idea of which phraseme types are actually used by the learners. As far as collocations are concerned, we found out that lexical collocations *strictu sensu* occur only rarely (especially on lower CEFR levels), for which it seemed useful to adopt a broader concept of collocation, including also sequences such as *andare a casa* ('to go home') and *guardare su/in internet* ('to look/search on the internet'), which seem freely combined but in which at least the preposition is idiosyncratically bound to the actual verb. As to other phraseme (sub)categories, we decided to adopt a "mixed classification" in terms of Burger (2007: 53). Besides various types of referential (both non-idiomatic and idiomatic) phrasemes, we took into consideration also communicative and structural phrasemes.

In our paper, we will illustrate which phrasemes we detected at various levels and how we proceeded in assigning them to different phraseological (sub)types as well as in the identification of possible error causes. In order to record several possible kinds of the latter, we provided for a separated category named "not exist" for those cases in which phrasemes existing in German were translated literally into not-existing Italian expressions (e.g. **queste vacanze diventano il martello* = interference from Germ. *dieser Urlaub wird der Hammer*).

References

Abel, Andrea; Vettori, Chiara; Wisniewski, Katrin (eds.) (2012): Gli studenti altoatesini e la seconda lingua: indagine linguistica e psicosociale. / Die

Südtiroler SchülerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung. Vol. 1 – Bd. 1. Bozen-Bolzano: EURAC.

Burger, Harald (^{3}2007): Phraseologie. Eine Einführung am Beispiel des Deutschen. Berlin: Schmidt.

CEFR (2001) = Council of Europe (2001): The Common European framework of reference for languages: Learning, teaching, assessment. Cambridge: CUP.

Đorović, Danijela (2013): "(Mis)understanding Italian phrasemes in Italian history texts." In: Konecny, Christine; Hallsteinsdóttir, Erla; Kacjan, Brigita (eds.): Phraseologie im Sprachunterricht und in der Sprachendidaktik / Phraseology in language teaching and in language didactics. Maribor: Mednarodna založba Oddelka za slovanske jezike in književnosti, Filozofska fakulteta [Zora; 94], 137-151.

Gonzáles Rey, María Isabel (ed.) (2013): Phraseodidactic Studies on German as a Foreign Language / Phraseodidaktische Studien zu Deutsch als Fremdsprache. Hamburg: Dr. Kovač [Lingua. Fremdsprachenunterricht in Forschung und Praxis; 22].

Gonzáles Rey, María Isabel (ed.) (2014): Outils et méthodes d'apprentissage en phraséodidactique. Fernelmont (Belgique): EME Editions.

Heid, Ulrich (2005): "Corpusbasierte Gewinnung von Daten zur Interaktion von Lexik und Grammatik: Kollokation – Distribution – Valenz." In: Lenz, Friedrich; Schierholz, Stefan J. (eds.): Corpuslinguistik in Lexik und Grammatik. Tübingen: Stauffenburg, 97-122.

Heid, Ulrich; Weller, Marion (2010): "Corpus-derived data on German multiword expressions for lexicography." In: Vatvedt Fjeld, Ruth; Torjusen, Julie Matilde (eds.): Proceedings of the 15th Euralex International Congress, Oslo, 7-11 August 2012. Oslo: Department of Linguistics and Scandinavian Studies of the University of Oslo, 331-340.

Kühn, Peter (1987): "Deutsch als Fremdsprache im phraseologischen Dornröschenschlaf. Vorschläge für eine Neukonzeption phraseodidaktischer Hilfsmittel." In: Fremdsprachen Lehren und Lernen 16, 62-79.

Kühn, Peter (1992): "Phraseodidaktik. Entwicklungen, Probleme und Überlegungen für den Muttersprachenunterricht und den Unterricht DaF." In: Fremdsprachen Lehren und Lernen 21, 169-189.

Lorenz-Bourjot, Martine; Lüger, Heinz-Helmut (eds.) (2001): Phraseologie und Phraseodidaktik. Wien: Praesens [Beiträge zur Fremdsprachenvermittlung / Sonderheft; 4].

Paquot, Magali; Granger, Sylviane (2012): "Formulaic Language in Learner Corpora." In: Annual Review of Applied Linguistics 32, 130-149.

Scinetti, Luca (2013): "L'uso delle collocazioni: due gruppi di apprendenti a confront." In: Italiano LinguaDue 2, 109-131.

Steyer, Kathrin (2013): Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht. Tübingen: Narr.

Sułkowska, Monika (2013): De la phraséologie á la phraséodidactique. Études théoriques et pratiques. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Voghera, Miriam (2004): "Composizione: Polirematiche." In: Grossmann, Maria; Rainer, Franz (eds.): La formazione delle parole in italiano. Tübingen: Niemeyer, 56-69.

A COMPARATIVE STUDY ON LEXICAL BUNDLES IN IDENTITY CONSTRUCTION IN L1 AND L2 ACADEMIC WRITING: A CORPUS-BASED APPROACH

Ruixue Liu
Northwestern Polytechnical University
yusnowhite@mail.nwpu.edu.cn

Xueai Zhao
Northwestern Polytechnical University
xazhao@nwpu.edu.cn

Academic writing not only ensures academics to delivery their ideational "content", but also enables them to present themselves and shape their social identity (Hyland, 2002). As an important part in academic writing, identity constructions has been probed by researchers home and abroad in terms of first person pronouns for self-mention or the stance markers in metadiscourese. However, few researches have investigated the phraseological patterns in identity construction in academic writing. Thus, this paper adopted a corpus-based approach to first explore the patterns of lexical bundles in identity construction, then to examine their variations in L1 and L2 academic writing. Two corpora were established, respectively including 60 English academic papers from Chinese students (L2 learners) and 60 from scholars of L1 learners, which were all from the field of Aeronautics and Computer Science. In this study, a model of lexical bundles was proposed according to Hyland (2005)'s research on phraseological pattern of *that* and other influential findings on identity construction: the lexical bundles were first categorized into three types based on their evaluation resources (human, abstract entity and concealed) and then further divided into 6 categories depending on their expression with modal verbs or not (Xu, 2011). For example, the phraseological

pattern referring to evaluation resource of “human”, without model verb could be *I/we +VP/V(that) X* (e.g **We propose that**…) while that referring to concealed one with model verbs could be *It+Modal+ be + V+ed that X* (e.g. **It can be concluded that**…).The analysis showed that the lexical bundles about “human” appear much more frequently than the others in terms of evaluation resources. Besides, this study also indicated that L2 learners tend to underuse lexical bundles in identity construction, especially the ones about “human” and “concealed”, but overuse certain verbs and adjectives (e.g. **believe, consider, well-known, clear**), and are lack of alternation when stating findings or comments. The possible causes for these variations between L1 and L2 academic writing might be the differences between these two groups of academics in their experience and confidence on research and ways taught in writing. As to the findings of this comparative study, we hope that it can raise Chinese students’ consciousness in establishing authorship through the utilization of various phraseological patterns in their academic writing.

References

HYLAND, K. (2002). Authority and invisibility: authorial identity in academic writing. *Journal of Pragmatics*, 34(8), p.1091.

HYLAND, K. (2005). Hooking the reader: a corpus study of evaluative *that* in abstracts. *English for Specific Purpose*, 24(2), p.123.

XU, F. (2011). A corpus-based study on lexical bundles in identity construction in Chinese students’ academic writing. Foreign Languages Research, 127(3), p.57.

THE ADJECTIVIZATION OF POSTERIOR FRENCH NOUNS IN BINOMINAL EXPRESSIONS: A CORPUS-BASED STUDY

François Maniez

Centre de Recherche en Terminologie et en Traduction

Université Lumière Lyon 2

francois.maniez@univ-lyon2.fr

French *N1+N2* constructions have been thoroughly studied by lexicologists in the past few decades, notably by Noailly (1990), who mentions the proliferation in advertising of such sequences as *événement minceur* or *cadeau saveur*. These constructions, which are common to all Romance languages (Fabre 1996, Villoing 2002, Savary 2004, Montermini 2008), are increasingly used in both written and oral discourse, and cannot all be lexicalized. They are very often formed by ellipsis of a preposition (*version (sur) papier*, *études (de) marketing*, *allocations (de) chômage*), and observation of current usage shows that the shortened version is gradually replacing the original wording in such sequences. However, the semantic relationships between the two nouns in these constructions are multifarious (Arnaud 2001) and cannot always be attributed to ellipsis of a preposition, as some cases clearly reflect adjectival use of N2 (*rôle clé*, *solution miracle*, *présentateur vedette*). In the literature, these constructions have indeed been mostly analyzed as either the result of regular

N+N compounding or the syntactic combination of a noun with a second noun converted into an adjective. Because N2 has some basic adjectival properties in such structures, some have even argued (Martinho 2013) that it is equivalent to an adjectival modifier.

In this article, we investigate the status of French *N+N* constructions in which the *N2* productively combines with different *N1*s (e.g. *assurance chômage* 'unemployment insurance'*, allocations chômage* 'unemployment benefits'). Our work involves the use of two corpora.

The Corpus of Journalistic French (a corpus of early twenty-first century newspaper articles drawn from French dailies such as *Le Monde*, *L'Humanité* and *La Dépêche du Midi*, described in Chambers, 2005) is used in its part-of-speech tagged version to extract all the tokens of its *N+N* constructions. Among those, we analyze the most productive N2s (*clé*, *culte*, *fantôme*, *fétiche*, *miracle*, *phare*, *record*, *symbole*, *vedette*) and track the evolution of their use over the past thirty years using the Google Books NGram Viewer. We show that most of those N2s gradually tend to behave as adjectives over time as regards agreement in number whereas some other nouns in our corpus do not (*conseil*, *crédit*, *maladie*, *marketing*, *vie*, *vieillesse*) and advance hypotheses that might help explain this phenomenon.

Using recent examples drawn from usage observed on the Web (mostly in informal usage witnessed on blogs and forums), we also show that following perceived or institutional lexicalization of the N1 N2 sequence, such nouns are now increasingly used as adjectives in predicative structures (*l'affluence était record*, *ces répliques sont cultes*). This trend would tend to show that the adjectivization of N2 nominal epithets goes beyond the usual behavior of relational adjectives (which are theoretically non-predicative and to which N2s have been previously assimilated (Demonte 1999, Martinho 2013) in the literature on the subject) and that such nouns are gradually acquiring the status of qualifying adjectives.

References

ARNAUD, P. (2001). Relations sémantiques N1-N2 dans les composés timbre-poste. In: H. Paugam-Moisy, V. Nyckees, J. Caron-Pargue eds. *La Cognition entre individu et société*. Paris: Hermès-Sciences. pp. 105-117.

DEMONTE, V. (1999). A Minimal Account of Spanish Adjective Position and Interpretation. In: J. Franco, A. Landa & J. Martín eds. *Grammatical Analyses in Basque and Romance Linguistics.* Amsterdam, John Benjamins.

FABRE, C. (1996). *Interprétation automatique des séquences binominales en anglais et en français. Application à la recherche d'informations,* thèse de doctorat, Université de Rennes.

MARTINHO, F. (2013). Noms épithètes dans les expressions binominales. In: *Linguística: Revista de Estudos Linguísticos da Universidade do Porto*; Vol. 8, ppp. 39-67.

MONTERMINI F. (2008). La composition en italien dans un cadre de morphologie lexématique. In: D. Amiot ed. *La composition dans une perspective typologique*, Artois Presses Université, pp.161-187, Etudes linguistiques.

NOAILLY, M. (1990). *Le Substantif épithète*, PUF, Paris.

SAVARY, A. 2004. *Recensement et description des mots composés – méthodes et applications*, Thèse de doctorat en Informatique Fondamentale, Laboratoire d'Automatique Documentaire et Linguistique, Université Paris 7.

VILLOING, F. (2002). *Les mots composés [VN] N/A du français: réflexions épistémologiques et propositions d'analyse*, thèse de doctorat, Paris-X Nanterre.

EXPLORING THE FORMAL AND CONTEXTUAL STEREOTYPICALITY OF COLLOCATIONAL CHAINS

Piotr Pęzik

University of Lodz

piotr.pezik@gmail.com

Automatic Combinatorial (or Collocation) Dictionaries (ACDs) can be defined as databases of recurrent word combinations whose status as phraseological units is estimated from distributional criteria such as frequency, degree of binding or evenness of distribution (Kilgarriff & Rychlý 2010). It has been argued that, despite their imperfect precision, ACDs extracted from large reference corpora complement traditional collocation dictionaries in that they usually guarantee a higher recall of basic collocations (Pęzik 2014). At the same time, since a considerable proportion of phraseological units are made up of more than two obligatory word segments, ACDs which only record combinations of two words co-occurring in positionally or syntactically defined contexts have a principally limited coverage of idiomatic units.

This paper presents a method of extracting an ACD of "collocational chains" of two and three elements which can be linked by direct object and adjectival modifier dependencies. Collocational chains (or catenae) are defined here as recurrent "combinations of words the projections of which are continuous with respect to dominance" (Osborne, Putnam, Groß 2012: 354), cf. O'Grady (1998) which are characterized by some formal and contextual stereotypicality.

In order to extract a complete ACD of such chains from the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), we first process these two corpora with the Stanford Dependency Parser (Chen and Manning 2014). Next, we compute a variety of statistics for each of the relevant types of dependencies found, including their frequency, different measures of binding and dispersion. These statistics are then stored in the resulting ACD and used for retrieving recurrent collocational chains.

A simplified entry of this dictionary generated for the lemma *finger* using only the relative frequencies of the relevant catenae is shown in Table 1. The entry contains three-word approximations of figurative idioms such as *have green fingers* (Am. English: *have a green thumb*) or *point (an) accusing finger* as well as open collocations (Cowie, Mackin, McCaig 1993) such as *have long fingers*.

V	**AMOD**	**DOBJ**	**BNC Frequency**	**COCA Frequency**
point	accusing	finger	16	38
have	green	fingers	5	0
run	long	finger(s)	5	32
follow	point	finger	4	15
point	accusatory	finger	3	11
have	long	finger(s)	3	26

Table 1 Recurrent V -[:dobj] -> N -[:amod] -> JJ catenae for the noun “finger”.

The validation of the phraseological status of such chains according to the criteria used for two-word combinations (e.g. Mel’čuk 2001) may be problematic. For example, the chain *run (one’s) long fingers* can be seen to consist of a blend of two types of collocations: a) the open, possibly type-bound (Martin 2008) collocation *long fingers* and 2) the more restricted combination of *run + finger(s)*, in which the sense of *run* is more restricted to the noun *finger(s)*. Arguably, what makes this chain “collocational” or “idiomatic” is not only its

partially fixed internal structure, but also the fact that it functions in stereotypical contexts, such as imaginative writing narratives. By providing hyperlinks to their original contexts of occurrence, the format of the “higher-order” ACDs extracted in this study makes it possible to explore both formal and contextual stereotypicality of such collocational chains.

References

CHEN, DANQI, AND CHRISTOPHER D MANNING. 2014. “A Fast and Accurate Dependency Parser Using Neural Networks.” In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 740–50.

COWIE, ANTHONY PAUL, RONALD MACKIN, AND I. R MCCAIG. 1993. Oxford Dictionary of English Idioms. Oxford; New York: Oxford University Press.

KILGARRIFF, ADAM, AND PAVEL RYCHLÝ. 2010. “Semi-Automatic Dictionary Drafting.”. In A Way with Words : Recent Advances in Lexical Theory and Analysis : A Festschrift for Patrick Hanks, edited by Gilles-Maurice De Schryver, 299–312. Kampala: Menha Publishers.

O’GRADY, WILLIAM. 1998. “The Syntax of Idioms.” Natural Language & Linguistic Theory 16, no. 2 (1998): 279–312.

OSBORNE, TIMOTHY, MICHAEL PUTNAM, AND THOMAS GROß. 2012. “Catenae: Introducing a Novel Unit of Syntactic Analysis: Catenae: Introducing a Novel Unit of Syntactic Analysis.” Syntax 15, no. 4 (December 2012): 354–96. doi:10.1111/j.1467-9612.2012.00172.x.

MARTIN, WILLY. 2008. “A Unified Approach to Semantic Frames and Collocational Patterns.” In Phraseology: An Interdisciplinary Perspective, edited by Sylviane Granger and Fanny Meunier, 51–65. Amsterdam ; Philadelphia: John Benjamins Pub.

MEL’ČUK, IGOR. 2001. “Collocations and Lexical Functions.” In Phraseology : Theory, Analysis, and Applications, edited by Anthony Paul Cowie, 23–54. Oxford [etc.]: Oxford University Press.

PĘZIK, PIOTR. 2014. “Graph-Based Analysis of Collocational Profiles.” In Phraseologie Im Wörterbuch Und Korpus (Phraseology in Dictionaries and Corpora), edited by Vida Jesenšek and Peter Grzybek, 227–43. ZORA 97. Maribor, Bielsko Biała, Budapest, Kansas, Praha: Filozofska fakuteta.

LOOKING FOR MULTIWORD TERMS IN A COMPARABLE BILINGUAL CORPUS

Ivanka Rajh

Zagreb School of Economics and Management

irajh@zsem.hr

In specialized languages, e.g. the language of marketing, there is a strong tendency towards the creation of rules or principles for the designation of new concepts. Researchers aim for a new term to reflect the characteristics of a new concept, thus the main principles of term formation are transparency and consistency (Sager 1990: 57). Regarding the circumstances in which new terms are created, Sager differentiates between primary and secondary term formation (*ibid*: 80-85). Primary term formation is monolingual and accompanies the formation of a new concept. For example, *customer lifetime value* in English, which is the main language in which new marketing concepts and terms are created. On the other hand, secondary term formation occurs as a result of a monolingual revision of the existing terminology or as a result of the transfer of knowledge into another language community. For example, the equivalents of the above term are *doživotna vrijednost kupca* in Croatian and *življenjska vrednost kupca* in Slovenian. The main difference between the two is that in primary term formation there is no pre-existing term, although there may be rules for term formation in a particular field, while in secondary term formation an already existing term may influence the formation of a new term. Furthermore, in the transfer of knowledge and creation of new terminology there are differences among language communities regarding their economic and linguistic situation. Cabré (1998:18) defines non-dominant languages and

cultures as those that are scientifically and technologically dependent on dominant cultures from which they import new knowledge and the accompanying terminology.

Croatian and Slovenian are examples of non-dominant languages, where borrowing (especially from English) is the usual first step in finding an equivalent of a foreign term (*marketing* and *marketing-miks* in Croatian, *marketing* and *marketing-mix* in Slovenian). The next step is translation, which is a precondition for the further creation of derivatives according to domestic term formation rules, multiword terms being one of the possible outcomes of that process (for the above terms, *tržišno poslovanje* and *splet marketinga* in Croatian and *trženje* and *trženjski splet* in Slovenian). Since Croatian and Slovenian are related languages, multiword terms are structured in almost the same way, with some syntactic variation across six basic patterns (Hudeček and Mihaljević 2009; Vintar 2008). A study of twelve morphosyntatic patterns was conducted in a comparable Croatian and Slovenian corpus made up of nine university-level marketing textbooks using the Sketch Engine.

The results show that in the texts written by Croatian and Slovenian authors, the formation of multiword terms correspond to the ones described in the terminological literature. However, in translations from English, some terminological solutions are quite different not only from the ones described in the literature, but also between the two languages, which may be accounted for by translators' individual characteristics. These solutions are assessed against Croatian, Slovenian and international terminological principles as well as against the methods used in the corpus study.

References

CABRÉ, M. T. (1999). *Terminology: theory, methods and applications.* Amsterdam: John Benjamins.

HUDEČEK, L. AND MIHALJEVIĆ, M. (2009). *Hrvatski terminološki priručnik*. Zagreb: Institut za hrvatski jezik i jezikoslovlje

SAGER, J. C. (1990). *A practical course in terminology processing.* Amsterdam; Philadelphia: John Benjamins

VINTAR, Š. (2008). *Terminologija: terminološka veda in računalniško podprta terminografija.* Ljubljana: Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo.

CORPORA, THE WORLD WIDE WEB AND QUESTIONNAIRES AS SOURCES OF INFORMATION ON RECENT PHRASEOLOGICAL BORROWINGS: THE CASE STUDY OF THE POLISH UNIT *WYGLĄDAĆ JAK MILION DOLARÓW*

Joanna Szerszunowicz

University of Bialystok

joannaszersz@gmail.com

English as a modern lingua franca is a donor not only of words, but also of fixed expressions incorporated by other languages. One of them is Polish, in which many anglicisms, including phraseological borrowings, have been attested recently. An example of a new multiword loan unit is *wyglądać jak milion dolarów* (lit. to look like w million dollars), a calque of the idiom *to look like a million dollars*, commonly used, especially in the spoken variety of Polish and on the Internet. It has been observed that the unit is often modified (e.g. *Wyglądasz jak milion dolarów długu publicznego*, lit. You look like a million dollars of public debt), which shows that it is already well established in Polish. The phrase is marked with novelty and originality, since Polish similes describing good looks do not come from the domain of finance. In Polish phraseology good-looking people, both men and women, are compared to

supernatural god-like creatures, e.g. an angel (*wyglądać jak anioł*), a goddess (*wyglądać jak bóstwo*), persons of high social standing, e.g. a lord (*wyglądać jak panisko*), and characters from fairy tales, e.g. a princess (*wyglądać jak księżniczka z bajki*), a queen (*wyglądać jak królewna*). There are also units motivated by the stereotyped perception of visual representations, in which good-looking persons can be seen, e.g. a picture (*wyglądać jak z obrazka*) or a fashion magazine (*wyglądać jak z żurnala*). It means that the calque in question employs a completely different imagery, which contributes to its attractiveness in the Polish language. As the borrowing is recent, the unit is not included in any lexicographic sources, even the *Wikisłownik* (Wiktionary) does not register it. The question arises how to determine its status and characteristics. In order to provide a proper lexicographic description, the information on the properties of the Polish phrase has to be collected. The aim of the paper is to show how various sources can be used in order to create a linguo-cultural presentation of the phrase at issue. The first source of information is *Narodowy Korpus Języka Polskiego* [the National Corpus of Polish], which contains over 1.5 billion of words. Another is the World Wide Web, which is not a corpus proper, but which will also be discussed in terms of its usefulness for such research. It is consulted for two reasons: first, in general idioms are not so frequent in corpora, which is also the case with the analysed unit; second, a great variety of texts are available on the Internet, allowing for searching kinds of texts not available in the consulted corpus. Moreover, one will analyse the implementation of a questionnaire aimed at eliciting information on the unit from Polish native speakers. The information gathered from these three sources will be the basis of a proposal of a monolingual entry for the phrase *wyglądać jak milion dolarów*.

References

BAŃKO, M. (2004). *Słownik porównań* [A dictionary of similes]. Warszawa: Wydawnictwo Naukowe PWN.

COLSON, J.-P. (2007). The World Wide Web as a corpus for set phrases. In H. Burger, D. Dobrovol'skij, P. Kühn and N.R. Norrick, eds. 2007. *Phraseologie. Phraseology. Ein internationales Handbuch zeitgenössischer Forschung. An International Handbook of Contemporary Research*, Vol. 1. Berlin and New York: Walter de Gruyter. pp. 1071-1077.

DOBROVOL'SKIJ, D. (2011). Cross-Linguistic Equivalence of Idioms: Does It Really Exist?. In A. Pamies and D. Dobrovol'skij, eds., 2011. *Linguo-*

Cultural Competence and Phraseological Motivation. Baltmannsweiler: Schneider Verlag. pp. 7-24.
FIEDLER, S. (2014). Gläserne Decke *und* Elefant im Raum — *Phraseologische Anglizismen im Deutschen*. Berlin: Logos Verlag.
FURIASI, C., PULCINI, V., AND RODRÍGUEZ GONZÁLEZ, F., eds. (2012). *The Anglicization of European Lexis*. Amsterdam and Philadelphia: John Benjamins.
KILGARIFF, A. (2001). Web as Corpus. In G. Sampson and D. McCarthy, eds., 2004. *Corpus Linguistics: Readings in a Widening Discipline*. London & New York: Continuum. pp. 471-473.
SAILER, M., (2007). Corpus linguistic approaches with German corpora. In H. Burger, D. Dobrovol'skij, P. Kühn and N.R. Norrick, eds., 2007. *Phraseologie. Phraseology. Ein internationales Handbuch zeitgenössischer Forschung. An International Handbook of Contemporary Research*, Vol. 1. Berlin and New York: Walter de Gruyter. pp. 1060-1071.
STEYER, K. (2004). Kookurenz. Korpusmethodik, linguistisches Modell, lexikographische Perspektiven. In K. Steyer, Hrsg., 2004. *Wortverbindungen – mehr oder weniger fest*. Berlin and New York: Walter de Gruyter. pp. 87-116.

Phraseology: theoretical and practical approaches Fraseología: enfoques teóricos y prácticos

ESTUDIO FRASEOLÓGICO BASADO EN EL CORPUS CORBICON

Arsenio Andrades

Universidad Complutense de Madrid

arsenioa@ucm.es

El objeto de esta propuesta de comunicación es presentar un estudio fraseológico basado en la compilación de un corpus bilingüe inglés-español de textos jurídicos con la finalidad de identificar y clasificar las expresiones fraseológicas más comunes.

Los estudios de fraseología del lenguaje general (Corpas, 1997; Ruiz Gurillo, 1998; García-Page, 2008) distinguen diferentes tipos de estructuras fraseológicas (colocaciones, locuciones, binomios y expresiones formulaicas) pero son escasas las investigaciones dedicadas a poner de manifiesto las características fraseológicas específicas de los textos jurídicos.

Desde una perspectiva fraseológica, el lenguaje jurídico anglosajón se caracteriza por un elevado empleo de un tipo de estructuras que se conocen como binomios o dobletes (Mellinkoff, 1963; Bhatia, 1994; Tiersma, 1999; Andrades, 2013b), pero también incluye otras expresiones fraseológicas recurrentes que aparecen bajo la forma de colocaciones o fórmulas hechas. A pesar de la importante presencia de este tipo de estructuras en el lenguaje jurídico anglosajón no han sido objeto de la suficiente atención por parte de los estudios de traducción en el par de lenguas inglés-español.

La metodología de trabajo se basa fundamentalmente en la compilación y explotación de un corpus bilingüe comparable y especializado, el CORBICON

(Corpus Bilingüe de Contratos de Derecho Civil), compuesto por textos originalmente redactados en inglés y español (Andrades, 2013a). El análisis de los datos mediante el programa de concordancias Wordsmith 5.0 hace posible la identificación y extracción de las principales fórmulas fraseológicas.

Un estudio contrastivo de los datos obtenidos y el cotejo de las diferencias y semejanzas de las unidades fraseológicas identificadas en los subcorpus inglés y español permiten presentar una propuesta de clasificación fraseológica aplicable al tipo de documentos jurídico examinado, que en futuros trabajos podría extrapolarse y adaptarse al discurso jurídico en general.

El establecimiento de una taxonomía fraseológica jurídica puede ser un instrumento útil a la hora de proponer distintas estrategias para abordar la traducción de las estructuras fraseológicas y facilitar la labor de búsqueda de equivalencias en el ámbito jurídico. En este sentido el corpus nos proporciona información que no se encuentra en los diccionarios (Bowker, 2002) ya que, como ocurre con la fraseología en general, su presencia suele ser escasa tanto en los diccionarios (Carvalho, 2008) como en las obras terminológicas (Cabré, 1998).

Este tipo de estudios basados en corpus subraya la importancia del conocimiento de la fraseología para el traductor jurídico (Monzó y Hoyo, 1998; Lorente, 2002; Aguado de Cea, 2007) y pretenden colmar la carencia de herramientas de trabajo eficaces en la traducción jurídica (Borja, 2004).

References

AGUADO DE CEA, G. (2007). La fraseología en las lenguas de especialidad. In: E. Alcaraz, José Mateo Martínez & Francisco Yus Ramos (eds.). *Las lenguas profesionales y académicas*. Barcelona: Ariel. pp. 53-65.

ALCARAZ, E. (1994). *El inglés jurídico. Textos y documentos*. Barcelona: Ariel.

ALCARAZ, E. (2000). *El inglés profesional y académico*. Madrid: Alianza Editorial.

ALCARAZ, E., CAMPOS, M. A. AND MIGUÉLEZ, C. (2001). *El inglés jurídico norteamericano*. Barcelona: Ariel.

ALCARAZ, E. AND HUGHES, B. (2002). *Diccionario Bilingüe de Términos Jurídicos: inglés-español, español-inglés*. Barcelona: Ariel.

ANDRADES, A. (2013ª). *Estudio contrastivo de las unidades fraseológicas especializadas (UFE) en un corpus comparable bilingüe de documentos jurídicos vinculantes en lengua inglesa y española*. Tesis doctoral inédita. Madrid: Universidad Complutense de Madrid.

ANDRADES, A. (2013b). La importancia de los binomios en la traducción jurídica. In: Ortega Arjonilla, E. (Dir.). *Translating Culture/Traduire la Culture/Traducir la cultura*. Vol. 3. Granada: Comares.

BHATIA, V. (1994). Cognitive structuring in legislative provisions. In: J. Gibbons (ed.), *Language and the Law*. Londres: Longman. pp. 136-155.

BLACK, H. C. (1891/1991). *Black's Law Dictionary*. St. Paul, Minn.: West Publishing.

BORJA, A. (2004). La investigación en traducción jurídica. In: M. A. García Peinado y E. Ortega (eds.). *Panorama actual de la investigación en traducción e interpretación*. Granada: Atrio. pp. 415-426.

BOWKER, L. (2002). *Computer-Aided Translation Technology: A Practical Introduction*. Otawa: University of Otawa Press.

CABRÉ, M. T. (1998). Las fuentes terminológicas para la traducción. In: P. Fernández Nistal y J. M. Bravo Gozalo, *La traducción: orientaciones lingüísticas y culturales*. Valladolid: Universidad de Valladolid.

CARVALHO, L. (2008). Translating contracts and agreements: a *corpus* linguistic perspective. In: S. E. O. Tagnin y O. A. Vale. *Avanços da Linguística de Corpus no Brasil*. São Paulo: Humanitas.

CORPAS, G. (Ed.) (2003). *Diez años de investigación en fraseología: análisis sintáctico-semánticos, contrastivos y traductológicos*. Frankfurt/Madrid: Vervuert/Lingüística Iberoamericana.

CORPAS, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt: Peter Lang.

GARCÍA-PAGE, M. (2008). *Introducción a la fraseología española*. Barcelona: Anthropos.

GRANGER, S. (2009). Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. *Journal of Shanghai Jiaotong University*, [online] Available at: <http://sites.uclouvain.be/cecl/archives/Granger_Crosslinguistic_research.pdf> [Accessed 10 January 2015]

GUSTAFSSON, M. (1975). *Binomial expressions in present-day English: a syntactic and semantic study*. Turku: Turun yliopisto

GUSTAFSSON, M. (1984). The syntactic features of binomial expressions in legal English, *Text*, 4 (1-3), p. 123.

LORENTE, M. (2002). Terminología y fraseología especializada: del léxico a la sintaxis. In: G. Guerrero y L. F. Pérez Lagos (eds.), *Panorama actual de la terminología*. Granada: Comares. pp. 159-180.

MELLINKOFF, D. (1963). *The Language of the Law*. Boston: Little, Brown & Co.

MONZÓ, E. AND HOYO, E. (1998). La traducció dels textos jurídics al DOGV, *Fòrum de Recerca*, 3. [online] Available at: <http://sic.uji.es/publ/edicions/jfi3> [Accessed 10 January 2015]

READ, J. AND NATION, P. (2004). Measurement of formulaic sequences. In: N. Schmitt. (ed.). *Formulaic Sequences*. Amsterdam: John Benjamins. pp. 23-35.

SCOTT, M. (2012). Wordsmith Tools 6.0. [online] Available at: <http://www.lexically.net/wordsmith/index.html> [Accessed 10 January 2015].

SINCLAIR, J. (2005). Corpus and Text - Basic Principles. In: M. Wynne (ed.). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, [online] Available at:

<http://ota.ahds.ac.uk/documents/creating/dlc/chapter1.htm> [Accessed 10 January 2015].

TEUBERT, W. (2005). My Version of Corpus Linguistics. *International Journal of Corpus Linguistics*, 10, p. 1.

TIERSMA, P. (1999). *Legal Language*. Chicago/London: The University of Chicago Press.

ZANETTIN, F. (1998). Bilingual Comparable Corpora and the Training of Translators. *Meta*, XLIII, 4, [online] Available at: [http://www.erudit.org/erudit/meta/v43n04/] [Accessed 10 January 2015]

“NICE CLEAN SPRAYS OF BLOOD”: SUBTITLING ANOMALOUS COLLOCATES IN CRIME TV SHOWS

Blanca Arias

Universitat Pompeu Fabra

blanca.arias@upf.edu

More than a million people in Spain turn on their TVs every night and watch hit American crime television series (Noxvo 2004-2015). Like any genre, crime entails recognition: These audiovisual texts are easily recognizable by audiences through the visual channel (eg. costumes, spaces, weapons) and the audio channel (eg. doors howling, shouts, sirens). The corpus study reported in this paper, which is currently in progress, shows how this kind of series also portrays identifiable linguistic traits. Typically, the first set of elements that springs to mind in this sense are terminological units, which frequently occur in the medical descriptions of forensic analyses, or in interrogation sequences where, for example, suspects are made aware of their rights (cf. Arias-Badia and Brumme 2014). There is, however, a second salient set of linguistic features that often goes unnoticed as key to the genre, i.e. creative uses of language.

This paper explores the preliminary results obtained after applying/adapting corpus pattern analysis and the distinction made by Hanks (2004, 2013) between norms and exploitations to a study of three different American crime TV shows, i.e. *Dexter* (2006), *The Mentalist* (2008) and *Castle* (2009).

Specifically, the research hereby reported focuses on the way in which one specific kind of exploitation, i.e. anomalous collocates, is rendered in the Spanish DVD subtitles of the shows. On a first stage, the paper aims to establish which unusual collocations in the source text (ST) could be classified as genre-specific, where some eligible candidates would be *neat monster*, *good old-fashioned murderer, nice clean sprays of blood*, *to work dead bodies* or *gruesome souvenir*; and which anomalous collocates rather portray general stylistic markers of the speech of a character, such as *to rush genius*, *to drink a meal* or *to sign life away*.

After a proposal for the classification of anomalous collocates from the ST, the paper describes the process whereby these creative units get translated into the target text (TT). Therefore, as a theoretical framework, the undertaken analysis incorporates the translation procedures traditionally acknowledged by authors from the field of Translation Studies, such as Vinay and Darbelnet (1958), Reiss (1971) or Newmark (1988).

While previous work (Arias-Badia, in press) has demonstrated that terminology is frequently preserved in the TT, the reason to turn attention to creative language is that this kind of language present in the ST disappears or is 'neutralized' in up to a 40% of the instances in the TT. The research reported in this paper intends to find out whether anomalous collocates fall within the category of exploitations that more typically tend to be neutralized in the subtitling process.

References

ARIAS-BADIA, B. (in press). Towards a Methodology for the Analysis of Neutralisation in Spanish Subtitling. In: G. Corpas, M. Seghiri, R. Gutiérrez, M. Urbano, eds. *New Horizons in Translation and Interpreting Studies. Proceedings of the 7th AIETI Conference*. Geneva: Tradulex.

ARIAS-BADIA, B., AND BRUMME, J. (2014). Subtitling stereotyped discourse in the crime TV series *Dexter* (2006) and *Castle* (2009). In: *JoSTrans. The Journal of Specialised Translation* 22. pp. 110-131.

Castle (2009). La primera temporada completa. English original version; Castilian Spanish subtitles. DVD. [s. l.]: ABC Studios. Distributed by Walt Disney Studios Home Entertainment. Episodes 1-2.

Dexter (2006). Primera temporada. English original version; Castilian Spanish subtitles. DVD. Madrid: Showtime. Episodes 1-2.

El mentalista (2008). Primera temporada completa. English original version; Castilian Spanish subtitles. DVD. [s. l.]: Warner Bros. Episodes 1-2.

HANKS, P. (2004). The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography*, 17(3). pp. 245-274.
HANKS, P. (2013). *Lexical Analysis. Norms and Exploitations.* Cambridge, MA: MIT Press.
NEWMARK, P. (1988). *A textbook of Translation*. London: Prentice Hall.
Noxvo (2004-2015). *Formula TV*. Data audited by OJDinteractiva Auditoría Medios Online, [online]. Available at http://www.formulatv.com [Accessed 11 April 2015].
REISS, K. (1971). *Möglichkeiten und Grenzen der Übersetzungskritik*. Munich: Max Hueber.
VINAY, J. P. and DARBELNET, J. (1958). *Stylistique compare du français et de l'anglais.* Paris: Didier.

SPORTS METAPHORS IN ENGLISH LEGAL DISCOURSE

Vesna Cigan
University of Zagreb
vesna.cigan@gmail.com

Darija Omrčen
University of Zagreb
darija.omrcen@kif.hr

Sports metaphors are figurative expressions frequently used in business, political and legal discourse. The concepts of *sprint, marathon* and *raising the bar* that originate in track-and-field as a source domain, a boxing-related *below the belt*, *kick-off* in soccer, or *game, set, match* emanating from tennis are well known both to general public as well as to particular domain experts.

As for the legal discourse, they are used by judges, researchers and journalists alike, thus flavouring the otherwise rather dry and strict discourse of court decisions, their newspaper commentaries or research articles of law-related subject matters.

Past research on the subject matter mostly had two foci. On the one hand, it focused on the appropriateness of using sports metaphors in judicial context (Oldfather, 1994), for example, when explaining a court decision, or when identifying judges with umpires (Zelinsky, 2010). This was done by analysing sports metaphors' meaning both in the source domain, i.e. sports, and in the target domain, i.e. law. Some authors view this matter in the sense that sports metaphors are superfluous and that they jeopardize the unambiguousness of legal discourse in general. Others juxtaposed, to a greater or lesser extent, this point of view thus advocating the usage of metaphors (e.g. Benforado, 2011). On the other hand, past research concentrated on sports metaphors used in various types of legal context, e.g. judicial opinions, opening statements, etc.

(Boyd, 2014). Such analyses were substantiated by listing numerous examples of using sports metaphors in various types of texts.

The analysis in this paper will rely on sports metaphors' semantic functions in law-related target domains. Following the methodology applied by Boyd (2014) and Abrams (2010), sports metaphors will first be classified by sport from which they originate, followed by the explanation of their meaning in source and target domains. Further, sports metaphors will be scrutinized by the type of text in which they have been used – be it, for example, a court decision, a newspaper/magazine commentary or a law-related professional text (e.g. Sampsell-Jones, 2010). Analysis will be directed towards a sample comprising sports metaphors in which those that have not been analysed previously in past research will prevail. Some examples of metaphors will also be discussed upon cross-tabulating sport and the type of text in which they were found. Finally, attention will be drawn to cultural differences that govern the usage of certain sports metaphors.

References

ABRAMS, D.E. (2010). Sports in the courts: the role of sports references in judicial opinions. *Villanova Sports & Entertainment Law Journal*, 17(1), pp. 1-57.

BENFORADO, A. (2011). Color commentators of the bench. *Florida State University Law Review*, 38(3), pp. 451-479.

BOYD, M.E. (2014). Riding the bench – a look at sports metaphors in judicial opinions. *The Harvard Journal of Sports and Entertainment Law*, 5(2), pp. 245-264.

OLDFATHER, C.M. (1994). The hidden ball: a substantive critique of baseball metaphors in judicial opinions". [online] *Faculty Publications.* Paper 475. Available at: http://scholarship.law.marquette.edu/facpub/475 [Accessed 20 March 2015].

SAMPSELL-JONES, T. (2010). On silence: a reply to professors Cribari and Judges. *Faculty Scholarship.* Paper 167. Available at: http://open.wmitchell.edu/facsch/167 [Accessed 20 March 2015].

ZELINSKY, A.S.J. (2010). The justice as commissioner: benching the judge-umpire analogy. *The Yale Law Journal Online*, 119, pp. 113-125.

THE ADAPTATION OF ANGLICISMS – PHRASEOLOGICAL UNITS IN CROATIAN ECONOMIC TERMINOLOGY

Ivo Fabijanić

University of Zadar

ivo.fabijanic@unizd.hr

Lidija Štrmelj

University of Zadar

lstrmelj@unizd.hr

In our previous studies on contact between English and Croatian in economic and medical terminology, several new trends in the adaptation of anglicisms were noticed. Firstly, the contact has become more profound at a general, lexical level and secondly, the same trend of their adaptation is evident at a narrower (more specialized) phraseological level. At the former one, the anglicisms were analysed as polylexemic units, adapted to Croatian in a variety of ways, and showing different degrees of their adaptation, i.e. through the primary or secondary adaptation, and realized according to the concepts of the zero, compromise and complete transmorphemization of nominal syntagms. In this research, referring to the latter phraseological level, we are interested in finding practical solutions for a) the classification of anglicisms – phraseological units in Croatian, on the material of economic terms, within the following groups of PUs: phrasal compound idioms, unilateral idioms, phraseological nominations, and restricted collocations, and b) determining the degrees of their adaptation in the Croatian language. Due to the fact that the PUs we are interested in, can be grouped within the confines of different specific types, a type of feasibility test will be conducted to support the decision-making process in their classification. The test will provide data about various features of PUs

(e.g. their structural features, semantic stability, idiomaticity, etc.) on the basis of which, they will be classified into the appropriate class of the aforementioned types of PUs. In the second part of the research, after having done the typological part, the analysis of approximately one hundred phraseological units will be conducted, which, according to our preliminary studies, show a high probability of typological classification of adapted anglicisms in a way similar to that by which nominal syntagms were analysed. Our preliminary study has shown that the principal degrees of adaptation of anglicisms – phraseological units might be defined within the following frame of degrees: 1) the zero degree of adaptation for the PUs which retained their original structure and the order of structural elements in the receiving language, together with the original degree of idiomaticity (e.g. *flat tax* > *flat tax*, *floor broker* > *floor broker*, *front runner* > *front runner*); 2) the degree of compromise adaptation for the PUs which retained their original structure and morphological elements of the giving language (e.g. *charter party* > *čarter party*, *fleet manager* > *fleet manager, hot bunking* > *hot bunking*, *lead manager* > *lead manager*) with a relatively variable degree of idiomaticity, and 3) the degree of complete adaptation for the PUs whose original structure and the order of elements hasn't been retained, i.e. the degree of variability/alteration in replicas has substantially risen, resulting either in a changed order of their structural elements or in different changes of their morphological structure (e.g. *financial pyramid* > *financijska piramida*, *credit line* > *kreditna linija*, *output effect* > *efekt outputa*), together with a relatively variable degree of idiomaticity.

References

AJDUKOVIĆ, J. (2012). *Radovi iz lingvističke kontaktologije*. Beograd: Foto-futura.

ALLERTON, D.J. NESSELHAUF, N. AND SKANDERA, P. eds. (2004). *Phraseological Units: basic concepts and their application*. Basel: Schwabe.

FABIJANIĆ, I. (2009). *Anglizmi u ruskoj i hrvatskoj ekonomskoj terminologiji*. doktorska disertacija u rukopisu. Zadar.

FABIJANIĆ, I. (2011). Reinterpretacija transmorfemizacije anglizama – imeničkih sintagma u ruskome i hrvatskom jeziku. *Fluminensia*. 23(1), pp. 67-83.

FABIJANIĆ, I. AND MALENICA, F. (2013). Abbreviations in English Medical Terminology and their Adaptation to Croatian. *JAHR*. 4(7), pp. 71-105.

FIEDLER, S. (2007. *English Phraseology: a Coursebook*. Tuebingen: Narr.

FINK-ARSOVSKI, Ž. (2002). *Poredbena frazeologija: pogled izvana i iznutra*. Zagreb: Filozofski fakultet.

FIRTH, J. R. (1957/1986). A Synopsis of Linguistic Theory, 1930-55. In: Selected Papers of J.R. Firth 1952-59. Ed. by F.R. Palmer (1968). London: Longman, pp.168-205.
ROOS, E. (2001). *Idiom aund Idiomatik: ein sprachliches Phaenomen im Lichte der kognitiven Linguistik und Gestaltheorie*. Achen: Shaker.

IDENTIFICATION AND ACQUISITION OF MULTI-WORD TERMS IN BUSINESS ENGLISH DOMAINS

Tatiana Fedulenkova

Vladimir State University

fedulenkova@list.ru

Context contributes much to identification of multi-word terms and their variants. If a L2 learner encounters a statement as follows *'My foot! A Dutch treat of dumb Dora in the laughing academy!'*, their first thoughts might be about a person's limb connected in a way with some exquisite dish provided by a feminine named Dora, temporarily unable to speak, who might be studying or giving a party at a kind of educational establishment where they are used to laugh a lot. The puzzle is clear for those who happen to know that *my foot* is a comment expressing disbelief (Gulland), *a Dutch treat* denotes an outing, entertainment, social gathering, etc where each person pays his own share of the expenses (Cowie), *dumb Dora* stands for a giddy woman (Spears) and *the laughing academy* is an informal name for a lunatic asylum and institution for the care of mentally handicapped, or mentally ill, people (Cowie). The absence of the context fails the adequate comprehension of the set of multi-word terms.

So identification of multi-word terms is a back-breaking process which, on the one hand, needs context specializing their meaning, on the other hand, requires knowledge.

This is of especial importance in the field of business and finance where a mistake or even a mere a slip is sure to cost one much because the divergence

of the form and the meaning of the term may lead one to a crucial misunderstanding, as with terms in the domains: (a) business *a sleeping partner* – a person who provides a percentage of the capital of a business but who does not have a part in the management of a business (Seidl), also *gravy train, little dragons, nest egg,* etc.; (b) commerce: *period of grace* – additional time that is allowed before a payment must be made (Longman); also *halo effect, blanket agreement,* etc.; (c) finance: *dead cat bounce* – an occasion when a share price or stock market rises a small amount after a large fall before falling further (Longman); also *call feature, green-shoe option*, etc.; (d) stock exchange: *bull position* – when you possess particular bonds, shares, currencies, believing that their value will rise (Longman); also *grey knight, black market,* etc.

The exuberance of newly coined multi-word terms in a majority of Business English domains may be explained by the analytical character of English. As far as acquisition of multi-word terms is concerned, one of the algorithms embraces a three-stage study of them, namely: a) the first stage aiming at acquisition of collocations: *general practice, sharp practice, best practice*, b) the second stage aiming at acquisition of multi-word terms with partial transfer of meaning: *blanket insurance, Robert's rules, pester power*, c) the third stage aiming at acquisition of multi-word terms having full transfer of meaning: *shark repellent, soft landing, poison pill, red tape,* etc.

The general and reliable rule is to appeal to dictionary definitions of multi-word terms.

References

COWIE, A.P., MACKIN, R., MCCAIG, I.R. (1984). Oxford Dictionary of Current Idiomatic English. Vol. 2: Phrase, Clause and Sentence Idioms. Oxford: Oxford University Press.

FEDULENKOVA, T. (2002). Idioms in Business English: Ways to Cross-cultural Awareness. In Giuseppina Cortese & Philip Riley (ed.), *Domain-specific English: textual practices across communities and classrooms,* 247-269. Bern; Berlin; Bruxelles; Frankfurt am Mein; New York; Oxford; Wien: Lang.

FEDULENKOVA, T. (2014). Basic Components of the Connotative Aspect in Phraseological Units: (As Seen by A.V. Kunin and his Disciples). In *Phraseology in Multilingual Society*, 34-47. Newcastle upon Tyne: Cambridge Scholars Publishing.

GULLAND, D.M., HINDS-HOWELL, D.G. (1994). The Penguin Dictionary of English Idioms. Harmondsworth: Penguin Books Ltd.

SEIDL, J., MCMORDIE, W. (1978). English Idioms and How to Use Them. Oxford: Oxford University Press.

SPEARS, R. (1991). Dictionary of American Slang. Lincolnwood: National Textbook Company.
URDANG, L. (ED.). (1996). Longman Dictionary of English Idioms. Harlow and London: Longman Group UK Ltd.

A CORPUS-BASED APPROACH TO LEXICOGRAPHY: TOWARDS A THESAURUS OF TATAR IDIOMS

Guzel Gizatova

Kazan State Agrarian University

guzelgizatova@hotmail.com

This paper deals with the principles of constructing a new ideographic dictionary of Tatar idioms (IDTI) based on corpus data. Idioms are arranged by their figurative meaning rather than alphabetically. The Tatar language spoken by 7 million people belongs to a Turkic branch of an Altaic family of languages[11]. The purpose of this paper is to explore Tatar idioms along two revealing lines of enquiry. The first line of enquiry suggests constructing an ideographic dictionary of Tatar idioms; the second line illustrates advantages of organizing such dictionary on an analysis of corpus data.

The need for a new dictionary is motivated by the fact that there is no ideographic description of Tatar idioms based upon the principle “from concept to sign”. Such approach enables to find all idiomatic word combinations of the language that express the given concept. Theoretical basis of constructing the dictionary is the strategy developed by Anatoli Baranov and Dmitri

[11]Information about the Tatar language (and recent idiom research of Tatar) can be found in: Gizatova, Guzel (2014): On vanishing Tatar idioms. In: Abstracts of the EUROPHRAS 2014 conference, 10-12 September 2014 in Paris Sorbonne University, 2014. pp. 22-23 (Proceedings in print); Gizatova, Guzel (2015): A nation without a language is a nation without heart: In: Piirainen, Elisabeth; Sherris, Ari (eds.): Language Endangerment. Disappearing Metaphors and Shifting Conceptualizations. Amsterdam/Philadelphia: John Benjamins, 179-202.

Dobrovol'skij[12] – the authors of the first profound pioneer Thesaurus of idioms in international lexicography. Apart from its theoretic relevance as an instrument of description of the mental lexicon, a new ideographic dictionary of Tatar idioms can be used for purposes of language acquisition and translation.

The main important difference of IDTI from existing Tatar idiom dictionaries is in its orientation on modern authentic data drawn from two corpora. The first one is the Text Corpus of the modern Tatar language, a 116-million-word corpus including written texts and the second one is the National Corpus of the Tatar Language, a 50-million-word corpus comprising both written and spoken language. Modern technologies allow handling large amounts of data, which are essential in presenting objective information on idiom use.

The article describes the ways in which corpus data can be used to present a more detailed and weighed analysis of idioms under consideration, show a range of syntactic patterns and unexpected variants which cannot be retrieved from the existing idiomatic, bilingual and monolingual dictionaries of the Tatar language. And more, the corpus data, which have been received in the process of research into two Tatar corpora, give evidence of the new perspectives about co-occurrence patterns, semantic and combinatorial properties of idioms and their frequency in various types of corpora (e.g., media versus literary or academic prose).

The results of corpus analysis give an opportunity to reveal new aspects of meanings of Tatar idioms, which have not been yet fixed in dictionaries or cases of their inaccurate definitions. The findings indicate that idioms can only be understood to full extent if they are considered in discourse in which they occur.

Switching over to a corpus-based approach in lexicography is a question of vital importance for Tatar linguistics.

[12] Baranov, A. N. and D.O. Dobrovol'skij (eds.) (2007). Thesaurus of Modern Russian Idioms.

METAPHORICAL PHRASEOLOGICAL UNITS OF SANITATION IN MARIANO RAJOY'S POLITICAL DISCOURSE

María José Hellín-García
The Military Collge of South Carolina
mhelling@citadel.edu

Metaphors are effective tools for political argumentation as they can contribute to persuade and influence other's beliefs and attitudes. Furthermore, their multi-functionality may fulfill intended purposes to achieve the speaker's intentions with his or her audience. This corpus-based study focuses specifically on the metaphorical phraseological units that conceptualize the notion of sanitation. I particularly investigate how these phraseological units serve as a persuasive means to frame the political debate on economy. Furthermore, this study also explores what the most frequent metaphorical units of sanitation are, and how they are mediated in the argumentation. The corpus of investigation includes political speeches from the current Prime Minister of Spain, Mariano Rajoy, during 2011 and 2012. I argue that this conscious metaphorical phraseological units of sanitation is Rajoy's main argumentative strategy to justify his political action in undertaking new drastic reforms to revitalize the economy in Spain. I also claim that the need for a radical 'cleaning up' of the economic system is a rhetorical strategy to discredit the previous socialist government of Rodríguez Zapatero. As this study investigates the role of phraseological units in relation to metaphor within a political corpus, the method of analysis adopted combines a cognitive and a pragmatic perspective. The cognitive perspective is based on Conceptual Metaphor Theory (Johnson 1987,

Lakoff and Johnson 1980, 1999), and the pragmatic perspective is based on Critical Metaphor Analysis (Charteris-Black 2004, 2005). This study contributes to the understanding of the relation of phraseological units and metaphor in discourse as a potential argumentative tools in politics.

References

Charteris-Black, Jonathan. (2004). Corpus Approaches to critical metaphor analysis. New York: Palgrave MacMillan.

Charteris-Black., Jonathan. (2005). Politicians and rhetoric: the persuasive power of metaphor. New York: Palgrave MacMillan.

Johnson, Mark. (1987). The body in the mind. Chicago: The University of Chicago Press.

Lakoff, George and Mark Johnson. (1980). Metaphors We Live By. The University of Chicago Press.

Lakoff, George and Mark Johnson. (1999). Philosophy in the Flesh: The Embodied Mind and Its Challenge to the Western Thought. Cambridge: Perceus Publishing.

BRAZILIAN CULTURE THROUGH ITS METAPHORS: A MULTILINGUAL CONTRASTIVE APPROACH

Rosemeire Monteiro-Plantin
Universidade do Ceará
meire@ufc.br

Antonio Pamies-Bertrán
Universidad de Granada
antonio.pamies@gmail.com

Lei Chunyi
Universidad de Granada
leichunyi@hotmail.com

It is a well known fact that plants are one of the most productive source domains for conceptual metaphors, after human body and animals, and that their role in language is mediated by several kinds of cultural symbolism (Dobrovol'skij & Piirainen 2005; Pamies & Tutáeva 2011; Pamies 2014). We intend to show that mandioca("cassava") is a cultureme very representative of Brazilian Portuguese.

We analyze the importance of this plant in the construction of cultural identity of Brazilian people, taking into account etymological, historical, anthropological, and, more specifically, linguistic factors, especially as reflected in phraseology. This basic component of Brazilian diet is the center of a complex network of values and social relations, which makes it an ethnobotanic identitary icon, due to its anthropological and symbolic dimensions.

From the linguistic point of view, we collect and analyze collective mental images related to the concept of mandioca, as embedded in numerous metaphorical and/or metonymic processes such as the dialectal idioms*ser pior do que o caso da mandioca, comer com farinha, farinha pouca meu pirão primeiro, tirar farinha (com alguém), comer mingau pelas beiradas, ser*

farofeiro, etc. which refer to Brazilian realities, values and beliefs. Looking for total or partial parallelisms, we compare this phenomenon with culturemes from other languages and cultures, such as bread in Europe and rice in Asia.

References

DOBROVOL'SKIJ, D. O. (1998). "On Cultural Component in the Semantic Structure of Idioms". In: Ďurčo, P. (ed.). *Phraseology and Paremiology*. Bratislava: Akadémia PZ, 55-61.

DOBROVOL'SKIJ, D. O. & PIIRAINEN, E. (2005). *Figurative Language: Cross-cultural and Cross-linguistic Perspectives*. Amsterdam: Elsevier.

LUQUE DURÁN, J. d. D. & LUQUE NADAL, L. (2008). "Cómo las metáforas recurren a conocimientos ontológicos y culturales. Fundamentos teóricos del Diccionario Intercultural e Interlingüístico". In: Korhonen, J. et al. (eds.): *Phraseologie - Global - Areal - Regional.* Tübingen: G. Narr.

LUQUE NADAL, L. (2009). "Los culturemas: ¿unidades lingüísticas, ideológicas o culturales?", *Language Desig*n, 11: 93-120.

MONTEIRO, R. S. (2011). "Gastronomismos fraseológicos: um olhar sobre fraseologia e cultura". In: Ortiz Alvarez, M. L.; Huelva Unterbäumen, E. (eds.). *Uma (re)visão da teoria e da pesquisa fraseológicas*. Campinas: Pontes & Universidad Nacional de Brasília: 249-276.

PAMIES, A. (2007). "El lenguaje de la lechuza. Apuntes para un diccionario intercultural", In: Luque J.d.D. & Pamies, A. (eds.): *Interculturalidad y lenguaje I. El significado como corolario cultural*. Granada: Granada Lingvistica / Método: 375-404.

PAMIES, A. (2008). "Comparaison inter-linguistique et comparaison interculturelle". In: Michel Quitout (ed.) *Traduction, proverbes & Traductologie.* Paris: Éditions L'Harmattan, pp. 143-156.

PAMIES, A. (2011). "Motivación cultural y botanismos gastronómicos". In: Ortiz Alvarez, M. L. & Huelva Unterbäumen, E. (eds.). *Uma (re)visão da teoria e da pesquisa fraseológicas*. Campinas: Pontes & Universidad Nacional de Brasília: 49-68.

PAMIES. A. (2014). "Provérbios fitonímicos e plantas proverbiais". In: S. Silva (ed.) *Fraseologia & Cia: entabulando diálogos reflexivos,* vol II. Campinas (S.P): Pontes: 79-104.

PAMIES, A. & LEI, Chunyi (2014). *L'intraduisible? Dîtes-le avec des fleurs: Botanismes figuratifs et spécificité culturelle*. In: Dalmas, M.; Piirainen, E. & Filatkina, N. (eds.): *Figurative Sprache / Figurative Language / Langage figuré: Festgabe für Dmitrij O. Dobrovol'skij*: Tübingen: Stauffenburg (Linguistik, Band 83): 19-40.

STUDY ON TRANSLATION ERRORS IN KOREAN-SPANISH PHRASEOLOGICAL EXPRESSIONS BY MACHINE TRANSLATION AS PART OF LOCALIZATION

So Young Park

Hankuk University of Foreign Studies

soyoungines@naver.com

The Korean government is providing massive support for translation of Korean cultural content, with an aim of promoting the “Korean Wave” or “Hallyu” around the world. Against this backdrop, localization has been introduced in place of traditional translation where a translator performs translation from one language into another and then a native speaker of the target language proofreads or revises it when necessary. Localization in this context is more than a linguistics transfer: it involves a transfer of contexts, culture and sentiments of the end users of the localized products. It comprises not just traditional linguistic activities such as translation and proofreading, but also linguistic and cultural research, considerations of technical aspects, and selection of the type of medium. It also involves different agents from clients to agencies and linguistic professionals who work on localization with their specific linguistic knowledge. During the localization process, when the source text is mainly composed of short sentences as in games and animations, these

linguistic professionals make use of machine translation, a technology provided by many IT companies such as Google, in order to improve work efficiency. In such cases, the translation between languages that share similar origins and linguistic features has a good chance of success as in translation between Spanish, French, and Italian or between Korean and Japanese. On the other hand, limitations do exist to what can be translated between languages like Spanish and Korean that display large cultural and linguistic differences. Therefore, with the current level of technology, linguistic transfer, that is, translation error is inevitable in the localization process. In particular, when the translation requires knowledge on sociocultural elements of the source language, such as multiword phraseological expressions, it is very difficult to produce a successful translation.

This study conducts a contrastive analysis of a piece of long animation series translated from Korean into Spanish for children, focusing on translation errors produced by machine translation. The data consists of 1,500 words of the script, or roughly 20 minutes of the animation, which has been translated first into English and then into Spanish by a native local translator. The study found that machine translation left traces of errors on lexical, grammatical and sociolinguistic levels. It categorizes errors in phraseological expressions translated from Korean into Spanish as a basis for criteria for proofreading and revising translations. The results are expected to help translators in their efforts to improve translation quality as well as serve as useful references for translation trainees in their training of revising translation as part of the localization process.

ITALIAN MULTIWORD ADVERBS: DISTRIBUTIONAL FEATURES AND FUNCTIONAL PROPERTIES. A CORPUS BASED ANALYSIS

Valentina Piunno

Roma Tre University

valentina.piunno@uniroma3.it

According to typological studies, languages can employ several lexical devices to express the adverbial modification (among others, Van der Auwera 1988, Hengeveld 1992 and references therein); furthermore, the category of adverbs is semantically, morphologically and syntactically heterogeneous (among others, Van der Auwera 1988, Givón 2001, Haspelmath 2001).

In Italian, as well as in other Romance language, the adverbial function can be realized by different lexical elements, such as single lexical items or adverbials constructions – mainly nominal phrases and prepositional phrases or propositions – (among others, for Italian, Ramat and Ricca 1994, Lonzi 2001; for Spanish, Bosque and Demonte 2009, Bosque 2010; for French, Gross 1990). As far as Italian adverbial constructions are concerned, the most common syntactic configuration is the prepositional phrase (Voghera 2004, Piunno 2013). This investigation will deal with prepositional phrases covering an adverbial function and showing a high degree of cohesion between their constituents (hereinafter *Multiword Adverbs-PPs*), such as: *di solito* 'usually', *a occhio e croce* 'more or less', *per filo e per segno* 'chapter and verse'.

From a functional point of view, Italian Multiword Adverbs-PPs can fulfil different syntactic functions: they can appear as verb modifiers (1), as adverb modifiers (2), as adjective modifiers (3) or as sentence modifiers (4):

(1) Lucia esce *di rado*
'Lucia goes out rarely'
(2) Marina arriverà *al più tardi* domani
'Marina will arrive tomorrow at the latest'
(3) Ho mangiato un panino *a dir poco* eccellente
'I ate a simply great sandwich'
(4) *A parte gli scherzi*, il film era davvero noioso
'Joking aside, the movie was really boring'

From a distributional point of view, depending on their functional properties, Italian Multiword Adverbs-PPs can cover different syntactic positions, i.e.: a) after the adjective or the adverb, b) between the subject and the verb, c) between the auxiliary and the verb, d) after the verb, e) at the beginning of a sentence, f) at the end of a sentence.

This investigation analyses a range of about one thousand Multiword Adverbs-PPs extracted from the Italian corpus of written language *La Repubblica* (Baroni et al. 2004).

On the one hand, the quantitative corpus analysis will demonstrate that this analytical resource is not only very productive, but it is also particularly exploited in Italian (Piunno 2013).

On the other hand, the computational analysis will reveal to which extent the distribution of Multiword Adverbs-PPs may vary depending on their functional properties, i.e.: a) when they modify a verb, their location is preferably after it; b) when they modify an adjective or an adverb, they are usually located before them; c) if they function as sentence modifiers, their syntactic distribution is variable (they are usually located in starting or ending position).

The corpus-based investigation will allow us to provide a classification in subcategories, according to the distributional properties as well as to the functional and semantic value of sets of Multiword Adverbs-PPs.

References

BARONI, M. *et al.* (2004). Introducing the La Repubblica corpus: a large, annotated, TEI(XML)-compliant corpus of newspaper Italian", In: M. T. Lino *et al.*, eds. *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004, Lisbon, may 26-28). Paris: ELRA - European Language Resources Association. pp. 1771-1774.

BOSQUE, I. (2010). *Nueva gramática de la lengua española*, Madrid: Real Academia Española.

BOSQUE, I. AND DEMONTE, V. eds. (2009). *Gramática descriptiva de la lengua española*. Madrid: Espasa.

CINQUE, G. (1999). *Adverbs and Functional Heads. A Cross-Linguistic Perspective*, Oxford University Press.

GIVÓN, T. (2001). *Syntax. An introduction*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

GROSS, M. (1990). *Grammaire transformationnelle du français*. 3. Paris: Asstril.

HASPELMATH, M. (2001). Word Classes and Parts of Speech. In: P. Baltes and N. Smelser, eds. *International Encyclopedia of the Social and Behavioral Sciences*. Amsterdam: Pergamon. pp. 16538–16545.

HENGEVELD, K. (1992). *Non-verbal Predication: theory, typology, diachrony*. Berlin: Mouton de Gruyter.

La Repubblica: http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica

LONZI, L. (2001). Il sintagma avverbiale, In: L. Renzi, and G. Salvi, *Grande grammatica italiana di consultazione*. vol. 2. Bologna: Il Mulino. pp. 341-412.

PIUNNO, V. (2013). *Modificatori sintagmatici con funzione aggettivale e avverbiale.* PhD Thesis. Roma: Università Roma Tre.

RAMAT P. AND RICCA D. (1994). Prototypical adverbs: on the scalarity / radiality of the notion of adverb*, Rivista di Linguistica.* 6 (2). pp. 289-326.

VAN DER AUWERA, J. (ed.) (1998). *Adverbial Constructions in the Languages of Europe*. Berlin/New York: Mouton de Gruyter.

VOGHERA, M. (2004). Polirematiche. In: M. Grossmann and F. Rainer, eds. *La formazione delle parole in italiano*. Tübingen: Max Niemeyer Verlag. pp. 56-69.

CONTRIBUTION OF MULTI-ELEMENT FEATURES IN AUTOMATIC TEXT CLASSIFICATION FOR AUTHORSHIP ATTRIBUTION

Antonio Rico Sulayes
Universidad de México
antonio.rico@uabc.eda

The use of multi-element units in computational linguistics has a long tradition that is reflected on both automatic text classification and the related forensic linguistic task of authorship attribution. Frequently performed by means of automated processes, the task of authorship attribution is aimed at assigning an anonymous piece of text to a subject within a list of potential authors. In order to achieve this goal, a constant proposal of new classificatory features has characterized the research on this task for a number of decades (Rudman, 1998). More recently, the number of features used in authorship attribution has exploded as all sorts of multi-element units have been introduced in the task, such as character, word and POS n-grams (Rico Sulayes, 2014). Easy to tag by computational tools, these multi-element units can produce long lists of features even in the small corpora –text collections authored by some set of subjects– which are standard in authorship attribution. The present study uses contributions to organized crime-related online forums to randomly create a number of corpora with an increasing number of subjects. Using these corpora to run several hundreds of experiments, two types of different classificatory features are tested: a rather short, previously selected list of multi-element

function words and a large list with all word unigrams in a given corpus. The combined set of all these features is fed to a suite of the most common and successful machine learning classifiers in this task. As will be reported here, the best results averaged by some classifier over all corpora are obtained after reducing the list of features by statistical techniques. In the significantly reduced sets of features that render the best performance, there are only a few of the multi-element function words previously selected; however, a further analysis of the features selected from the list of unigrams shows that many of these elements are either part of or they represent themselves multi-element units from a phraseological point of view (Corpas, 2013; Shanavas, 1996).

References

CORPAS PASTOR, G. (2013). Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In: I. Olza and E. Manero, eds. *Fraseopragmática*. Berlin: Frank & Timme. pp. 335-373.

RICO SULAYES, A. (2014). Técnicas de reducción, algoritmos resistentes al ruido o ambos. Opciones para el manejo de rasgos clasificatorios en la atribución de autoría. *Research in Computing Science*, 80, pp. 43-53.

RUDMAN, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31, pp. 351-365.

SHANAVAS, S. A. (1996). *Structure of a computational lexicon of Malayalam*. Ph. D thesis, Jawaharlal Nehru University. Available at:<http://shodhganga.inflibnet.ac.in/handle/10603/35036> [Accessed 28 February 2015].

LA MARCACIÓN DE LAS UNIDADES FRASEOLÓGICAS A PARTIR DEL EXAMEN DE CORPUS

Ana María Ruiz Martínez
Universidad de Alcalá
ana.ruiz@uah.es

Los diccionarios proporcionan información acerca de la particularidad del uso de las unidades fraseológicas por medio de etiquetas o marcas que acompañan a la definición. Por esta razón, tanto los diccionarios de lengua como los fraseológicos vienen siendo un instrumento de consulta obligada para que el traductor, el estudiante o el usuario de una lengua conozca el rasgo formal, informal, vulgar, etc. que tiene una determinada unidad fraseológica. Sin embargo, ni en los propios diccionarios ni en los trabajos que se ocupan de su redacción es habitual encontrar una explicación detallada de los criterios que utilizan los lexicógrafos para asignar las marcas correspondientes a cada unidad, lo que acarrea que nos podamos encontrar con que una misma unidad fraseológica presente etiquetas distintas en diferentes diccionarios o que algunos lexicógrafos se limiten a reproducir una marca de acuerdo con una tradición en la que esta no había quedado definida de una manera clara. Esta falta de sistematicidad y concreción en el empleo de algunas marcas la hemos observado de manera específica con la marca *lit* (literario), a partir del análisis que hemos llevado a cabo de diferentes clases de unidades fraseológicas extraídas del *Diccionario fraseológico documentado del español actual* (2004). El hecho de que este diccionario ofrezca mediante la citada marca información sobre el nivel de uso formal y elevado de las unidades fraseológicas

(información diafásica), además de otras indicaciones de distinta naturaleza (uso propio de obras literarias o de la lengua escrita), y el hecho de que para un buen número de ejemplos de unidades no exista una correspondencia con las informaciones ofrecidas por otros diccionarios de la lengua española, nos ha llevado a organizar nuestra investigación en torno a dos objetivos principales: 1) revisar los problemas que ocasiona a la fraseología la marca *literario* a tenor de los diferentes valores que puede aglutinar esta etiqueta dentro de la lexicografía española; y 2) exponer por qué los corpus de lengua deben convertirse en una herramienta de consulta imprescindible para lexicógrafos, lingüistas, traductores, fraseólogos o profesores de idiomas, a la hora de precisar cuáles son realmente las particularidades de uso que tienen algunas unidades fraseológicas, dado que el rápido acceso a una ingente cantidad de datos pone a nuestra disposición la posibilidad de realizar un análisis exhaustivo de las situaciones comunicativas, los tipos de textos, los discursos, etc. que facilitan la marcación de una determinada unidad.

References

BAKER, P. (2006). *Using Corpora in Discourse Analysis*. London / New York: Continuum.

BINON, J. et al. (2004). Tendances et innovations récentes en lexicographie pédagogique. La contribution des dictionnaires d'apprentissage. En: P. Battaner y J. Decesaris, eds. *De Lexicografia. Actes del I Symposium Internacional de Lexicografia (Barcelona, 16-18 de maig de 2002)*. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. pp. 53-79.

CAMACHO BARREIRO, A. M. (2008). Las marcas sociolingüísticas en una muestra de la lexicografía cubana: tipología y evolución. *Revista de Lexicografía*, XIV, pp. 43-58.

COLSON, J. P. (2008). Cross-linguistic phraseological studies: An overview. En: S. Granger y F. Meunier, eds. *Phraseology. An interdisciplinary perspective*. Amsterdam / Philadelphia: John Benjamins Publishing Company. pp. 191-206.

COLSON, J. P. (2010). The Contribution of Web-Based Corpus Linguistics to a GlobalTheory of Phraseology. En: S. Ptashnyk, E. Hallsteindóttir y N. Bubenhofer, eds. *Corpora, Web and Databases. Computer-Based Methods in Modern Phraseology and Lexicography*. Hohengehren: Schneider Verlag. pp. 23-35.

ETTINGER, S. (1982). La variación lingüística en lexicografía. En: G. Haensch et al. (eds. *La lexicografía. De la lingüística teórica a la lexicografía práctica*. Madrid: Gredos. pp. 359-394.

GARRIGA, C. (2003). La microestructura del diccionario: las informaciones lexicográficas. En: A. M. Medina Guerra, coord. *Lexicografía española*. Madrid: Ariel. pp. 103-126.

GÓMEZ MARTÍNEZ, M. y CARRIAZO RUIZ, J. R. coords. (2010). *La marcación en lexicografía histórica*. San Millán de la Cogolla: Cilengua.

GRANGER, S. y MEUNIER, F. (2008). Introduction. En: S. Granger y F. Meunier, eds. *Phraseology. An interdisciplinary perspective*. Amsterdam / Philadelphia: John Benjamins Publishing Company. pp. XIX-XXVIII.

GRIES, S. (2008). Phraseology and linguistic theory. En: S. Granger y F. Meunier, eds. *Phraseology. An interdisciplinary perspective*. Amsterdam / Philadelphia: John Benjamins Publishing Company. pp. 3-25.

OLÍMPIO DE OLIVEIRA SILVA, M. E. (2007). *Fraseografía teórica y práctica*. Frankfurt am Maim: Peter Lang.

PHILIP, G. (2008). Reassessing the canon: 'Fixed' phrases in general reference corpora. En: S. Granger y F. Meunier, eds. *Phraseology. An interdisciplinary perspective*. Amsterdam / Philadelphia: John Benjamins Publishing Company. pp. 95-108.

SINCLAIR, J. (2008). Preface. En: S. Granger y F. Meunier, eds. *Phraseology. An interdisciplinary perspective*. Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. XV-XVIII.

WOTJAK, G. (2006). *Las lenguas, ventanas que dan al mundo*. Salamanca: Ediciones Universidad de Salamanca.

ZGUSTA, L. (1988). *Lexicography Today: an annoted bibliography of the theory of lexicography*. Tübingen: Max Niemeyer Verlag.

PROPUESTA DE SUBTITULADO PARA PERSONAS SORDAS Y PERSONAS CON DISCAPACIDAD AUDITIVA DE LA SERIE *THE BIG BANG THEORY*

Esther Sedano Ruiz

Universidad de Málaga

esther.sedano@gmail.com

En este trabajo se aborda la propuesta de subtitulación y la posterior adaptación para sordos y personas con discapacidad auditiva de la serie televisiva norteamericana *The Big Bang Theory*. Esta serie se enmarca dentro de las comedias de situación (*sitcom*), donde encontramos elementos humorísticos que se sirven del humor, coloquialismos o frases hechas y que son importantes para la trama o para la caracterización de los personajes, por lo que es necesario reflejarlos de forma adecuada.

La traducción audiovisual ha disfrutado de un crecimiento exponencial en el ámbito académico en los últimos años. Sin embargo, la accesibilidad en los medios en este campo en concreto no se ha desarrollado tanto como otros hasta el momento. Asimismo, nos parece que la complejidad y la necesidad inminente de la subtitulación para personas sordas y con discapacidad auditiva tienen unas características merecedoras de estudio. Por esta razón, nuestra motivación está basada en la necesidad que las personas sordas o con discapacidad auditiva tienen de cara a ser partícipes de información que, gracias a las nuevas tecnologías, inunda la sociedad actual y enfatizar la

necesidad de mejorar los métodos existentes para la integración de estas personas en la que ya es conocida como la sociedad de la información.

Por otro lado, creemos que el tema de la subtitulación y adaptación para las personas sordas o con discapacidad auditiva es relevante debido a la gran difusión que en los últimos años están teniendo las series en versión original y que provienen de otros países. De este modo, la mayoría de los DVD solo incluyen subtítulos adaptados para sordos en versión original, en la mayoría de casos en inglés, por lo que el espectador español no puede acceder a este material. En el presente trabajo pretendemos hacer un recorrido por todas las fases de preparación y etapas de traducción y subtitulación hasta llegar al producto final. Para ello, seguiremos la norma UNE 153010 publicada en mayo de 2012, que contiene especificaciones sobre la presentación de este tipo de subtitulado, aúna algunas de las prácticas en particular y así, como medio de apoyo para facilitar la accesibilidad, especifica los requisitos y recomendaciones sobre su presentación.

En este estudio, en primer lugar presentaremos el estado de la cuestión siguiendo, entre otros, los trabajos de Lorenzo y Pereira (2000) y Díaz Cintas (2003), y enmarcaremos el subtitulado dentro de la Traducción Audiovisual (TAV), sus características y las fases de subtitulación, tal y como las describe Chaume Varela (2004) y Díaz Cintas y Remael (2007). A continuación, nos centraremos en el subtitulado para sordos y personas con discapacidad auditiva, sus características especiales y aspectos técnicos como la posición de cada elemento en la pantalla y su formato, como indican autores como Orero (2007) y Neves y Lorenzo (2007). Por último, analizaremos los resultados obtenidos del corpus objeto de estudio siguiendo sobre todo las clasificaciones de Diaz Cintas (2007) y que se dividirán en cuatro apartados: especificaciones técnicas de la propuesta de subtitulación y adaptación, retos de la traducción, restos de la adaptación del guión y desafíos técnicos. El objetivo final es presentar una propuesta para un material audiovisual adaptado a las necesidades y expectativas de las personas sordas o con discapacidad auditiva, de esta forma pretendemos eliminar las posibles barreras que existan y aportar nuestro grano de arena a la lucha por la accesibilidad en todos los ámbitos.

References

AENOR (2012). *Subtitulado para personas sordas y personas con discapacidad auditiva.* UNE 153010, Madrid: AENOR. Available at: <http://www.implantecoclear.org/documentos/accesibilidad/UNE_153010_2012.pdf> [Accessed 21 February 2015]

CHAUME VARELA, F. (2004). *Cine y Traducción.* Madrid: Cátedra.

CHAUME VARELA, F. (2004). Film Studies and Translation Studies: Two disciplines at stake in Audiovisual Translation. *Meta: Le Journal des Traducteurs*, 49 (1), pp. 12-24.

CÓMITRE, I. (1997). Algunas consideraciones sobre la traducción del texto audiovisual. In J.M. Santamaría, E. Pajares, V. Olsen, R. Merino & F. Eguiluz, eds., *Trasvases culturales: literatura, cine y traducción.* País Vasco: Departamento de Filología inglesa y alemana de la Universidad del País Vasco. pp. 89-95.

CÓMITRE, I. (2009). Retraducción y traducción audiovisual: estudio descriptivo del doblaje y subtitulado de Peau d'âne de J. Demy. In M.J. Varela Salinas, ed., *Panorama actual del estudio y enseñanza de discursos especializados*. Bern: Peter Lang. pp.165-181.

DÍAZ CINTAS, J. (2001). *La traducción audiovisual: el subtitulado.* Salamanca: Almar.

DÍAZ CINTAS, J. (2003). *Teoría y práctica de la subtitulación. Inglés-español.* Barcelona: Ariel Cine.

DÍAZ CINTAS, J. (2005). Audiovisual translation today: A question of accessibility for all. *Translation Today,* [e-journal] *4*, 3-5.

DÍAZ CINTAS, J. (2007). Por una preparación de calidad en accesibilidad audiovisual. *Trans, Revista de traductología,* 11*,* 56-59*.* Available through Roehampton University, London.

DÍAZ CINTAS, J. (2007). Traducción audiovisual y accesibilidad. In C. Jiménez Hurtado, ed., *Traducción y accesibilidad. Subtitulación para sordos y audiodescripción para ciegos: nuevas modalidades de Traducción Audiovisual* (9-23). Frankfurt: Peter Lang.

DÍAZ CINTAS, J. (2008). La accesibilidad en los medios de comunicación audiovisual a través del subtitulado y la audiodescripción. *Revista Virtual Cervantes.* Roehampton University, Londres.

DÍAZ CINTAS, J. AND REMAEL A. (2007). *Audiovisual translation: subtitling (Translation Practices Explained Series).* Manchester: St. Jerome Publishing.

KARAMITROGLOU, F. (2000). *Towards a Methodology for the Investigation of Norms in Audiovisual Translation*. Amsterdam: Rodopi.

LORENZO GARCÍA, L. AND PEREIRA RODRÍGUEZ A.M. (2000). *Traducción subordinada (II) El subtitulado (inglés-español/galego).* Universidade de Vigo, Vigo: Servicio de Publicacións.

MAYORAL ASENSIO, R. (2001). *Aspectos epistemológicos de la traducción.* Castellon: Universidad Jaume I.

NEVES, J. (2005). *Audiovisual Translation: Subtitling for the Deaf and Hard-of-Hearing.* [Doctoral thesis] University of Surrey-Roehampton, London. Available at: <http://roehampton.openrepository.com/roehampton/bitstream/10142/12580/1/neves%20audiovisual.pdf> [Accessed 23 January 2015].

NEVES, J. (2007). There is research and research: Subtitling for the Deaf and hard of hearing. En C. Jiménez (Ed.), *Traducción y accesibilidad. La subtitulación para sordos y la audiodescripción para ciegos: nuevas modalidades de Traducción Audiovisual.* Frankfurt: Peter Lang.

NEVES, J. AND LORENZO L. (2007). La subtitulación para s/Sordos, panorama global y prenormativo en el marco ibérico. *Trans, Revista de traductología,* 11, 31-44*.*

ORERO, P. (2007). La accesibilidad en los medios: una aproximación multidisciplinar. *Trans, Revista de traductología, 11, 11-14.*

REMAEL A. (2004). A place for film dialogue analysis in subtitling courses. In Orero P., ed. *Topics in Audiovisual Translation.* Amsterdam: John Benjamins. pp. 103-126.

THE TRANSLATION INTO ENGLISH OF BRAZILIAN ANTHROPOLOGICAL SPECIALIZED PHRASEOLOGICAL UNITS: A STUDY OF THE FORMATION OF A TRANSLATIONAL HABITUS BASED ON CORPORA ANALYSIS

Talita Serpa

São Paulo State University

talitasrp82@gmail.com

Diva Cardoso de Camargo

São Paulo State University

divaccamargo@gmail.com

Intending to investigate the social and translational linguistic behaviors of two translators in face of obstacles imposed by cultural barriers in translation, we analyzed a parallel corpus of Social Anthropology of Civilization sub-area, composed by the works, *O processo civilizatório* (1968) and *O povo brasileiro* (1995), written by Darcy Ribeiro, as well as by their translations into English, performed by Meggers and Rabassa, respectively. We also used two comparable corpora of Anthropology in Portuguese and in English, and a support corpus composed mainly of dictionaries of Social Sciences and Anthropology. The main objectives that guided this research were: to observe the translation of specialized phraseological units in Darcy Ribeiro's works; to analyze translators' linguistic and cultural behavior through the analysis of resources used by them in their translations. With these purposes in mind, we

based our study on Camargo's interdisciplinary proposal (2007) by adopting, for the electronic collection and processing of data, the theoretical and methodological framework of Corpus-Based Translation Studies (BAKER, 1995, 1996, 2000), of Corpus Linguistics (Berber Sardinha, 2004, 2010), of Terminology (Barros, 2004; BEVILACQUA, 1999, 2004; CABRÉ, 1998, 1999), and of Lexicology (CORPAS PASTOR, 1996; COLSON, 2004; SABINO, 2011). Concerning the classification and analyzes of data gathered from our corpora, we based our research on the works of Sociology of Translation (SimeonI, 1998; GOUANVIC, 2005), in addition to the concepts of *social capital*, *habitus* and *symbolic exchanges* proposed by Bourdieu (1980). The methodology adopted in our investigation required the use of the program *WordSmith Tools*, which has provided the resources for collection of data and for the observation of cultural and textual aspects. Considering translational behavior, the results obtained from our parallel corpus showed that the translators presented socio-cultural similarities and differences made by different lexical resources: 1) use of literal translations and inversions; and 2) use of omissions. The results also pointed to the nominalization of verbs in the translation of the especialized units, a factor that may allow the Target Culture reader to understand the differences of meanings contained in anthropological terms and expressions, especially in relation to Brazilian universe , such as in: "integrar as massas marginais" →*integration of marginal groups;* "transfigurar as etnias originais" →*transformation of earlier ethnos*; "desenvolver núcleos urbanos" →*development of urban nuclei*; e "colonizar povos" →*colonization of people*. An intense use of "normalization" (BAKER, 1996, 1999) was also observed when the lexical units were related to cultural marked elements, for example: "colher as roças" →*to plant garden plots;* "subjugar caudilhos" →*to subdue local leaders*; "cultivar rocinha" →*to cultivate a small plot.* Recurrence in using these features allowed us to verify the possible formulation of a translational *habitus* that can be associated to Translation Studies. The use of the resources provided by Corpus Linguistics and Lexicology contributed to the theoretical and practical analysis, and allowed the process of awareness of the social role played by translators, through different lexical choices endowed with different social meanings, which represent a trend in works that aim to study the formation of the "Brazilian people."

References

BAKER, M. (1995). Corpora in translation studies: an overview and some suggestions for future research. *Target,* V.7, n. 2, p. 223-243.

_____. (1996). Corpora in translation studies: the challenges that lie ahead. In: SOMERS, H. ed. (1996, *Terminology, LSP and translation studies in language enginnering:* in honour of Juan C. Sager. Amesterdã: John Benjamins, p. 177-186.

_____.(2000). Towards a Methodology for investigation the style of literary translation. *Target*, Amsterdã, V. 12, n. 2, p. 241-266.

BARROS, L.A. (2004). *Curso básico de terminologia*. São Paulo: USP.

BEVILACQUA, C.R. (2004). *Unidades Fraseológicas Especializadas Eventivas – descripción y reglas de formación en el ámbito de la energía solar*. Tese. Barcelona: IULA /Universidade Pompeu Fabra.

_____. (1999). *Unidades fraseológicas especializadas: estado de la cuestión y perspectivas*. Trabalho de pesquisa. Barcelona: IULA / Universidade Pompeu Fabra.

BERBER SARDINHA, T. (2004). *Linguística de corpus.* São Paulo: Manole.

_____. (2010, Como usar a Linguística de Corpus no Ensino de Língua Estrangeira–por uma Linguística de Corpus Educacional brasileira. In: TAGNIN, S.E.O.; VIANA, V. (2010). *Corpora no Ensino de Línguas Estrangeiras***.** São Paulo: HUB Editorial, p. 293-348.

BOURDIEU, P. (1980). *Questions de sociologie*. Paris : Éd. de Minuit.

CABRÉ, M.T. et al. (1998) La Terminología hoy: replanteamiente o diversificación. Terminologia e integração. *Organon*, n.26, v.12, Instituto de Letras da UFRGS, Porto Alegre.

_____. (1999) *La terminología: representación y comunicación; elementos para uma teoría de base comunicativa y otros artículos*. Barcelona: Universitat Pompeu Fabra, IULA.

COLSON, J.P. (2004) Phraseology and computational corpus linguistics: from theory to a practical example. In: BOUILLON, H. ed. (2004, *Langues à niveaux multiples. Hommage au Professeur Jacques Lerot à l'occasion de son éméritat*. Bibliothèque des Cahiers de l'Institut de Linguistique de Louvain, 113. Louvain-la-Neuve, Peeters, p. 35-45.

CORPAS PASTOR, G. (1996). *Manual de fraseología española*. Madrid: Gredos.

GOUANVIC, J. (2005). A Bourdieusian Theory of Translation, or the Coincidence of Practical Instances: Field, ‘Habitus”, Capital and ‘Illusio”. 11 (2), p. 147-166.

RIBEIRO, D. (1968). *O processo Civilizatório.* Rio de Janeiro: Editora Civilização Brasileira S.A.

_____. *The Civilizational Process. (*1968*,* Translated from Portuguese by Betty M. Meggers. Washington: Smithsonian Institution Press.

_____. *O povo brasileiro:* a formação e o sentido do Brasil, 1995, São Paulo: Companhia das Letras.

_____. (2000) *The Brazilian People:* formation and meaning of Brazil*,* Translated from Portuguese by Gregory Rabassa. Gainesville: University Press of Florida**.**

SABINO, M. A. (2011). O campo árido dos fraseologismos. **Revista Signótica**. V. 23, n. 2, p. 385-401.

SIMEONI, D. (1998). The Pivotal Status of the Translator's Habitus. *Target* 10 (1), p. 1-39.

A MIXED APPROACH FOR AUTOMATIC SUB-SENTENTIAL ALIGNMENT OF ENGLISH–ARABIC PARALLEL CORPORA

Abdelghani Yahiaoui
University Lumière Lyon2
abdelghani.y@hotmail.fr

The automatic sub-sentential alignment is very important for the large parallel corpora operation available online for several purposes including machine translation, phraseological studies, multilingual information retrieval and multilingual terminology. We suggest in this paper a mixed approach focusing on the treatment of the Arabic language specificities to improve the quality of automatic sub-sentential alignment of English–Arabic parallel corpora. This alignment quality is even less satisfactory in comparison to the quality of alignment of pairs of languages with Latin writing. This is due, in part, to the different writing system between Arabic and English, but more specifically to some Arabic language specificities, in particular its agglutinative character where we often find words that represent the meaning of syntagmas or whole sentences. This phenomenon makes the word-to-word alignment ineffective, because in this case, an Arabic word corresponds to several English words. To optimize the alignment, we will focus on the resolution of this problem through a mixed approach that makes it possible to obtain an alignment by group of words (one-to-many and many-to-many).

Our approach uses two main approaches of automatic alignment. Firstly, a linguistic approach where we will use a dictionary based on an Arabic

morphological analyzer to obtain the different morphological segmentations of the Arabic words into basic linguistic components (suffix, radical, prefix), then do the alignment basing on translations of these components. The use of an Arabic morphological analyzer allows control of two important characteristics of this language, the agglutination, and the absence of vowels, Furthermore, the use of a dictionary enables the alignment of a large part of corpus with reliable information. With this first approach, we can have one-to-one alignment where the source language is Arabic and the target language is English. The second approach we will use is a statistical approach based on IBM translation models. For this purpose, we will use the GIZA++ tool that implements its models and also the hidden Markov models (HMM). With GIZA++ tool, we can have one-to-many alignments for each language. Therefore, for many-to-many alignment, we will apply the alignment operation in both directions (Arabic-English) and (English-arabic), then we will use heuristics as the intersection to merge the results and get an alignment between groups of words (many-to-many), and this is very important for syntagmas and multiword units alignment.

We will carry out the experimentation of our approach on English–Arabic parallel corpora, which has been selected from the open parallel corpora platform “OPUS”. Furthermore, we will apply our developed process on two corpora already aligned at the sentence level, comprising of a corpus constructed from multilingual documents of the United Nations and a corpus constructed from translated movie subtitles. We will evaluate thereafter the quality of alignment results and how our approach can improve the automatic sub-sentential alignment of English–Arabic parallel corpora.

References

ABDULHAY, A. (2012). Constitution d’une ressource sémantique arabe à partir d’un corpus multilingue aligné. Theses. [online]. Université de Grenoble. Available from: https://tel.archives-ouvertes.fr/tel-00836764 [Accessed 31 Mar 2015].

BROWN, P.F., PIETRA, V.J.D., PIETRA, S.A.D. and MERCER, R.L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*. 19 (2), pp. 263–311.

CHEN, Y. and EISELE, A. (2012). MultiUN v2: UN Documents with Multilingual Alignments. In: *LREC* [online]. pp. 2500–2504. Available from: http://lrec.elra.info/proceedings/lrec2012/pdf/641_Paper.pdf [Accessed 31 Mar 2015].

FRUNZA, O. and INKPEN, D. (2010). Identification and disambiguation of cognates, false friends, and partial cognates using machine learning techniques. *International Journal of Linguistics*. 1 (1), p. E2.

IGNAT, C. (2009). Improving Statistical Alignement And Translation Using Highly Multilingual Corpora. Theses. [online]. Université de Strasbourg. Available from: https://tel.archives-ouvertes.fr/tel-00405733 [Accessed 31 Mar 2015].

KOEHN, P. (2005). Europarl: A parallel corpus for statistical machine translation. In: *MT summit* [online]. pp. 79–86. Available from: http://mt-archive.info/MTS-2005-Koehn.pdf [Accessed 31 Mar 2015].

LARDILLEUX, A. (2009). L'alignement sous-phrastique multilingue pour les nuls. In: *7e Manifestation des Jeunes Chercheurs en Sciences et Technologies de l'Information et de la Communication* [online]. Avignon, France. Available from: https://hal.archives-ouvertes.fr/hal-00439810 [Accessed 30 Mar 2015].

OCH, F.J. and NEY, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*. 29 (1), pp. 19–51.

OCH, F.J. and NEY, H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*. 30 (4), pp. 417–449.

OKITA, T., MALDONADO GUERRA, A., GRAHAM, Y. and WAY, A. (2010). Multi-word expression-sensitive word alignment. In: [online]. Association for Computational Linguistics. Available from: http://doras.dcu.ie/15801/ [Accessed 31 Mar 2015].

OZDOWSKA, S. (2006). ALIBI, un système d'ALIgnement BIlingue à base de règles de propagation syntaxique. [online]. Université de Toulouse II-Le Mirail. Available from: http://w3.erss.univ-tlse2.fr/textes/theses_hdr/ozdowska-these06.pdf [Accessed 30 Mar 2015].

SEGURA, J. and PRINCE, V. (2011). Two Memory-Based Methods for Phrase Alignment. *5th Language and Technology Conference* [online]. Available from: http://hal-lirmm.ccsd.cnrs.fr/lirmm-00838090 [Accessed 31 Mar 2015].

SEMMAR, N. and SAADANE, H. (2014). Etude de l'impact de la translittération de noms propres sur la qualité de l'alignement de mots à partir de corpus parallèles français-arabe. In: *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles* [online]. Marseille, France: Association pour le Traitement Automatique des Langues. pp. 268–279. Available from: http://www.atala.org/taln_archives/TALN/TALN-2014/taln-2014-long-024.

TIEDEMANN, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In: *LREC* [online]. pp. 2214–2218. Available from: http://lrec.elra.info/proceedings/lrec2012/pdf/463_Paper.pdf [Accessed 31 Mar 2015].

TOMEH, N., ALLAUZEN, A., and YVON, F. (2011). Estimation d'un modèle de traduction à partir d'alignements mot-à-mot non-déterministes. In: *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles* [online]. Montpellier, France: Association pour le Traitement Automatique des Langues. Available from: http://www.atala.org/taln_archives/TALN/TALN-2011/taln-2011-long-037.

WISNIEWSKI, A.A.-G. Modèles discriminants pour l'alignement mot à mot. [online]. Available from: http://www.atala.org/IMG/pdf/TAL-2009-3-06-Allauzen [Accessed 30 Mar 2015].

ZIMINA-POIROT, M. (2004). Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles. Theses. [online]. Université de la Sorbonne nouvelle - Paris III. Available from: https://tel.archives-ouvertes.fr/tel-00008311 [Accessed 30 Mar 2015].