



Computational and Corpus-based Phraseology:
Recent Advances and Interdisciplinary Approaches

Proceedings of the Conference

Volume II (short papers, posters and student workshop papers)

November 13-14, 2017
London, UK



EUROPHRAS
EUROPÄISCHE GESELLSCHAFT FÜR PHRASEOLOGIE

Sketch  Engine



tradulex
multilingual communication

ISBN 978-2-9701095-2-5



2017. Editions Tradulex, Geneva

©European Association for Phraseology EUROPHRAS

©University of Wolverhampton (Research Group in Computational Linguistics)

©Association for Computational Linguistics – Bulgaria

This document is downloadable from www.tradulex.com and
<http://rgcl.wlv.ac.uk/europhras2017/>

Preface

As the late and inspiring John Sinclair (1991, 2007) observed, knowledge of vocabulary and grammar is not sufficient for someone to express himself herself idiomatically or naturally in a specific language. One has to have the knowledge and skill to produce effective and naturally-phrased utterances which are often based on phraseological units (the idiom principle). This contrasts with the traditional assumption or open choice principle which lies at the heart of generative approaches to language. As Pawley and Syder (1983) stated more than three decades ago, the traditional approach cannot account for nativelike selection (idiomaticity) or fluency.

Language is indeed phraseological and Phraseology is the discipline which studies phraseological units (PUs) or their related concepts referred to (and regarded largely synonymous) by scholars as multiword units, multiword expressions, fixed expressions, set expressions, phraseological units, formulaic language, phrasemes, idiomatic expressions, idioms, collocations, and/or polylexical expressions. PUs or multiword expressions (MWEs), are ubiquitous and pervasive in language. They are a fundamental linguistic concept which is central to a wide range of Natural Language Processing and Applied Linguistics applications, including, but not limited to, phraseology, terminology, translation, language learning, teaching and assessment, and lexicography. Jackendoff (1977) observes that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words (Jackendoff 1977). Biber et al. (1999) argue that they constitute up to 45% of spoken English and up to 21% of academic prose in English. Sag et al. (2002) comment that they are overwhelmingly present in terminology and 41% of the entries in WordNet 1.7 are reported to be MWEs.

PUs do not play a crucial role only in the computational treatment of natural languages. Terms are often MWEs (and not single words), which makes them highly relevant to terminology. Translation and interpreting are two other fields where phraseology plays an important role, as finding correct translation equivalents of PUs is a pivotal step in the translation process. Given their pervasive nature, PUs are absolutely central to the work carried out by lexicographers, who analyse and describe both single words and PUs. Last but not least, PUs are vital not only for language learning, teaching and assessment, but also for more theoretical linguistic areas such as pragmatics, cognitive linguistics and construction grammars. All the areas listed above are nowadays aided by (and often driven by) corpora, which makes PUs particularly relevant for corpus linguists. Finally, PUs provide an excellent basis for inter- and multidisciplinary studies, fostering fruitful collaborations between researchers across different disciplines, which are, for the time being, unfortunately still largely unexplored.

The e-proceedings feature the short and poster papers presented at the conference "Computational and Corpus-based Phraseology: recent advances and interdisciplinary approaches" (Europhras 2017) as well as the papers from the student seminar which accompanies Europhras 2017. (Regular papers and papers written by the invited speakers are published in a separate Springer LNAI volume). This e-proceedings volume comes with ISBN and DOI numbers assigned to every contribution.

The conference, which is organised jointly by the European Association of Phraseology (Europhras), the Research Institute in Information and Language Processing of the University of Wolverhampton, and the Association for Computational Linguistics – Bulgaria, and sponsored by Europhras, the Sketch Engine, The European Language Resources Association (ELRA) and the University of Wolverhampton, provides the perfect opportunity for researchers to present their work, fostering interaction between (and

joint work by) scholars working in disciplines as diverse as natural language processing, translation, terminology, lexicography, languages learning, teaching and assessment, and cognitive science, to name only a few. In other words, Europhras 2017 provides an excellent basis for interdisciplinary research and for collaboration between researchers across different areas of study related to phraseology, which for the time being is underexplored.

The conference programme is thematically organised into different sessions which demonstrate the breadth of the topics represented at Europhras 2017 and illustrate the application of phraseology in (and its links to) disciplines such as translation, cross-linguistic studies, lexicography, terminography, language learning, theoretical and descriptive linguistics, natural language processing, computational linguistics, corpus linguistics, cognitive studies, cultural studies, specialised languages, technical writing and academic writing.

Every submission to the conference was evaluated by 3 reviewers – i.e. members of the Programme Committee consisting of 46 scholars from 23 different countries, or 12 additional reviewers from 8 countries, who were recommended by the Programme Committee. The conference contributions were authored by a total of 86 scholars from 24 different countries. These figures attest to the truly international dimension of Europhras 2017.

I would like to thank all colleagues who made this truly interdisciplinary and international event possible. In the first place, I would like to acknowledge Kathrin Steyer, the President of Europhras, whose initiative was to organise a Europhras conference in London. I would like to thank our delegates, who have travelled from countries all across the globe to attend this conference, thus providing a living acknowledgement of this special event. I am grateful to all members of the Programme Committee and the additional reviewers for carefully examining all submissions and providing substantial feedback on all papers, helping the authors of accepted papers to improve and polish the final versions of their papers. A special thanks goes to the invited speakers – the keynote speakers of the main conference Ken Church, Gloria Corpas, Dmitrij Dobrovol'skij, Patrick Hanks, Miloš Jakubíček, the invited speakers of the 2 accompanying workshops Carlos Ramish and Jean-Pierre Colson and the tutorial co-speaker Ondřej Matuška. Words of gratitude go to our sponsors – Europhras, the Sketch Engine, ELRA and the University of Wolverhampton including Jan Gilder from the Project Support Office. Many thanks also to João Esteves-Ferreira for publishing the e-proceedings with Tradulex.

Last but not least, I would like to use this paragraph to acknowledge the members of the Organising Committee, who worked very hard during the last 12 months and whose dedication and efforts made the organisation of this event possible. I would like to mention (in alphabetical order) the following colleagues whom I would like to highlight for competently carrying out numerous organisational tasks and being ready to step in and support the organisation of the conference whenever needed. My big 'thank you' goes out to Amanda Bloore, Martina Cotella, Arianna Fabbri, April Harper, Sara Moze, Nikolai Nikolov, Ivelina Nikolova, Rocío Sánchez González, Andrea Silvestre Baquero, Shiva Taslimipoor and Victoria Yaneva.

References

- Biber, D., Finegan, E., Johansson, S., Conrad, S. and Leech, G. 1999. *Longman Grammar of Spoken and Written English*. Longman, Harlow.
- Jackendoff, R. 2007. *Language, consciousness, culture: Essays on mental structure*. The MIT Press.
- Monti, J., Seretan, V., Corpas Pastor, G. and Mitkov R. (forthcoming) "Multiword Units in Machine

Translation and Translation Technology." In Mitkov, R. Monti, J., Corpas Pastor, G. and Seretan V. (Eds.) *Multiword Units in Machine Translation and Translation Technology*. John Benjamins.

Pawley, A. and Syder, F.H. 1983. "Two puzzles for linguistic theory: nativelike selection and nativelike fluency". In Richards J.C. and Schmidt R.W. (Eds.) *Language and communication*. London: Longman.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the third international conference on intelligent text processing and computational linguistics (CICLING 2002)* (pp. 1-15). Mexico City, Mexico.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (2008). Preface. In Granger, S., & Meunier, F. (Eds.), *Phraseology. An interdisciplinary perspective*. Amsterdam: John Benjamins publishers.

Ruslan Mitkov, Conference Chair
 London 13.11.2017

Organisers:

Europhras 2017 is jointly organised by the European Association for Phraseology EUROPHRAS, the University of Wolverhampton (Research Institute of Information and Language Processing) and the Association for Computational Linguistics – Bulgaria.

Conference Chair:

Ruslan Mitkov, University of Wolverhampton, UK

Programme Committee:

Julio Bernal, Caro and Cuervo Institute, Colombia
Douglas Biber, Northern Arizona University, USA
Nicoletta Calzolari, Institute for Computational Linguistics, Italy
María Luisa Carrió-Pastor, Polytechnic University of Valencia, Spain
Sheila Castilho, Dublin City University, Ireland
Kenneth Church, IBM Research, USA
Jean-Pierre Colson, Université catholique de Louvain, Belgium
Gloria Corpas, University of Malaga, Spain
František Čermák, Charles University in Prague, Czech Republic
Anna Čermáková, Charles University, Czech Republic
Dimitrij Dobrovolskij, Russian Academy of Sciences, Russian Language Institute, Russia
Jesse Egbert, Northern Arizona University, USA
Thierry Fontenelle, Translation Centre for the Bodies of the European Union, Luxembourg
Kleanthes K. Grohmann, University of Cyprus, Cyprus
Patrick Hanks, University of Wolverhampton, UK
Ulrich Heid, University of Hildesheim, Germany
Miloš Jakubíček, Lexical Computing and Masaryk University, Czech Republic
Kyo Kageura, University of Tokyo, Japan
Valia Kordoni, Humboldt University of Berlin, Germany
Simon Krek, University of Ljubljana, Slovenia
Pedro Mogorrón Huerta, University of Alicante, Spain
Johanna Monti, Naples Eastern University, Italy
Sara Moze, University of Wolverhampton, UK
Preslav Nakov, Qatar Computing Research Institute, HBKU, Qatar
Michael Oakes, University of Wolverhampton, UK
Marija Omazić, University of Osijek, Croatia
Petya Osenova, Sofia University, Bulgaria
Magali Paquot, Université catholique de Louvain, Belgium
Giovanni Parodi Sweis, Pontifical Catholic University of Valparaíso, Chile
Alain Polguère, University of Lorraine, France
Carlos Ramisch, Marseille Laboratory of Fundamental Computer Science, France
Ute Römer, Georgia State University, USA
Agata Savary, François Rabelais University, France

Barbara Schlücker, The University of Bonn, Germany
Violeta Seretan, University of Geneva, Switzerland
Kathrin Steyer, Institute of German Language, Germany
Yukio Tono, Tokyo University of Foreign Studies, Japan
Cornelia Tschichold, Swansea University, UK
Benjamin Tsou, City University of Hong Kong, China
Agnès Tutin, University of Grenoble, France
Aline Villavicencio, Federal University of Rio Grande do Sul, Brazil
Eveline Wandl-Vogt, Austrian Academy of Sciences, Austria
Tom Wasow, Stanford University, USA
Eric Wehrli, University of Geneva, Switzerland
Stefanie Wulff, University of Florida, USA
Michael Zock, Marseille Laboratory of Fundamental Computer Science, France

Additional Reviewers:

Verginica Barbu Mititelu, Romanian Academy, Research Institute for AI, Romania
Archana Bhatia, Language Technologies Institute, CMU, USA
Ismail El Maarouf, Adarga Limited, Oxford University Press, UK
Voula Giouli, Institute for Language and Speech Processing, “Athena” RIC, Greece
Václava Kettnerová, Charles University, Czech Republic
Rogelio Nazar, Pontifical Catholic University of Valparaíso, Chile
Irene Renau, Pontifical Catholic University of Valparaíso, Chile
Ioannis Saridakis, University of Athens, Greece
Inguna Skadina, University of Latvia, Latvia
Shiva Taslimipoor, University of Wolverhampton, UK
Veronika Vincze, Hungarian Academy of Sciences, Hungary
Victoria Yaneva, University of Wolverhampton, UK

Invited Speakers:

Kenneth Church, Johns Hopkins University, USA
Gloria Corpas, University of Malaga, Spain
Dmitrij Dobrovolskij, Russian Academy of Sciences, Russian Language Institute, Russia
Patrick Hanks, University of Wolverhampton, UK
Miloš Jakubíček, Lexical Computing and Masaryk University, Czech Republic

Invited Speakers of Europhras 2017 workshops:

Jean-Pierre Colson, Université catholique de Louvain, Belgium
Carlos Ramisch, Marseille Laboratory of Fundamental Computer Science, France

Invited Speakers of SketchEngine Tutorial:

Miloš Jakubíček, Lexical Computing and Masaryk University, Czech Republic
Ondřej Matuška, Lexical Computing Ltd, Czech Republic

Organising Committee:

Amanda Bloore, University of Wolverhampton, UK
Martina Cotella, University of Wolverhampton, UK
Arianna Fabbri, University of Genoa, Italy
April Harper, University of Wolverhampton, UK
Sara Moze, University of Wolverhampton, UK
Rocío Sánchez González, University of Malaga, Spain
Andrea Silvestre Baquero, Polytechnic University of Valencia, Spain
Shiva Taslimipoor, University of Wolverhampton, UK
Victoria Yaneva, University of Wolverhampton, UK

Organisers of the Student Research Workshop:

Victoria Yaneva, University of Wolverhampton, UK
Shiva Taslimipoor, University of Wolverhampton, UK

Programme Committee of the Student Research Workshop:

Sara Moze, University of Wolverhampton, UK
Michael Oakes, University of Wolverhampton, UK
Petya Osenova, Sofia University, Bulgaria
Irene Renau, Pontifical Catholic University of Valparaíso, Chile
Irina Temnikova, Qatar Computing Research Institute, Qatar
Aline Villavicencio, Federal University of Rio Grande do Sul, Brazil

Table of Contents

Short Papers

<i>A Comparison of Three Metrics for Detecting Cross-Linguistic Variations in Information Volume and Multiword Expressions between Parallel Bitexts</i>	
Éric Poirier	1
<i>Hybrid Methods for the Extraction and Comparison of Multilingual Collocations in Languages for Specific Purposes</i>	
Guadalupe Ruiz Yepes	11
<i>Phraseological Units in Horror Comics: Comparative Study of the Translation into English, French and Spanish from a Multimodal Corpus</i>	
María del Carmen Baena Lupiáñez	19
<i>Exploring Automated Essay Scoring for Nonnative English Speakers</i>	
Amber Nigam	28
<i>Automatic Annotation of Verbal Collocations in Modern Greek</i>	
Vasiliki Foufi, Luka Nerima and Eric Wehrli	36
<i>Corpus Linguistic Exploration of Modern Proverb Use and Proverb Patterns</i>	
Kathrin Steyer	45
<i>Life Values Reflection in Idioms: Corpus Approach</i>	
Seda Yusupova	53
<i>Metaphors of Economy and Economy of Metaphors</i>	
Antonio Pamies-Bertrán and Ismael Ramos Ruiz	60
<i>A Note on Controlled Compositions in Japanese EFL Classes for Intermediate Learners: With a Focus on Reordering Questions</i>	
Shimpei Hashio and Nobuyuki Yamauchi	70
<i>How Does Data Driven Learning Affect the Production of Multi-Word Sequences in EAP Students' Academic Writing?</i>	
Melissa Larsen-Walker	78
<i>Phraseology in Teaching and Learning Spanish as a Foreign Language in the USA.</i>	
Victoria Llongo	87
<i>Extracting Formulaic Expressions and Grammar and Edit Patterns to Assist Academic Writing</i>	
Jhih-Jie Chen, Jim Chang, Mei Hua Chen, Jason Chang and Ching-Yu Yang	95
<i>Improving Requirement Boilerplates Using Sequential Pattern Mining</i>	
Maxime Warnier and Anne Condamines	104

Intonational PERiods (IPE) and Formulaic Language: A Genre-based Analysis of a French Speech Database

Maria Zimina and Nicolas Ballier 113

Poster Papers

Google N-grams Viewer and Food Idioms

Sarah Virginia Carvalho Ribeiro and Paula Lenz Costa Lima 122

The Effects of Learner Variables on Phraseological Proficiency

Kathrin Kircili 127

Synonymy Between Theory and Practice: The Corpus-Based Approach to Determining Synonymy in Lexicographic Descriptions in Croatian

Goranka Blagus Bartolec 132

A Lexical Database for the Analysis of Portuguese MWES

Sandra Antunes 137

A Contrastive Analysis of Antonymous Prepositional Pairs in Croatian and Russian

Ivana Matas Ivanković 143

An Objective Method of Identifying Teachworthy Multi-word Units for Second Language Learners

James Rogers 148

English Multi-word Expressions as False Friends between German and Russian: Corpus-driven Analyses of Phraseological Units

Iyubov Nefedova 154

Towards the Generation of Bilingual Chinese-English Multi-word Expressions from Large Scale Parallel Corpora: An Experimental Approach

Benjamin K. Tsou, Derek F. Wong and Ka Po Chow 162

Phraseological Meaning and Image

Roza Ayupova 169

Student Research Workshop

Language and Power in Czech Corpora

Irene Elmerot 174

Teasing Apart Russian Idioms And Homonymic Compositional Expressions

Marina Pchelina and Jae-Woong Choe 178

Observations on Phonetic and Metrical Patterns in Spanish-language Proverbs

Jordi Martínez, Gemma Bel Enguix and Liliana Torres Flores 182

<i>Towards a Corpus-lexicographical Discourse Analysis</i>	
Emma Franklin	190
<i>Digital Storytelling and the 21st Century Classroom: a powerful tool in phraseological units learning</i>	
Annalisa Raffone	197

A Comparison of Three Metrics for Detecting Cross-Linguistic Variations in Information Volume and Multiword Expressions Between Parallel Bitexts¹

Éric Poirier¹

¹ Université du Québec à Trois-Rivières, Québec, Canada
eric.poirier@uqtr.ca

Abstract. This paper presents the results of a comparison of three metrics for measuring cross-linguistic variations in information volume between parallel segments of a bilingual corpus. The performance of each metric is compared with the results of a human annotation of multiword expressions (MWEs) in each segment. The first metric measures characters in source and target segments and compares the variation, if any, with the expected character count ratio based on averages for the entire source and target texts. The second metric follows the same method except that it measures graphical word count (function and content words combined) in target and source segments. The third metric involves an analysis obtained via the content word precision (CWP) algorithm coded in Python. The purpose of the comparison is to determine which metric is closer to the human annotation and is, therefore, a better indicator of a large spectrum of MWEs.

Keywords: cross-linguistic phraseology, detection of multiword expressions, information volume variation, content word precision algorithm.

1 Introduction

As a contribution to computational studies of multiword expressions (MWEs), this paper presents the results of a comparison, albeit small-scale, of three metrics for detecting MWEs in parallel segments of a bilingual corpus. The untested underlying hypothesis of the comparison is that information volume variation in parallel segments (as established by content word imbalance) correlates with the presence in source or target of a large spectrum of MWEs. Hence, MWEs should occur in parallel segments where there is a cross-linguistic information volume difference, as determined by differences in the number of content words in source and target.

The study of MWEs in parallel segments is of the utmost importance to cross-linguistic phraseological studies as described in Colson (2008), and hence to transla-

¹ I wish to thank the reviewers for their comments and suggestions. I also wish to thank my colleague Paul John from UQTR for his input in reading the final version of the paper. Of course, any remaining omissions or errors are mine.

tion studies and phraseological studies. Determining which metric is most accurate in detecting MWEs will also contribute to recent works in corpus-based phraseology such as Granger and Paquot (2008) and will allow the mining of a large spectrum of MWEs (such as clusters, lexical bundles, n-grams, recurrent sequences) or even new classes of MWEs. Just like any corpus-based approach, our new bilingual approach is designed to complement other traditional monolingual approaches such as appear in volumes 1 and 2 of the *Oxford Dictionary of Current Idiomatic English* (Cowie & Mackin, 1975; Cowie, Mackin & McCaig, 1983).

As a preliminary phase of a larger project, this paper presents the results of a small-scale analysis of the first 25 segments of the bilingual text selected. The text chosen for the evaluation of metric performance is the parallel English-French Inaugural Address of J.F. Kennedy (January 20, 1961). The French version used in our corpus is the official French translation provided on the multilingual pages of the website. The source text and its translation were automatically aligned in source and target segments with Logiterm, a proprietary software creation tool of bitexts using HTML output format.

2 Definitions

The notion ‘volume of information’ is defined as the number of content words in each parallel language segment. This definition makes it possible to determine a ratio of information volume in source and target segments as measured by the number of content words they each contain. For purposes of comparing the three metrics with regard to detection of MWEs, the information precision of the translation (IPT) is the ratio of target segment information volume to source segment information volume, as shown in figure 1.

Figure 1. Formula for calculating IPT

$$\text{IPT} = \frac{\text{Number of content words in source segment}}{\text{Number of content words in target segment}}$$

This ratio is an enhancement² of the BLEU metric with modified n-gram precision, as proposed by Papineni et al. (2002), and it is used for the assessment of machine translation outputs in comparison with a human translation. IPT measures the variation of information volume in translation.³ Bilingual segments having the same in-

² Instead of taking into account all words of a segment as with BLEU, the IPT ratio focuses exclusively on content words. This approach is more appropriate to translation, which is a meaning-based process. That is, it makes sense in the assessment of translation quality to favor content words over function or grammatical words, which contribute less to the information conveyed in a segment.

³ Interestingly, information volume is associated with adequacy, which is one of two qualities of a good translation based on human judgment, the other being fluency, as defined by

formation volume (as measured by an IPT of 1) are isomorphic, while bilingual segments having a different IPT score (either lower or higher than 1) are anisomorphic.

Stylistic effects excluded, information volume has also been used to define basic translation errors, such as addition or omission (Delisle, Lee-Jahnke & Cormier, 1999). Most of the time, a change in information volume (i.e., addition or omission of meaning) implies the addition or omission of a content word, but this is not always the case, particularly in segments containing phraseological units or MWEs. That is, many occurrences of parallel segments unequal to 1 cannot be attributed solely to these two basic translation errors. Other language or textual constraints occurring in the translation process cause variations in IPT. Provided the translation is correct, an information imbalance is an indication of linguistic or textual constraints which are often, but not necessarily, caused by a large spectrum of idiomatic and phraseological units and MWEs. Our untested hypothesis regarding a correlation between information volume imbalance and the presence of MWEs in parallel segments is based on this assumption.

According to our hypothesis, MWEs may be detected in segments that are anisomorphic (having an IPT score other than 1). As Franco Aixelà (2015) suggests, anisomorphisms have multiple origins and are not limited to linguistic structures, such as interpretative, pragmatic and cultural anisomorphisms. In a bilingual parallel corpus, the divergence of a segment’s IPT score from 1 derives not only from linguistic anisomorphisms, but also from various textual anisomorphisms. The latter include idiosyncratic constraints in the translation of MWE in target text, which the metrics may contribute to detecting. MWEs detected in anisomorphic segments may include for example expressions such as {globe; monde entier} and {(the) free; États libres}.

3 Methods

To identify the most effective method of detection of MWEs in parallel segments, the performances of character count metric, of word count metric and of the content word precision algorithm (Poirier, 2014) are compared to a human annotation of the same bilingual segments in a parallel corpus. Inter-annotator agreement on phraseological units in the extract of the parallel bitext was not checked but will be in a subsequent extended research project.

Described below, the human annotation of the bilingual segments served as the gold standard measure of information volume similarity or discrepancy to which the three other metrics were compared.

Snover, Madnani, Dorr and Schwartz (2009). Since adequacy measures whether the translation conveys the correct meaning, measuring adequacy requires the decomposition of sentence meaning into smaller parts (phrase and word level meaning). Measuring correct meaning in a translation therefore involves alignment of information content at the phrase and word level between parallel segments.

3.1 Human annotation protocol

Content word pairs in source and target segments were aligned manually with the meaning-based heuristic described in Poirier (2016). The heuristic requires that each content word in the source segment be manually aligned to at least one content word or to a null token (\emptyset) in the target segment and reciprocally that each content word in the target segment be manually aligned to at least one content word or to a null token in the source segment. Therefore, anisomorphic segments may contain the null token associated to either 1, 2 or more content words in the opposite segment. These alignments have been coded as 1-to- \emptyset , \emptyset -to-1, 2-to- \emptyset , \emptyset -to-2, and so on.

The first segment of our corpus was the following salutation at the beginning of the address: [English source] *Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice president Nixon, President Truman, Reverend Clergy, fellow citizens*; [French target] *Monsieur le Vice-président Johnson, Monsieur le Président, Monsieur le Président de la Cour suprême, Monsieur le Président Eisenhower, Monsieur le Vice-président Nixon, Monsieur le Président Truman, révérend clergé, chers concitoyens*.

The bilingual segment contains 13 1-to-1 aligned pairs (a) such as {Johnson; Johnson}, {Mr.; Monsieur}, and {Speaker; Président}. There are no 1-to- \emptyset anisomorphic aligned pairs (b), although there are four occurrences of the same \emptyset -to-1 aligned pair (c) { \emptyset ; Monsieur}. For these alignments, there is no counterpart in English for the formal use of *Monsieur* in French, which is a recurrent element in French sequences of a phraseological nature. The bilingual segment contains three many-to-many anisomorphic aligned pairs (d): a 2-to-3 pair as in {Chief Justice; Président Cour suprême}⁴ and two occurrences of the same 2-to-1 aligned pair: {Vice President; Vice-président}⁵.

The manual count of content word balance between source and target in the first segment can be schematized as $a+b+c+d$, that is, $13+0+0+6 = 19$ in English and $13+0+4+5 = 22$ in French. This count shows that the segment is anisomorphic in terms of information volume. Assuming the proposed correlation of anisomorphic segment to MWE is correct, this segment should contain an MWE, which can be confirmed with the aligned pairs above, since these contain some MWEs in a broad sense. In comparing the performance of the three metrics, success constitutes the detection of an anisomorphic segment in which an MWE has been detected with the human annotation.

⁴ The aligned pair is cited without its function words because the items aligned are content words.

⁵ This anisomorphism is simply due to a morphological spelling difference which might be perceived as arbitrary on a theoretical/linguistic level, but which still counts, on a practical level in translation, as an incorrect usage. Although minor, this error is typical of misuse of MWE that translators are expected and required to recognize and use properly.

3.2 Character count metric

This metric compares the number of characters in source and target segments. Since character uses are generally not correlated with meaning (as opposed to information volume or content word counts), a global ratio of characters in source to target text is first compiled to normalize, if any, the expected discrepancy of character counts by segment. The value of the ratio could vary from one text and one translator’s style to another. In the corpus, the character count (in Word 2013, but any other software or version of Word applied consistently to all the segments could be used) of the whole text in source and target languages shows a ratio of 1.23 between English source text and its French translation. That is, there are 6199 characters (without spaces) in the source text and 7664 characters (without spaces) in the target text (with the calculation being very similar when spaces are taken into account). The ratio used to qualify segments as isomorphic was based on a character count difference of 1.23 between English and French segments. To allow for a standard deviation from this ratio, arbitrary but adjustable isomorphic minimum and maximum thresholds were calculated as a 10% difference from the expected character count ratio of 1.23.⁶ Minimum thresholds were rounded down to the nearest whole number, and maximum thresholds were rounded up. For the first segment of our corpus (as cited in the previous section), these thresholds are described in table 3 below.

Table 3. Character count metric results for the first segment

Characters in source	Characters in target	Target expected count (1.23 ratio)	Isomorphic threshold minimum	Isomorphic threshold maximum	Character count metric prediction
151	233	185.73	167	204	anisomorphism

For the first segment, the comparison of the actual character counts in the target segment with the expected character count, as calculated with the normalized ratio for the entire text, predicts that the target segment is anisomorphic with respect to the source segment. In other words, the actual character count in target segment is much higher (233) than the expected character count (204), as determined by the normalized ratio.

⁶ There is empirical evidence in professional translation that French translations are approximately 15% longer than their English source text in terms of words, but we have not found scientific works confirming this principle in terms of character length. A recent study on POS-Editing (Béchara, Ma & van Genabith, 2011) found that in a set of 52,383-word English-French segment pairs, the average segment length was 13 words in English and 15 words in French, which confirms the 15% variation in words.

3.3 Word count metric

This metric compares the number of all content and function words present in source and target segments. For this metric, words are defined as any chain of characters in a segment that are separated by a space. As with the character count metric, a ratio of words in source to target entire text was first calculated to normalize the expected discrepancy of word counts between source and target segments. A ratio of 1.09 was calculated with 1363 words (in Word 2013) for the entire source text and 1485 words (in Word 2013) for the whole target text. To allow for a standard deviation from this ratio, arbitrary but adjustable isomorphic minimum and maximum ratio thresholds were calculated as a 5% difference from the expected word count ratio of 1.09. The deviation interval of 5% is an approximation of tolerance, taking into account the fact that word counts are partly based on function words (not correlated with lexical meaning and information content) and partly based on content words (correlated to meaning and information content). Depending on the type of texts and translations, this deviation interval could be modified as needed. The 5% standard deviation defines a minimum tolerance of $1.09 - 0.0545 = 1.0355$ word count ratio and a maximum tolerance of $1.09 + 0.0545 = 1.1445$ word count ratio for isomorphic segments. As for the character count metric, thresholds were rounded up and down to the nearest whole number. For the first segment of our corpus, these thresholds are described in table 4 below.

Table 4. Word count result for the first segment

Words in source	Words in target	Target expected count (1.09 ratio)	Isomorphic threshold minimum	Isomorphic threshold maximum	Word count metric prediction
19	30	20.71	19	23	anisomorphism

The comparison of the actual word count in target segment with the expected word count predicts that the first target segment is anisomorphic compared with the source segment. In other words, the actual word count in the target segment is much higher (30) than the maximum word count threshold (23), as compiled with the normalized ratio for the whole text.

3.4 CWP algorithm

The CWP algorithm is coded in Python, as described in detail in Poirier (2014). This method is an automatized version of the human annotation method and is based on the deletion in parallel segments of function words, which, in English and French (as in most languages), form a closed set of items⁷. A comparison is made between the

⁷ The algorithm works with any language which uses distinct and separate function and content word sets. It is thus operational for English, French and Spanish corpora. For other lan-

number of content words in the source and target segments. As for the manual annotation, no specific threshold was used as an expected difference in content word counts for the target segments. Tolerance margins are not required for short parallel segments containing few content words, but might be useful when segments are longer. There is no known standard as to when to use maximum and minimum thresholds to consider that source and target segments are isomorphic in terms of content words. This uncertainty, and the added complexity needed for the algorithm to take into account segment length in the comparison of content words, made us defer the implementation of these options to a future version of the algorithm.

For the first segment of our corpus, the content word counts in source and target segments, as determined by the CWP algorithm, are described in table 5 below.

Table 5. CWP result applied to the first segment

Content words in source	Content words in target	CWP metric prediction
19	22	anisomorphism

The comparison of the content word count in target segment with the content word count in source segment gives the same result as the human annotation (19, 22 and an anisomorphic parallel segment) because both used a method based on the distinction between function and content words. However, the performance of the algorithm over the 25 segments is not identical to the human annotation, with more details on these differences provided below in the next section.

4 Results

Among the first 25 segments manually annotated, only five were of isomorphic nature, according to the human annotation. This suggests that volume information imbalance is very common in translation and, correlatively, if our untested assumption is correct, that the text analyzed contains numerous occurrences of MWEs (19 segments out of 25 would include one or more MWEs).

The table below shows the success rate of the three methods in the detection of anisomorphic segments and potential MWEs.

Table 6. Success rates compared with the human annotation (25 segments)

<i>Character</i>	<i>Word</i>	<i>CWP</i>
------------------	-------------	------------

languages, the distinction between function and content words has not been tested for the application of the CWP algorithm. Most typological classifications of languages (such as isolating, agglutinative and inflecting) are not concerned with the possibility or even the feasibility of separating function and content words. In the current initial state of development of the algorithm, the concepts of function and content words seem to be universal, and the limitations of these concepts presently seem to be more related to the use of similar or equivalent words classified differently (see discussion section) than to the existence of languages not having function and content words.

	<i>count</i>	<i>count</i>	
<i>Correct detection</i>	13/25	15/25	22/25
<i>Detection rate (%)</i>	52%	60%	88%

As we can see from Table 6, the CWP algorithm, at 88%, has the highest success rate in the detection of anisomorphic segments. Although interesting, the CWP results are not always identical to the human annotation. In three occurrences, the CWP algorithm did not properly detect an isomorphic or an anisomorphic parallel segment. One occurrence is an anisomorphic segment analyzed as isomorphic (false negative), and in the two other occurrences, an isomorphic segment is wrongly analyzed as anisomorphic (false positive). Two of these failures are due to a difference between French and English in word categorization and to a function word such as “all”, used as quantifier (function word), being translated by “toutes”, used as an indefinite adjective (content word). The third wrong detection is due to the use of “dare” as an auxiliary (specifically, modal) verb in English. The algorithm processed it as a full lexical verb, while the human annotation aligned it with a null token (not translated) in French, as in segment 25 [*We dare not tempt them with weakness* = *Nous ne les tenterons pas par notre faiblesse*].

5 Discussion

Two of the failures of the CWP algorithm may be due to arbitrary uses and categorizations of function and content words in English and French. Resolving this issue seems impossible, since each language classifies semantically equivalent words in different categories, so it is probably best viewed as a demonstration of language idiomaticity in grammatical uses and analysis of specific words. Adopting this view just pushes the limits of MWEs a bit further towards the recognition of irreducible, incompatible, and hence phraseological, uses of lexical items between languages, even if on their meaning side, these expressions are transparent and easily translatable.

The third failure of the CWP algorithm seems to be different in that it relies on a correct analysis of a particular verb acting as a function word (modal auxiliary) in some contexts and as a content word (main verb) in other contexts⁸. In this case, it seems that adding an automatic POS tagging step in the CWP algorithm may help to resolve this specific issue. However, automatic POS tagging is not 100% accurate, and the figures are even lower when compiled on a segment level.⁹

⁸ As in *I dared not move* [modal] vs *I dared him to move* [main verb].

⁹ Giménez and Marquez (2004) report an accuracy of close to 97% on a word count basis. Converting this level of accuracy at the segment level makes it less impressive since it can be reasonably argued that most segments (and sentences) are often comprised of 10 words or more. For example, for ten segments of 10 words, the 97% accuracy implies that as much as three segments out of ten (30 % of segments) could contain a POS tagging inaccuracy pro-

Although the CWP algorithm permits detection of a large spectrum of MWEs, a significant limitation of the automatic comparison of information volume is that literally translated MWEs (such as “tomber comme des mouches” = “to drop like flies”, “donner le feu vert” = “give the green light”) cannot be detected by a discrepancy in information volume, even though the examples given are both well-known MWEs from a monolingual and traditional approach in phraseological studies. However, this limitation of the CWP algorithm may well be viewed positively as a supplemental approach to traditional monolingual mining of MWEs. Another limitation of the CWP algorithm is that MWEs entirely constituted of function words and having no content words, such as “this much”¹⁰, cannot, by definition, be detected with the algorithm.

6 Conclusion and further studies

Comparison of the three metrics in detecting information volume variation shows that the results of the CWP algorithm are closer to the manual annotation of the same parallel segments. The CWP algorithm can thus help to detect a large spectrum of corpus-based MWEs, albeit with the limitations detailed in the discussion section. Analysis of failures has shown that implementing a POS tagging module to enhance the content word analysis with the CWP could marginally increase its success rate on a segment-level basis.

The implementation of POS tagging and a contextual analysis of ambiguous words are left to a future study on a larger corpus. In such an experimental design it would be advisable to use human annotation only for segments in which MWEs have been detected to confirm or invalidate the presence of MWEs after the application of the CWP algorithm. Inter-rater agreement could also be measured at the same time. Further studies are also required in the assessment of two types of anisomorphic segments in relation with MWEs: those in which source segments contain more content words than do target segments and those in which target segments contain more content words than do source segments.

Automatic or semi-automatic detection of MWEs in parallel segments constitutes a significant contribution to translation studies for teaching and translation quality assessment purposes, as well as to cross-linguistic phraseology as part of bilingual lexicography and usage-based linguistics. Extending MWE categories to textual constraints could help translation studies specialists distinguish between free translation shifts due to circumstantial and creative reasons and language-constraint translation shifts due to idiomatic textual uses in languages. The possibility of separating and studying both categories of translation shifts represents an advance in translation studies.

vided the three words inaccurately tagged out of 100 are distributed in three different segments.

¹⁰ This MWE was found in our corpus and detected as such with the manual annotation (but not the CWP). According to the Oxford English dictionary (online), it means “The fact about to be stated”.

The detection of MWEs in parallel segments also contributes to monolingual and traditional studies in phraseology and linguistics by extending the limits of language idiomaticity and uncovering new items in established categories of MWE or suggesting new classes of MWE.

References

1. Béchara, H., Ma, Y., and van Genabith, J. (2011, September). Statistical post-editing for a statistical MT system. In *MT Summit* (Vol. 13, pp. 308-315).
2. Colson, J.-P.: Cross-linguistic phraseological studies: an overview. In: Granger, S, Meunier, F. (eds) *Phraseology: An interdisciplinary perspective*, pp. 191–206. John Benjamins Publishing, Amsterdam/Philadelphia (2008).
3. Cowie, A. P. and Mackin, R. (1975). *Oxford Dictionary of Current Idiomatic English. Volume 1: English Idioms*. Oxford Univ. Press.
4. Cowie, A. P., Mackin, R., and McCaig, I. R. (1983). *Oxford Dictionary of Current Idiomatic English. Volume 2: English Idioms*. Oxford Univ. Press.
5. Delisle, J., Lee-Jahnke, H. and M. C. Cormier (eds.) (1999). *Terminologie de la traduction Translation Terminology Terminología de la traducción Terminologie des Übersetzung*. Coll. FIT, volume 1, Amsterdam / Philadelphia: John Benjamins
6. Franco Aixelà, J. (2015). “Anisomorfismos y traducción” in *Enciclopedia Abierta de Estudios de Traducción e Interpretación*, AIETI. <online: <http://www.aieti.eu/Enciclopedia/ANI-iconoses/index.html>>. Retrieved August 24th, 2016.
7. Giménez, J., and Marquez, L. (2004). Fast and accurate part-of-speech tagging: The SVM approach revisited. *Recent Advances in Natural Language Processing III*, 153-162.
8. Granger, S., Paquot, M.: Disentangling the phraseological web. In: Granger, S, Meunier, F. (eds) *Phraseology: An interdisciplinary perspective*, pp. 27–50. John Benjamins Publishing, Amsterdam/Philadelphia (2008).
9. Inaugural Address of President John F. Kennedy, Washington, D.C., January 20, 1961, <online: <http://www.jfklibrary.org/Research/Research-Aids/ReadyReference/JFK-Quotations/Inaugural-Address.aspx>> with official French translation Discours d'investiture du Président John Fitzgerald Kennedy provided <online: <http://www.jfklibrary.org/JFK/HistoricSpeeches/Multilingual-Inaugural-Address/MultilingualInaugural-Address-in-French.aspx>>. Retrieved July 2nd, 2016.
10. Papineni, K., Roukos, S., Ward, T., Zhu W.J. (2002). BLEU: a method for automatic evaluation of machine translation, in: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 311–318.
11. Poirier, E.: A method for automatic detection and manual localization of content-based translation errors and shifts. In *Journal of Innovation in Digital Ecosystems*, vol. 1, issue 1-2, pp. 38-46. Elsevier (2014).
12. Poirier, E.: Meaning-based content word alignment heuristic. In: Chbeir, R., Agrawal, R., Biskri, I.: In *MEDES Proceedings of the 8th International Conference on Management of Digital EcoSystems*, pp. 208-214, ACM, New York, (2016).
13. Snover, M., Madnani, N., Dorr, B. J., and Schwartz, R. (2009, March). Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 259-268). Association for Computational Linguistics.

Hybrid Methods for the Extraction and Comparison of Multilingual Collocations in the Language for Specific Purposes of Marketing

Guadalupe Ruiz Yepes

Heilbronn University, Am Europaplatz 11, 74076 Heilbronn, Germany
Guadalupe.ruiz-yepes@hs-heilbronn.de

Abstract. This paper presents a *status quo* of an ongoing research with the main focus on the cross linguistic comparison of collocations. The main aim is to compare the collocations extracted from a multilingual comparable corpus – German, English, Spanish – in the language for specific purposes of marketing using hybrid methods. To achieve this objective the collocations are defined and clearly distinguished from other phrases, before describing the methods which have so far been used to extract collocations. Special attention will be paid to hybrid methods that combine linguistic and statistical information. After comparing these methods, one of them will be applied to the corpora. The findings will be analyzed and conclusions drawn in order to list differences in the structure of German, English and Spanish collocations in the field of marketing. The results and findings of this piece of research can be applied to improve the performance in the field of Second Language Acquisition, Translation and Translator training.

Keywords: collocations, languages for special purposes, marketing, corpus linguistics, phraseology of marketing, translation and translator training.

1 Introduction

According to Fluck (1996: 11), there has been much discussion in Linguistics on the exact nature of languages for specific purposes (LSP) and a widely-accepted definition does not exist. However, most linguists in this field share the opinion that LSP is a variety of common language, has developed from common language and uses the grammatical means of common language (Arntz/Picht/Schmitz 2014). The formation of collocations in LSP follows therefore similar rules as in common language. The formation of collocations is very often domain specific, since words which do not participate in a collocation in everyday language often do form part of a collocation in a LSP, i.e. the noun “file” collocates with the verbs “create”, “delete”, “save” in texts about computers, but not necessarily in other contexts (McKeown and Radev, 2000: 510). This conclusion opens the doors of research to the field of LSP phraseology which is still in its infancy compared to common language phraseology.

The main aim of this paper is to compare the collocations extracted from a multilingual comparable corpus, but also to compare different methods for the extraction of

collocations. The corpus we are working with is composed of marketing texts. It is a comparable corpus composed by three corpora in three languages (German, Spanish and English) which are comparable because they belong to the same text type; all three corpora consist of articles on marketing topics, published in scientific journals. Firstly collocations will be defined in order to distinguish them from other word combinations such as idioms or free word combinations. The next step will be to describe existing methods for the extraction of collocations. Finally one of these methods will be applied on the corpora, findings will be analyzed and the consequences of these findings for translation and foreign language training will be highlighted.

2 What is a “collocation”?

There are numerous definitions for the phenomenon “collocation”. Originally the term “collocation” was used in a very broad sense to describe the “general event of recurrent word co-occurrence” (Seretan, 2011: 13). But this statistical view has been replaced later by a linguistically-motivated one, in which the items in a collocation are syntactically and semantically related. In recent studies it has been suggested to use the term “co-occurrence” for the recurrent co-appearance of two words, while the term “collocation” is reserved for the phraseological (linguistic) approach. This distinction between “co-occurrences” and “collocations” seems to be accepted and will be adopted in this paper.

Since the term collocation was introduced by Firth in 1957, researchers from all kind of disciplines have dealt with it. Consequently, there are definitions of “collocation” in phraseology, computational linguistics, corpus linguistics, etc. Each discipline tries to define “collocation” to meet its needs. As a consequence, the term is very vague and not clearly defined, but despite this lack of clarity, two traditions can be distinguished: one follows Firth’s empirical postulate within the British contextualism, the other has its origin in the German-French lexicography (Hausmann 1984). Unlike British contextualism where the main focus is on how frequently two words are combined, Hausmann’s focus is on the semantic interrelationship of these words.

According to Bartsch, a good definition of collocation has to consider both aspects:

The aim of devising a definition of collocation is twofold. In the first instance, it serves as the basis for the computer-aided extraction of collocation candidates, i.e. potential collocations, from the corpus. In a second step, a definition of collocations based on qualitative linguistic criteria must be devised to enable the systematic description of the structural and functional properties of collocations and the distinction between different collocational structures (Bartsch, 2004: 68).

2.1 Distinction between the Term “Collocation” and other Word combinations

In Phraseology, collocations are at the interface between free word combinations and idioms. According to McKeown and Radev (2000: 508) “an idiom, [...], is a given rigid word combination to which no generalities apply; neither can its meaning be determined from the meaning of its parts nor can it participate in the usual word-order variations”. On the other hand, according to Cowie (1981: 223-235) a free word combination can be described using the general rules of grammar, for example, considering the semantic constraints on the words which appear in a certain syntactic relation with a given headword. “Collocations fall between these extremes and it can be difficult to draw the line between categories” (McKeown and Radev, 2000: 508).

2.2 Criteria for the identification of collocations

To determine if and to which extent a word co-occurrence is really a collocation, idiomaticity and stability, together with other aspects must be verified. With this purpose in mind, we present four criterions to identify collocations after the implementation of quantitative methods.

First criterion: verify that the word co-occurrence is not an idiom. As stated above an idiom is a word combination whose meaning cannot be determined from the meaning of its parts. Therefore the first criterion for identifying a collocation is: “word combination whose overall meaning *can* be derived from the meaning of each word”. But idioms are not only on the semantic level fixed associations of words. In contrast to collocations they also can display a fixed syntactic behavior not allowing modifiers, the passive voice, etc.

Second criterion: verify the psycholinguistic stability of the word co-occurrence. Stability can occur as formal, lexical, syntactic, psycholinguistic, pragmatic or semantic stability. Of particular significance for collocation is the psycholinguistic stability which says that collocations like other lexemes are firm parts of the mental lexicon and can be reproduced. Therefore the second criterion is: “collocations are stored and recalled as a unit from our mental lexicon”.

Third criterion: verify the pragmatic stability of the word co-occurrence. The pragmatic stability which depends on the recurrence of words being combined is of similar significance for collocations (Caro Cedillo, 2006: 41). The third criterion is: “the institutionalization of a collocation is determined by its frequent usage” (Seretan, 2011: 16).

Forth criterion: verify the word co-occurrence does not admit the substitution of one of its components by a synonym without altering the meaning. Research by Pearce (2002) into collocations extraction produced a method based on the substitutions for synonyms. His method is based on the assumption that in a free word combination, it is possible to substitute one of its components by a synonym without altering too much the meaning. “If a phrase does not permit such substitutions then it is a collocation” (Pearce, 2002: 1533). Therefore, the forth criterion is: “A collocation can only be

identified as such if the speaker has got several collocators available which can be combined with a base from the semantic point of view, but only one of these collocators is preferred in use”.

3 Extraction of Collocations from Corpora

Just as there are different definitions of collocation, there are also different methods of extracting collocations from corpora. The following methods have been used so far:

3.1 Methods based on co-occurrence considerations

Depending on research interests and research purposes, different association measures were developed. So the corpus linguist has to decide in favour of certain association measures. According to Evert (2009: 1236), there is no “perfect” association measure. Therefore several ones should be used for a study in order to get different results which can be compared. There are two groups of association measures that pursue opposing goals: “effect-size measures” (mutual information, Dice, odds, ratio) and “significance measures” (z-score, t-score, simple II, chi-squared, log-likelihood). The linguist using “effect-size measures” is looking to find “How much does observed co-occurrence frequency exceed expected frequency” and when using “significance measures” “How unlikely is the null hypothesis that the words are independent?” (Evert, 2009: 1228).

3.2 Methods based on Collocation Patterns

Hausmann’s collocation typology (1989: 1010) distinguishes six types of collocations: noun + noun, adjective + noun (as an object), verb + noun (as a subject), adverb + verb and adverb + adjective. Weller and Heid (2010: 3195) call these types “collocation patterns”, once extracted from corpora they call them “collocation candidates”. In order to extract collocations from corpora using “collocation patterns” the corpora have to be annotated at least with POS-tags. However, a higher performance can be reached if the corpora are also syntactically analyzed (parsed).

3.3 Hybrid methods

Pamies and Pazos (2005: 317-329) compare different association measures and come to the conclusion that mathematical methods alone are not enough and prefer instead the use of “hybrid” methods (2005: 327). There are methods which use a combination of co-occurrence calculation and linguistic criteria in the form of collocation patterns. That is, a hybrid system combines statistical methods and multilingual parsing for detecting accurate collocational information. But, in which order are the filtering methods to be used? There are different approaches:

- First statistical methods are used and – for refining the filtering – additional collocation patterns are applied to the achieved results (Smadja 1993).

- First collocation patterns are used and in the second step the statistical methods are applied to the achieved results (Krenn (2000), Evert (2005), Seretan and Wehrli (2006)).

Smadja (1993) is the most representative researcher for the first approach. He developed a system called Xtract that retrieved word pairs using a frequency-based metric in the first place. The metric computed the z-score of a pair of words. In addition to the metric, Xtract used three additional filters based on linguistic properties. As a final step, an evaluation of the retrieved collocations was carried out by a lexicographer in order to estimate the number of the true lexical collocations retrieved.

Another example for the first approach is ConcGram 1.0. This software was developed as an “inclusive” search engine for phraseological units and works on the basis of co-occurrence considerations. It is left to the linguist which of these co-occurrences are significant word combinations and which are chance word co-occurrences. After entering a command, the software compiles a list of “unique words” which are the basis for showing “ConcGrams”, whereas ConcGrams are both adjacent and non-adjacent word co-occurrences which can appear in any order in the corpus. As soon as the ConcGram lists are compiled, statistical methods can be applied. They allow a reduction of the lists and provide clear information on non-relevant word combinations which can be ignored. The applied statistical methods are t-score and MI tests whose formulas are explained in detail by Barnbrook (1996: 88-106). When doing computer-assisted corpus analysis though, automatically compiled frequency tables are – despite using associations measures – not always directly usable, but need a human selection input. For this reason, when working with this kind of software the researcher has to apply manually collocation patterns in order to be able to extract collocations candidates. We also suggest to improve the process by adding work stages based on the criteria presented in section 2.2.

On the other hand, Seretan and Wehrli are the most representative researchers for the second approach. They consider that “syntactic analysis of source corpora is an inescapable **precondition** for collocations extraction” (2006: 1). The hybrid method Seretan and Wehrli (2006) developed relies on a deep parser called Fips (Wehrli, 2004) and can be seen as a two-stage process. Firstly the collocation candidates are identified by the parser while POS-tagging and parsing the text corpora. Secondly the candidates are scored and ranked using specific association measures (Seretan and Wehrli, 2006: 2). In this approach the parser is used in the first stage of the extraction in order to identify the collocation candidates and the criterion they employ firstly for the selection of the collocation candidates is the syntactic proximity. As Seretan and Wehrli explain, “as the parsing goes on, the syntactic word pairs are extracted from the parse structures created, from each head-specifier or head-complement relation. The pairs obtained are then partitioned according to their syntactic configuration” (2006: 2). An advantage of this method is that the pairs obtained are then partitioned according to the collocations patterns presented by Hausmann (1984). A major disadvantage, however, is the dependence on a specific linguistic theory. Finally, the log-likelihood test is applied.

4 Extraction and comparison of word co-occurrences

The terms “satisfaction” and “loyalty” occur extraordinarily frequently in the English corpus. The occurrence of the Spanish and German equivalents is similarly frequent. Therefore, certain co-occurrences of the words “satisfaction” and “loyalty” were extracted and compared with their German and Spanish analog co-occurrences using the software ConcGram. The words “satisfaction” and “loyalty” – in German “Zufriedenheit” and “Loyalität” and in Spanish “satisfacción” and “lealtad” – belong to the category of business psychology in the LSP of marketing. They occur particularly frequently with the words “consumer/customer”, “consumidor/cliente” and “Konsument/Kunde”. To verify if this word combinations were collocations at all the four criterions explained in section 2.2 were applied. Once checked that they met the criterions, a cross linguistic comparison was carried out.

Table 1. Collocations of “satisfaction” and “loyalty” in the English corpus

Base	Collocation	Frequency	Structure
loyalty	Customer loyalty	83	Noun + noun
	Consumer loyalty	7	
satisfaction	Customer satisfaction	137	Noun + noun
	Consumer satisfaction	0	

Table 2. Collocations of “Zufriedenheit”/“Loyalität/Bindung” in the German corpus

Base	Collocation	Frequency	Structure
Zufriedenheit	Kundenzufriedenheit	197	Copulative compound
	Konsumenten-zufriedenheit	1	
Bindung	Kundenbindung	286	Copulative compound
	Konsumentenbindung	0	
Loyalität	Kundenloyalität	19	Copulative compound
	Konsumentenloyalität	0	

Table 3: Collocations of “satisfacción” and “lealtad” in the Spanish corpus.

Base	Collocation	Frequency	Structure
lealtad	Lealtad del cliente	17	Noun + de + art. (pl.) + noun
	Lealtad del consumidor	31	
satisfacción	Satisfacción del cliente	79	Noun + de + art. (pl.) + noun
	Satisfacción del consumidor	61	

The comparison of the three languages English, German and Spanish has led to the conclusion that the collocations in one language are often expressed by other types of word combinations in the other languages. In the analyzed cases English collocations with the structure noun + noun are often expressed by compounds or possessive markers in the German language, and by prepositional phrases – either with or without article – in the Spanish language.

These findings provide also evidence of the fact that collocations do not preserve their meaning across languages. While the compound “Kundenloyalität” occurs in the corpus only 19 times, the compound “Kundenbindung” occurs 286 times in the same contexts as “customer loyalty” occurs in the English corpus. Therefore we can assume that “Kundenbindung” is the German equivalent for “customer loyalty” rather than “Kundenloyalität” which is the literal translation. Literal translations lead very often to unnatural and awkward sounding formulations.

Last but not least, these findings have also shown that while in the Spanish language the terms “consumidor” and “cliente” are used in the same way, in German and English the terms “Kunde” and “customer” are preferred. This is related to socio-linguistic and pragmatic factors – the words with the root “Konsum-“/“consum-” do have a negative semantic prosody or connotation in both English and German, but not in the Spanish language.

5 Consequences for translation and foreign language training

Among the numerous fields of applications of this study, translation training and foreign language training must be particularly highlighted. For translators, especially when translating into the foreign language, collocations are a frequent source of errors, all the more when it comes to detecting false friends. For the term “customer loyalty” the most obvious German translation would be “Kundenloyalität”. However, after a detailed research in the corpus, we have found out that the term “Kundenbindung” is preferred. These findings are of major importance for translators, especially when translating into a foreign language because collocations are unpredictable for non-native speakers of a language. The conducted corpus search has also shown that while in the Spanish language the terms “consumidor” and “cliente” are used in the same way, in German and English the terms “Kunde” and “customer” are preferred due to the semantic prosody of these words. This leads to a clear consequence for the translator: “consumidor” is not necessarily translated as “Konsument” or “consumer”, which would be the literal translation, but rather as “Kunde” and “customer”.

For foreign language didactics, these findings are equally valuable. The (foreign) language teacher has a base for explaining to the students that not only grammatical aspects are important for the use of a language, but that there are also uses which can neither be explained by grammatical nor semantical rules. Aspects which are related to the use of a language, which are pragmatic and can only be taught and learned in connection with the culture and the values of a society.

References

1. Arntz, R., Picht, H., Schmitz, K.-D.: Einführung in die Terminologearbeit. 7th edn. Georg Olms Verlag, Hildesheim (2014).
2. Barnbrook, G.: Language and Computers; A practical Introduction to the Computer Analysis of Language. Edinburgh University Press, Edinburgh (1996).
3. Bartsch, S.: Structural and functional properties of collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence. Narr, Tübingen (2004).
4. Caro Cedillo, A.: Fachsprachliche Kollokationen. Ein übersetzungsorientiertes Datenbank- modell Deutsch-Spanisch. Günter Narr Verlag, Tübingen (2004).
5. Cowie, A.P.: The Treatment of collocations and idioms in learner's dictionaries. *Applied Linguistics*, 2(3), 223-235 (1981).
6. Evert, S.: The statistics of Word Cooccurrences: Word Pairs and Collocations, PhD Thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart (2005).
7. Evert, S.: Corpora and collocations. *Corpus linguistics: an international handbook*. In: A. Lüdeling and M. Kytö. De Gruyter, Berlin, 1212-1248, (2009).
8. Firth, J. R.: A Synopsis of Linguistic Theory 1930-55. *Studies in Linguistics Analysis*. Oxford: The Philological Society, 1-32, (1957).
9. Fluck, H.R.: Fachsprachen. Einführung und Bibliographie. 5th edn. Francke, Munich (1996).
10. Hausmann, F.J.: Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. In: *Praxis des neussprachlichen Unterrichts* 31 (4), 395- 407 (1984).
11. Hausmann, F. J.: Praktische Einführung in den Gebrauch des Students Dictionary of Collocations. *Student's Dictionary of Collocations*. In E. Benson, M. Benson and R. Ilson, Copenhagen, iv-xviii, (1989).
12. Krenn, B.: The Usual Suspects: Data Oriented Models for the Identification and Representation of Lexical Collocations. PhD Thesis, DFKI & Universität des Saarlandes, Saarbrücken (2000).
13. McKeown, K. R., Radev, D. R.: Collocations. In: Dale, R., Moisl, H. and Somers, H. *A Handbook of Natural Language Processing*, Marcel Dekker, New York, 507-523, (2000).
14. Pamies, A., Pazos, J. M.: Extracción automática de colocaciones y modismos. In: Luque Durán, J., Pamies Bertrán, A.: *La creatividad del lenguaje: colocaciones idiomáticas y fraseología*, Granada Lingvistica, Granada (2005).
15. Pearce, D.: A comparative evaluation of Collocation Extraction Techniques. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas (2002).
16. Seretan, V.: *Syntax-Based Collocation Extraction*. Springer, Heidelberg (2011).
17. Seretan, V., Wehrli, E.: Accurate collocation extraction using a multilingual parser. In: *Proceedings of COLING/ACL*, (2006).
18. Smadja, F.: Retrieving collocations from text: Xtract. In: *Computational Linguistics*, 19 (1), 143-177 (1993).
19. Wehrli, E.: Un modèle multilingue d'analyse syntaxique. In: Auchlin, A. et al., editor, *Structures et discours – Mélanges offerts à Eddy Roulet*, Éditions Nota bene, Québec, 311-329, (2004).
20. Weller, M., Heid, U.: Corpus-derived data on German multiword expressions for lexicography. In: *Proceedings of the Euralex International Congress*, Leeuwarden (2010).

Phraseological Units in Horror Comics: Comparative Study of the Translation into English, French and Spanish from a Multimodal Corpus

María del Carmen Baena Lupiáñez

Universidad de Málaga
Av. de Cervantes, 2, 29016, Málaga, Spain
mariadelcarmenbaenalupianez@gmail.com

Abstract. The main objective of this study is to determine the real relationship between text and image in a comic. We will know the influence of the image and the market of the horror comics in the phraseological units and their translation. Therefore, we are going to study the paralinguistic elements of horror comics. We will create a multimodal corpus in English, French and Spanish to classify the linguistic and the paralinguistic elements of comics from a more graphic and technical point of view. We also propose to carry out a comparative study of the techniques of translation of comics, considering the elements that share and differentiate these comics. Computer tools of constitution and analysis of corpora will allow us to establish several classifications related to paralinguistic and linguistic elements, including phraseological units. In this way, the translator would find a more informative classification and would document and establish correlations between texts more easily.

Keywords: Comics, Paralinguistic Elements, Phraseological Units.

1 Introduction

It was in the 20th century when comics began to be published. These comics showed the canons that were established by the society at that moment: a virile, strong and courageous hero, a fragile and delicate woman, an innocent child and a truly wicked villain, among others.

Subsequently, comics evolved from a purely caricatured genre to much more varied themes. The use of its graphic and formal elements has also evolved. Today there are comics that are like true philosophical essays, or comics that are only iconic elements.

This kind of comics shows the importance that the image acquires in drawing.

This leads us to believe that text and image are two elements that cannot be separated because of their complementarities and interdependences.

In the field of translation, the translator must take into account these complementarities so that the target text has coherence, and the reader of the target language can follow the story without difficulty. The translator should also know the context in

which the comic takes place, the psychology, the purpose of each character, their gestures, their register and the paralinguistic elements of the image. This relation between image and text not only affects the use of phraseological units, but they also include cultural connotations that may be represented in the image. That makes more difficult the translation into the target language and culture. Thus, the phraseological units, and by extension, the complete text, have to be translated taking into account the intrinsic cultural connotation of the phraseological unit and its graphic representation within the vignette, thus avoiding inconsistencies between phraseological unit and image. Consequently, the translation of comics has a greater complexity that we could imagine, due to the multitude of heterogeneous elements to be taken into account during the translation process.

It is in this context that we consider that the translation of comics is a type of specialized translation. It has its own communication codes insofar as the textual and paratextual elements are to be taken into consideration in the translation.

Thus, the translator must not only read the text, but also interpret the image that accompanies the text. He or she has to use the appropriate translation techniques according to the message conveyed, adapting the phraseological units and the image to the target culture if necessary.

2 Translation of comics

Despite all that has been explained and the importance of a good interpretation of paralinguistic elements, the Translation studies have not often addressed this subject in a direct and detailed way.

Thus, there are studies that focus on the analysis of comics [8, 9], on the specific characteristics of this genre [10] and on its semiotic aspects [2]. There are also studies that highlight the importance of the image in the translation of comics [12, 21]. We can also find some studies that analyse the gestures and facial expressions on human being in comics [11], but they do not establish a direct relationship between these elements and the translation. Authors such as Zanettin studied both linguistic corpus and comics, which indicates that it is possible to establish a relationship between corpora and comics. This aspect emphasises the complexity that exists to include the translation of comics into a single discipline insofar as comics combine image and text. The above-mentioned areas of translation share these characteristics.

Traditionally, Translation studies have considered the texts with images through the concept of “subordinate translation” [14]. However, with the evolution of comics, this concept is obsolete and corresponds no longer to the graphic representation of the current comics. Today, the concept of “paratranslation” is the most adapted to the translation of comics. We will explain this concept later.

3 Translation, comic and image

As Yuste Frías says [20], the image and all the elements that appear in it are not universal. It may have a different meaning from one language to another, from one

culture to another. For instance, the colour of the image is perceived in a different way according to the language, the culture, the context and the place of the communicative context of the document of departure and that of arrival. Perception is never a simple vision, because a perception involves knowledge, memory, imagination and the cultural environment of both the text to be translated and the story that translates. He points out that colour is a cultural phenomenon that every society, every civilisation, every life, defines and “translate” in a different way according to the spatial-temporal context.

According to Yuste Frías [20], it is necessary to stop the old opposition between the text and the image in translation, in order to stop believing that the translator should only deal with the text. He says that “the new iconotextual entity formed by the text-image pair is a mixed entity where the verbal element is 100% present in the comic and the visual element is 100% present as well” [20]. So, the translator never works on a textual or iconic percentage less than 100%.

As Yuste Frías notices, the author of the term “paratranslation” [20], it is a question of analysing all the elements that accompany the source text to create a target text more appropriate to the reader of the target culture, so that we can avoid misunderstandings or translation mistakes.

4 Adaptation of comics

Zanettin [21] establishes a typology of strategies of visual adaptation in comics (that concerns visual codes, that is, nonverbal), which goes from the publication format to the appearance of the characters and typology (change in the format of publication, in colours, in the images, in the layout; replacing, deleting or adding of pages; resizing, deleting, replacing or adding of bullets, and change of font in the labels).

Zanettin [21] also explains briefly a trendy in the translation and edition of comics: the strategy of foreignization, which means, the prevalence of the non-adaptation.

In conclusion, Zanettin [21] states that this technical translation process is carried out for cultural and commercial reasons, since comics have to be adapted in many occasions to the culture and the target language. A large number of copies can be sold in the countries where it is to be published and bring some benefits to the publishers. The culture and the characteristics of the market play a fundamental role in the components of the comics and in their translation.

5 Relationship between textual corpora and translation of comics

Zanettin [22], the same author who studies the visual adaptation in comics, also studies the textual corpus. Thus, there are authors who consider these two disciplines as specific and complementary to each other, although they are completely different.

Zanettin [22] considers that the textual corpora are very useful for the translator, and that its use is essential to get a good result. According to this author, textual cor-

pore are easy to do, they allow obtaining complete information about the text, and they allow analysing the terminology and the specialized phraseology of the subject.

This could be related to what Corpus Pastor calls “translation technologies” [4]. These are the systems integrated into a translator work environment. These elements interact with each other or are used sequentially and in chain. Corpus Pastor [4] argues that the translator should use, at least, tools such as a translation memory system and an associated terminology programme.

As for the discipline that would involve this study, the Corpus Linguistics, Stéphane Patin [18] points out that it is the area that studies the speech through a digital and structured corpus. In this way, empirical textual corpus that treats the same subject is used and is organised in a way that can be analyzed and interpreted in a structured way.

6 Phraseological units and translation of comics

According to Corpus Pastor [3], phraseological units are lexical units are formed by lexical units that have at least two words and a compound sentence at the most. They are characterized by frequent use, by the co-existence of its component parts; by its institutionalisation; by its idiomaticity and by the degree to which all these aspects appear in the relevant phraseological unit.

Thus, Paula Romero Ganuza [7], following the German model, divides the phraseological units as follows:

— Phraseology shorter than a sentence:

- **Phraseolexems:** They are totally or partially idiomatic, can be modified and are verbal, nominal, adjectival and adverbial units that show different functions.
- **Collocations:** They are constituted by two non-idiomatic components. They are composed by two elements: a base and a placement. The base is an auto-semantic word and determines the second element.
- **Constructions with supporting verbs:** These are non-idiomatic phraseological units that can be considered as a subgroup of the collocation. They are verbal complexes whose structure is verb + preposition + noun, and also as the structure “verb + noun”. In these structures the meaning of the verb is blurred or weakened, residing the semantic nucleus of the construction in the noun.
- **Routine forms:** They may be totally idiomatic, partially idiomatic or may not be idiomatic.
- **Proverbs or sayings:** They are micro-texts that do not possess the communicative function that was attributed to the phraseolexems.
- **Textual forms:** They are non-idiomatic phraseological units, similar to the collocations, but at the sentence level.

However, although phraseological units have been studied a lot, these units in comics and their translation have been studied just a little. There are studies made by Russian researchers, such as those by Pavlova Alla Eduardovna [5] that focus on the

use of phraseological units for the creation of a comic, or the study of Mercedes Ariza [1], which deals with the phraseological creativity in translation of comics. But the studies that combine phraseological units and comics are still very scarce.

7 Objectives

The main objective is to determine the real relation between text and image in a comic, which means, to know the degree of influence of the image and the commercial aspects of the horror comic in the phraseological units and their translation into English, French and Spanish.

Therefore, this project aims to study the paralinguistic elements of horror comics to translate their phraseological units. We are going to create a multimodal corpus in English, French and Spanish which allows to classify the linguistic (phraseological units) and the paralinguistic elements (image) of comics from a more graphic and technical point of view. In this way, the translator can find a more informative classification, can document and establish correlations between texts and images more easily.

In order to achieve this main objective, other ones will be established to complete this study.

It is also planned, thanks to the computer tools of constitution and analysis of corpus, to establish several classifications for the three languages:

- A classification of the paralinguistic elements that appear in all the comics and their function;
- A classification of phraseological units in horror comics;
- A classification of techniques of comics' translation, taking into account all the elements that have been addressed in the objectives set out above.

The project also proposes to carry out a comparative study of the techniques of translation of comics, considering the elements that share and differentiate these comics.

8 Methodology: creation of a multimodal corpus of comics

To achieve these objectives, the study will rely on the analysis of five horror comics:

- **The Crow** by James O'barr [17]: It is an American comic book published in 1989 by Caliber Press. This comic has approximately 244 pages.
- **Sin City** by Frank Miller [15]: It is an American graphic novel published between 1991 and 1999 by Dark Horse Comics. The comic has 7 volumes. The first and second have 208 pages; the third has 184 pages; the fourth has 240 pages; the fifth has 128 pages; the sixth has 160 pages, and the seventh has 320 pages.

- **V for Vendetta**, written by Alan Moore and designed by David Lloyd [16]: It is a series of ten comics gathered later in a graphic novel. It was published between 1982 and 1988 by Vertigo Comics. V for Vendetta has 304 pages.
- **The Suicide Forest**, written by Juan Antonio Torres (“El Torres”) and designed by Gabriel Hernández [19]: It is a comic book published between 2010 and 2011 by IDW Publishing. The comic has 104 pages.
- **The Walking Dead** by Robert Kirkman and Tony Moore [13]: These comic books were published in 2003 by Image Comics. It has 157 editions and 27 volumes, which have between 136 and 144 pages each one.

The interest in this kind of comics lies in the fact that the paralinguistic elements are very remarkable. Indeed, there are many symbolic elements and the characters are particularly expressive.

In addition, commercial comics and independent comics of the same genre were chosen to cover a wider field. A comic can change if it has more or less impact for commercial reasons. The format and the story change according to its impact on the market.

These comics will be established in a multimodal corpus. A multimodal corpus is defined as a corpus that links a TXT file with other files in HTML format, and whose database could record the images of the drawings [4]. Thus, systematic linguistic analyses, of a translational nature, can be carried out by means of linguistic and paralinguistic corpus textual alignment in order to appreciate the reciprocal correspondence of translation, as well as analyses more linked to the image, such as the socio-cultural aspects and the adaptation. Text alignment such as AntPconc, MkAlign or Le Trameur will be used.

Once the texts have been separated from the images, the phraseological units that appear and the function they fulfil will be extracted when combined with the rest of the text and with the image. We will use the table that Corpas Pastor [3] proposes to classify the phraseological units to analyse the most frequent phraseological units in horror comics and their function in them.

9 Results

For the first results, we have analyzed some parts of the Spanish, the French and the English versions of “The Suicide Forest” by El Torres, and “The Crow” by James O’barr. Both are comics written originally in English. A chart illustrating these first results is shown below. They all use a colloquial register, even vulgar or slang in the case of “The Crow”, and the passages that have been chosen were not violent, but when the main character and other characters talked to each other and not other actions happened. So the use of swear words or insults was much reduced.

“The Crow” is a black and white comic, which symbolises the violence of the story and the loneliness and sadness that feels the main character, while “The Suicide Forest” is a colourful comic. They use dark green to represent the forest and the dangers inside it, different kinds of blue and grey to express the loneliness of the characters and red in the violent passages. As it can be seen, they are two different comic

books that exemplify the same notions in different ways. These specific characteristics of both comics are represented in different ways in the translation into French and Spanish and in the original English version.

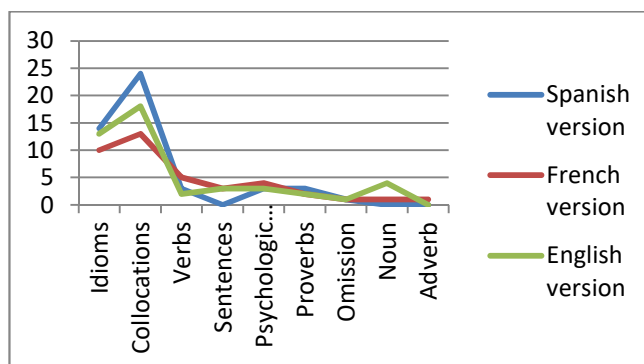


Fig. 1: First results of the study

As can be seen, the first results indicate a large use of idioms and collocations, and some uses of verbs, sentences, psychological formulae, proverbs, omissions, nouns and adverbs. The French version is the one that uses the shortest sentences, while the Spanish version uses the longest sentences. In the French version there is a tendency in using other forms instead of the phraseological units so that their sentences are shorter. However, in the Spanish version the use of phraseological units is larger, so their sentences are longer. The original English version uses very often some phraseological units, but it also replaces them by phrasal verbs or shorter units.

In the case of the collocations, it can be observed that in the three versions of both comics the formulas of “verb + noun” and “verb + preposition + noun” are the ones that repeat the most. With regard to the idioms, the verbal, adjectival and nominal idioms have been mostly found. The omissions are produced in different parts of the story, that means, the Spanish version uses the omission in a sentence that the French or the English do not, and vice versa.

In the Spanish version, it tends more to use a phraseological unit than to elaborate a complete sentence. It is the version that uses collocations (24/30) and idioms (14/30) the most, after the English version, where appears 18/30 of collocations and 13/30 of idioms. Hence no sentences, nouns or adverbs (0/14) have been found in the passages that have been analysed. In the case of the use of verbs (3/14), they have been found when in the French or English versions a phraseological unit has been used.

The French version is the only one that employs all the resources that appear in the chart. It uses more verbs (5/30) and psychological formulae (4/30) than other version. It is also the only one that uses an adverb (1/30) instead of a phraseological unit.

In the English version, the original one, more nouns can be found (4/30) than in their respective translations and the use of collocations (18/30) and idioms (13/30) is also very large.

In short, the Spanish version is the one that uses phraseological units the most, especially collocations and idioms, so the speech is always longer than in other versions. The French version has a use more balanced of the resources so that the sentences are shorter. It is the original English version that mixes these two trends.

References

1. Ariza, Mercedes: Creatividad fraseológica y traducción en la versión española de *Astérix chez les Bretons*, en Ruiz Miyares, L., Alvarez Silva, M. R. (eds.). *Comunicación social en el siglo XXI*, Vol. I, Santiago de Cuba: Centro de Lingüística Aplicada (2011).
2. Celotti, Nadine: The Translation of Comics as a Semiotic Investigator. En el libro de Zanettin (Ed.) *Comics in Translation*, pp. 33-48. London: Routledge (2008).
3. Corpas Pastor, Gloria: *Manual de fraseología española*. Madrid, Spain: Editorial Gredos (1996).
4. Corpas Pastor, Gloria: *Investigar con corpus de traducción: los retos de un nuevo paradigma*. Sweden: Peter Lang (2008).
5. Eduardovna, Pavlova Alla: Phraseological units as means of creation of comic M. in the novel *And. Bulgakov The Master and Margarita*. Dissertation. Sciences: (2003) 10/02/01.
6. Fleury, Serge: *Le Trameur* <<http://www.tal.univ-paris3.fr/trameur/>> (2011).
7. Romero Ganuza, Paula: La delimitación de las unidades fraseológicas (UF) en la investigación alemana y española. *Asociación de Jóvenes Lingüistas*, pp. 905-914 (2006).
8. Groensteen, Thierry: *The System of Comics*. Mississippi, USA: University Press of Mississippi (2009).
9. Groensteen, Thierry: *Comics and Narration*. Mississippi, USA: University Press of Mississippi (2013).
10. Gubern, R. y Gasca, L. (1988): *El discurso del cómic*. Madrid: Cátedra (1988).
11. Gubern Garriga-Nogués, Román: *De los cómics a la cinematografía*. Discurso llevado a cabo para la Real Academia de Bellas Artes de San Fernando (2013).
12. Kaindl, Klaus: Thump, Whizz, Poom: A framework for the study of comics under translation. *Target*, Núm. 11, pp. 263-288 (1999).
13. Kirkman, R., y Moore, T.: *The Walking Dead*. California: Image Comics (2003).
14. Mayoral Asensio, R., Kelly, D., y Gallardo, N.: Concepto de 'traducción subordinada' (cómic, cine, canción, publicidad). *Perspectivas no lingüísticas de la traducción (I)*. Pasado, presente y futuro de la lingüística aplicada en España. *Actas del III Congreso Nacional de Lingüística Aplicada de Valencia*, pp. 95-105 (1985).
15. Miller, Frank: *Sin City*. Milwaukie, Oregon: Dark Horse Comics (1991-1999).
16. Moore, A., y Lloyd, D.: *V de Vendetta*. Nueva York: Vertigo (DC Comics) (1982-1988).
17. O'Barr, James: *The Crow*. Northampton: Kitchen Sink Press (1993).
18. Patin, Stéphane: Apport de la textométrie dans l'analyse d'un corpus bilingue : la traduction pédagogique d'Atala par Simón Rodríguez». *HISTOIRE(S) de l'Amérique Latine*. Vol. 7. Art. 8 (2012).
19. Torres, J.A., y Hernández, G.: *El bosque de los suicidas*. San Diego, California: IDW Publishing (2010-2011).
20. Yuste Frías, José : Traduire l'image dans les albums d'Astérix. À la recherche du pouce perdu en Hispanie » dans RICHET, B. [éd.] *Le tour du monde d'Astérix*. Actes du colloque tenu à la Sorbonne les 30 et 31 octobre 2009, Paris : Presses Sorbonne Nouvelle, pp. 255-271 (2011).

21. Zanettin, Federico: Comics in Translation. London: Routledge (2008).
22. Zanettin, Federico: Corpora in translation practice. Yuste Rodrigo, Elia (Ed.) Language Resources for Translation Work and Research, LREC 2002 Workshop Proceeding. Las Palmas de Gran Canaria, pp. 10-14 (2002b).
23. Zimina, Maria et Fleury, Serge: MkAlign <<http://www.tal.univ-paris3.fr/mkAlign/>>. (2006).

Exploring Automated Essay Scoring for Nonnative English Speakers

Amber Nigam

Independent Scholar, New Delhi, India
ambarnigam12@gmail.com

Abstract. Automated Essay Scoring (AES) has been quite popular and is being widely used. However, lack of appropriate methodology for rating nonnative English speakers' essays has meant a lopsided advancement in this field. In this paper, we report initial results of our experiments with nonnative AES that learns from manual evaluation of nonnative essays. For this purpose, we conducted an exercise in which essays written by nonnative English speakers in test environment were rated both manually and by the automated system designed for the experiment. In the process, we experimented with a few features to learn about nonnative phraseology and its impact on the manual evaluation of the essays. The proposed methodology of automated essay evaluation has yielded a correlation coefficient of 0.750 with the manual evaluation.

Keywords: Automated Essay Scoring (AES), Natural Language Processing, Machine Learning, Latent Semantic Analysis (LSA), Random Forest.

1 Introduction

There are different versions of Automated Essay Scoring (AES) and lack of generalizability across different analyses and corpora prompts a question over the validity of one-size-fits-all AES.

Furthermore, nonnative analysis is differentiated from the native analysis on a few aspects. For example, it is difficult to detect context in essays that have errors specific to some nonnative usages. Moreover, a few valid nonnative spellings like Qutub Minar and Karur are not a part of standard English. This, however, does not render the usage of these words incorrect.

In this paper, we have discussed our methodology for nonnative essay evaluation. First, we have described the feature set that we used for our experiments. Second, we have discussed various adjustments that were made to the system to make it learn and account for nonnative phraseology from manual evaluation. Finally, we have discussed the results of our experiments.

2 Related Work

Although Automated Essay Scoring (AES) has been widely used in many of the real-world applications, there is very limited published work on rating nonnative speakers. Following analyses deal with nonnative speakers in one way or another.

e-rater system™, developed by Educational Testing Service (ETS), is one of the tools that automates scoring of English essays of native and nonnative speakers. In their analysis of e-rater, Jill Burstein, et al. (1999), reported that even when 75% of essays used for model building were written by nonnative English speakers, the features selected by the regression procedure were largely the same as those in models based on operational writing samples in which most of the sample were native English speakers. The correlations between e-rater scores and those of a single human reader were about .73. However, as mentioned in the paper, there were significant differences between final human reader score and e-rater score across language groups, and more data is needed to build individual models for different language groups to examine how this affects e-rater’s performance.

Another analysis in this field is by Sowmya Vajjala, (2016), which is the first multi-corpus study using TOEFL11SUBSET and First Certificate in English (FCE) datasets. For TOEFL11SUBSET, the best model achieved a prediction accuracy of 73% for classifying between three proficiencies (low, medium and high), using all the features. However, the study concludes by stating that the features do not seem to be completely generalizable across datasets.

Current research differs from the existing research in its feature set, unique methodology, and an essay corpus composed of essays written by candidates whose native language is Hindi. Besides, some of our features like lexical density and readability have either not been reported or reported with less significance in earlier analyses on nonnative AES. We have also successfully implemented grammar error correction for a better context detection (Alla Rozovskaya and Dan Roth, 2016) and an automated correction mechanism for whitelisting nonnative spellings that are not a part of standard English.

We are not referring to generic AES systems like by Dimitrios Alikaniotis, et al. (2016) and by Kaveh Taghipour, et al. (2016) because current paper is only concerned with nonnative AES.

3 Experiment

The experiment encompassed building an automated scoring model after learning from training essays (as shown in Figure 1). In all, there were more than 900 essays of length between 150 and 400 words across 7 unique topics (see Table 1). The mean

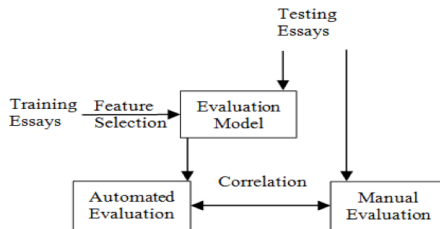


Figure 1: Process flow of the experiment

and the standard deviation for word count of these essays were 247 and 43 respectively. Essay topics were carefully chosen to be on commonly known issues so that they are easy to comprehend and write about. The test takers were all undergraduate students whose native language is Hindi to keep the analysis independent of test taker's native language. Each essay was manually scored on a scale of 1-10 by two raters, both Professors of English at a reputable Indian University, and they had .81 Cohen's kappa statistic (Cohen J, 1968) between their ratings. The raters were informed about the overall experiment and its intent, which is to holistically judge written English of the essays majorly on content, coherence, complexity, and adherence to rules.

We used LanguageTool (Daniel Naber, 2003; LanguageTool, 2012) for detecting grammatical errors. Other features were evaluated by the software developed for the experiment. Besides, the software was designed to detect cheating attempts such as repeating content, writing out of context, and excessive usage of irrelevant words. Feature selection and machine learning experiments were done using Waikato Environment for Knowledge Analysis (Weka) toolkit (Hall et al., 2009). For machine learning experiments, the split between training, validation, and testing sets was approximately 60:20:20.

Essay Topics
Corruption in politics
Role of ambition in career
Factors of motivation for an employee
Power leads to Corruption
Importance of leadership qualities
Should juveniles be tried as adults?
Adverse effects of climate change

Table 1: Essay topics used in the experiment

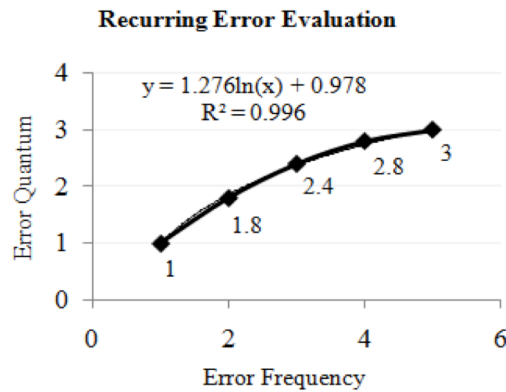


Figure 2: Variation of error quantum with repeated error frequency for a single word

4 Feature Set Selection

Feature	Correlation
Grammar Error Density	-0.396
Grammar Error Coverage	-0.295
Spelling Error Density	-0.146
Spelling Error Coverage	-0.137
Readability	0.326
Lexical Density	0.393

Table 2: Prominent features and their correlations with the manual score

We used Principal Component Analysis (PCA) to choose our final feature set from a larger set of features. Correlations between manual scores and the top few features were then evaluated (see Table 2) to understand the relative significance of the features in manual evaluation. The top features are described below:

4.1 Grammar error density (i.e. grammar errors per unit length)

In this paper, grammar error density refers to the grammatical error count per 100 words. We use density rather than simple error count to ensure that candidates who have written longer essays are not unduly penalized over those who have written shorter essays with the same grammar error density.

A granular categorization of grammar errors has helped in classifying the errors into severity bands based on their impact on manual evaluation. For example, it was observed that a subject-verb agreement error was generally more severely penalized than a style based error in the manual evaluation. Following are the buckets in which grammar errors have been classified:

Major errors like wrong form, incorrect tense, and agreement errors: This bucket includes severe grammar errors that degrade the quality of the essay and possibly cause serious comprehension issues. For instance, replacing “his” with “her” to form a sentence like “He is known for her intelligence.” would be detrimental to the semantics of the sentence.

Capitalization errors: Such errors occur when case of the character is not what it ought to be. This might happen when the first word of a sentence begins with a letter in lowercase. Proper nouns like Paris, Shakespeare also need to be capitalized. Also, personal pronoun “I” is always written in capital when used. Other pronouns are capitalized only if they begin a sentence.

Typography errors: These are also known as typographical errors and are not same as spelling errors. They result due to mechanical failure or slips of the hand or finger. For instance, “water” typed as “wster” due to S key being close to the A key. Typos

generally involve duplication, omission, transposition or substitution of a small number of characters.

Style based errors: These errors include usage of informal language or shorthand like using “u” instead of “you”.

Common replacement errors: Such errors happen in case of a sound-alike or look-alike word pair when one of the words is replaced by the other word in the pair. An example of such word pair is “affect” and “effect”.

Punctuation errors: These errors refer to incorrect usage of comma, semi-colon, colon, apostrophe, hyphen, etc. in a sentence or paragraph. Misplaced punctuation can sometimes alter the meaning of a sentence. For example, a sentence like “Jane finds inspiration in cooking, her family, and her dog.” without commas would be read as “Jane finds inspiration in cooking her family and her dog.”.

Miscellaneous errors: All other grammar errors are put under miscellaneous bucket. These include errors such as repetition of words, improper white space usage, etc.

4.2 Grammar error coverage

It is the count of type of grammar errors per 100 words in an essay. Grammar error coverage highlights the spread of errors across different grammar buckets.

4.3 Spelling error density (i.e. spelling errors per unit length)

Spelling error density of an essay is referred to as the number of spelling errors in the essay per 100 words.

Penalty for a recurring error is based on error quantum, which is evaluated by damping the error frequency (as shown in Figure 2), to lower incremental penalty for a single recurring error.

4.4 Spelling error coverage

It is the count of unique spelling errors per 100 words in an essay. The feature complements spelling error density by accounting for cases where a difficult spelling might be repeated in the essay due to a difficult topic.

4.5 Readability

Readability attempts to estimate the complexity of phraseology used in a text. Why is it relevant in nonnative speakers’ analysis?

It helps distinguish between the candidates who can articulate their thoughts into a syntactically correct and, if required, complex structure, and those who cannot. Because some nonnative speakers lack even the basic ability of constructing appropriate text, it becomes a very important feature. Our data shows that readability alone has a strong

Algorithm	Correlation
Random Forest	0.750
Random Subspace	0.738
Bagging	0.731
M5 Rules	0.706
Gaussian Processes	0.681

Table 3: Correlations between scores by machine learning algorithms and manual scores

correlation with manual scoring. We use Flesch–Kincaid grade level (Kincaid JP, 1975) that is evaluated on word count (WC), sentence count (SC), and syllable count (SyC) by the equation:

$$0.39 * \left(\frac{WC}{SC}\right) + 11.8 * \left(\frac{SyC}{WC}\right) - 15.59 \quad (1)$$

4.6 Lexical Density

It measures the ratio of lexical words to total words that include lexical and grammatical words (Ure, J 1971). Lexical words give a text its meaning and include nouns, adjectives, most verbs, and most adverbs. Grammatical words act as syntactic sugar and include pronouns, prepositions, conjunctions, etc. Lexical density gives a measure of the breadth of content in an essay. This is another factor that is strongly correlated with manual scoring.

The formula for lexical density (LD) is:

$$\frac{N_{lex}}{N} * 100 \quad (2)$$

where N_{lex} is number of lexical word tokens (nouns, adjectives, verbs, adverbs) and N is the total number of tokens in the text.

4.7 Context/Relevance

Latent semantic analysis (Foltz, et al., 1998) is used to detect the context of the topic using n-grams of phrases from word count 1 to 5. Our data shows that certain nonnative erroneous multi-word expressions (MWE) make context detection difficult and that we are better able to detect context in the essays with such errors through grammatical error correction (Alla Rozovskaya and Dan Roth, 2016). Context is detected by checking if cosine similarity between vector of the evaluated essay and vector of at least one corpus essay is more than a specified threshold.

4.8 Coherence

We use Grid model (Lapata and Barzilay, 2005; Barzilay and Lapata, 2008) to detect coherence. This attribute accounts for the flow and structuring of an essay.

5 Self-correction Mechanism

One of the unique selling propositions of the engine is that it pops unknown spellings/phrases once their cumulative frequency is beyond a pre-defined threshold. This has helped us add many nonnative spellings into our repository.

6 Results

We used machine learning algorithms to predict the scores of essays and Random Forest algorithm’s predictions were closest to the manual scoring as shown in the Table 3. The correlations that we report are statistically significant (given the parameters of the experiment) for a significance level of 0.01. Intuitively, error densities like grammatical error density and spelling error density that predict adherence to rules are among the top predictive features. Besides, lexical density and readability are also some of the important features, which underscores the importance of a lexical complexity metric for nonnative Automated Essay Scoring (AES).

7 Conclusion

In this paper, we presented a methodology of rating English essays of nonnative speakers that includes but is not limited to selecting a relevant feature set for the evaluation, categorizing grammar errors into finer types to learn about their importance from their distinctive treatment in contribution to the manual evaluation in nonnative context, treating essays for typical nonnative grammar errors (Alla Rozovskaya and Dan Roth, 2016) that improved context matching for essays with nonnative errors, and devising a self-correction mechanism to learn and promptly address nonnative spellings and styles. Our results show that these incremental adjustments have cumulatively helped in better alignment of automated evaluation with manual evaluation.

Acknowledgements We would like to thank Professor Dan Roth for his valuable feedback and guidance, especially in detecting context of grammatically incorrect essays. We would also like to acknowledge the efforts of Pragati Jaiswal whose valuable suggestions and proofreading were of great help to us. The efforts of Samiksha Sharma in furnishing data in the desired format are also acknowledged.

References

- Alla Rozovskaya and Dan Roth. Grammatical Error Correction: Machine Translation and Classifiers ACL (2016)
- Cohen, J.: Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychol. Bull.* 70, 213–220 (1968)
- Daniel Naber. A Rule-Based Style and Grammar Checker, Diploma Thesis, University of Bielefeld, 2003
- Dimitrios Alikaniotis, Helen Yannakoudakis, Marek Rei. Automatic Text Scoring Using Neural Networks. In *Proceedings of ACL*, pp. 715–725, 2016.
- Kaveh Taghipour, Hwee Tou Ng. A Neural Approach to Automated Essay Scoring. In *Proceedings of EMNLP*, pp. 1882–1891, 2016.
- Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25(2&3), 285-307.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.
- Jill Burstein and Martin Chodorow. Automated Essay Scoring for Nonnative English Speakers Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Memphis, Tenn: Naval Air Station; 1975.
- LanguageTool. Style and Grammar Checker. Retrieved 2013-09-04 from <http://www.languagetool.org/>; 2012
- Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, volume 5, pages 1085–1090.
- Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- Sowmya Vajjala. Automated assessment of non-native learner essays: Investigating the role of linguistic features
- Ure, J (1971). Lexical density and register differentiation. In G. Perren and J.L.M. Trim (eds), *Applications of Linguistics*, London: Cambridge University Press. 443-452.

Automatic Annotation of Verbal Collocations in Modern Greek

Vasiliki Foufi¹[0000-0002-8625-0734], Luka Nerima²[0000-0003-4247-7096] and Eric Wehrli³

¹ University of Geneva, Switzerland
Vasiliki.Foufi@unige.ch

² University of Geneva, Switzerland
Luka.Nerima@unige.ch

³ University of Geneva, Switzerland
Eric.Wehrli@unige.ch

Abstract. Identifying multiword expressions (MWEs) in a sentence and performing the syntactic analysis of the sentence are interrelated processes. In our approach, priority is given to parsing alternatives involving collocations, and hence collocational information helps the parser through the maze of alternatives, with the aim to lead to substantial improvements in the performance of both tasks (collocation identification and parsing), and in that of a subsequent task (automatic annotation). In this paper, we are going to present our system and the procedure that we have followed in order to proceed to the automatic annotation of Greek verbal multiword expressions (VMWEs) in running texts.

Keywords: Multiword expressions, verbal collocations, annotation, parsing.

1 Introduction

Multiword expressions (MWEs) are lexical units consisting of more than one word (in the intuitive sense of ‘word’). There are several types of MWEs, including idioms (a frog in the throat, break a leg), fixed phrases (per se, by and large), noun compounds (traffic light, cable car), discontinuous words (look up, take off), collocations (take a decision, mobile phone), etc. While easily mastered by native speakers, their detection and/or their interpretation pose a major challenge for computational systems, due in part to their flexible and heterogeneous nature.

In our research, MWEs are categorized in five subclasses: compounds, discontinuous words, named entities, collocations and idioms. While the first three are expressions of lexical category (N, V, Adj, etc.) and can therefore be listed along with simple words, collocations and idioms are expressions of phrasal category (NPs, VPs, etc.). The identification of compounds and named entities can be achieved during the lexical analysis, but the identification of discontinuous words (e.g. particle verbs or phrasal verbs), collocations and idioms requires grammatical data and should be viewed as part of the parsing process.

In this paper, we will primarily focus on verbal multiword expressions, and especially on verbal collocations in Modern Greek. Section 2 will give a brief review of MWEs and previous work. Section 3 will describe how our system handles MWEs, the way they are represented in its lexical database and will also be concerned with the treatment of collocation types which present a fair amount of syntactic flexibility (e.g. verb-object) especially in free word-order languages like Modern Greek. For instance, verbal collocations may undergo syntactic processes such as passivization, relativization, interrogation and even pronominalization, which can leave the collocation constituents far away from each other and/or reverse their canonical order. Section 4 will present the modifications made in order to adapt our system to the gold standard output provided in the framework of the PARSEME shared task¹ so as to evaluate our system and provide results (section 5).

2 Multiword Expressions: Related Work

In one of the first studies in automatic corpus-based collocation extraction, Church and Hanks [1] proposed the association ratio, a metric based on the information theoretic concept of mutual information. In parsing, the standard approach in dealing with MWEs is to apply a ‘words-with-spaces’ preprocessing step, which marks the MWEs in the input sentence as units which will later be integrated as single blocks in the parse tree built during analysis [2, 3]. This method is not really adequate for processing collocations because they have a high morphosyntactic flexibility. Alegria et al. [4] and Villavicencio et al. [5] adopted a compositional approach to the encoding of MWEs, able to capture more morphosyntactically flexible MWEs. Alegria et al. [4] showed that by using a MWE processor in the preprocessing stage, a significant improvement in the POS tagging precision is obtained. However, as argued by many researchers [6, 7], collocation identification is best performed on the basis of parsed material. This is due to the fact that collocations are co-occurrences of lexical items in a specific syntactic configuration. Additionally, Nasr et al. [8] have developed a joint parsing and MWE identification model for the detection and representation of ambiguous complex function words. Constant and Nivre [9] developed a transition-based parser which combines two factorized substructures: a standard tree representing the syntactic dependencies between the lexical elements of a sentence and a forest of lexical trees including MWE identified in the sentence.

Many studies have focused on the automatic identification of Modern Greek collocations. Frantzi [10] applied the C-value method on a specialized corpus in the aim to extract collocations and to enrich the specialized dictionary of Modern Greek. A hybrid method for the automatic extraction of nominal MWEs based on grammar rules and word lists but also on statistical evaluation of structures was constructed in the framework of the project ‘Ekfrasis’ [11]. Linardaki et al. [12] have created a dictionary of MWEs based on automatic extraction and manual validation of candidate

¹ <https://typo.uni-konstanz.de/parseme/index.php/2-general/202-parseme-shared-task-on-automatic-identification-of-verbal-mwes-edition-1-1>

MWEs. Finally, Samaridi and Markantonatou [13] have integrated fixed verbal MWEs in an LFG grammar of Modern Greek.

3 The Fips Parser

Fips [14, 15] is a multilingual parser, available for several languages, i.e. French, English, German, Italian, Spanish, Modern Greek, Romanian and Portuguese. It relies on generative grammar concepts and is basically made up of a generic parsing module which can be refined in order to suit the specific needs of a particular language.

The parsing procedure is a one pass (no preprocessing, no post-processing) scan of the input text, using rules to build up constituent structures and (syntactic) interpretation procedures to determine the dependency relations between constituents (grammatical functions, etc.), including cases of long-distance dependencies. One of the key components of the parser is its lexicon which contains detailed morphosyntactic and semantic information, selectional properties, valency information, and syntactico-semantic features that are likely to influence the syntactic analysis.

3.1 The Lexicon

The lexicon is built manually and contains fine grained information required by the parser. It is organized as a relational database with four main tables:

- words, representing all morphological forms (spellings) of the words of a language, grouped into inflectional paradigms;
- lexemes, describing more abstract lexical forms which correspond to the syntactic and semantic readings of a word (a lexeme corresponds roughly to a standard dictionary entry);
- collocations which describe MWE combining two lexical items, not counting function words;
- variants, which list all the alternative written forms for a word, e.g. the written forms, of British English vs American English, the spellings introduced by a spelling reform, presence of both literary and modern forms in Greek, etc.

3.2 Collocations in the Lexicon

In the introduction, we mentioned that in our research the MWEs are categorized in five subclasses, i.e. compounds, discontinuous words, named entities, collocations and idioms. Let's see how collocations are represented in the lexical database.

Collocations are defined as associations of two lexical units (not counting function words) in a specific syntactic relation (for instance adjective - noun, verb - noun (object), etc.). A lexical unit can be a word or a collocation. The definition is therefore recursive and enables to encode collocations that have more than two words. For in-

stance, the Greek collocation *κάνω απεργία πείνας* ('to make a hunger strike', literal translation) is composed of the word *κάνω* 'make' and the nominal collocation *απεργία πείνας* 'hunger strike'.

In addition to the two lexical units, a collocation entry encodes the following information: the citation form, the collocation type (i.e. the syntactic relation between its two components), the preposition (if any) and a set of syntactic frozenness constraints. Verbal collocations mainly occur in the following morpho-syntactic types:

- Verb-object where the object is a bare noun, e.g. *κάνω βόλτα* 'take a walk';
- Verb-object where the object is a nominal collocation, e.g. *έχω τηλεφωνική επικοινωνία* 'have a phone conversation' (literal translation);
- Verb-preposition-noun, e.g. *φέρω εις πέρας* 'carry out';
- Verb-adverb, e.g. *λαμβάνω υπόψη* 'take into account';
- Verb-adjective, e.g. *κάνω (κάτι) γνωστό* 'make (something) known'.

3.3 Parsing and Collocations

Collocation Identification Mechanism. The collocation identification mechanism is integrated in the parser. Collocations, if present in the lexicon, are identified in the input sentence during the analysis of that sentence. In this way, priority can be given to parsing alternatives involving collocations. Thus collocational information helps the parser through the maze of alternatives. To fulfil the goal of interconnecting the parsing procedure and the identification of collocations, we have incorporated the collocation identification mechanism within the constituent attachment procedure. Our parser, like many grammar-based parsers, uses left attachment and right attachment rules to build respectively left and right subconstituents. The grammar used for the computational modelling comprises rules and procedures. Attachment rules describe the conditions under which constituents can combine, while procedures compute properties such as long-distance dependencies, agreement, control properties, argument-structure building, and so on.

Treatment of Collocations. The identification of a collocation occurs when the second lexical unit of the collocation is attached, either by means of a left attachment rule (e.g. adjective-noun, noun-noun) or by means of a right-attachment rule (e.g. noun-adjective, noun-prep-noun, verb-object). In the example *Η κυβέρνηση έκανε λάθος* 'The government made a mistake', when the parser reads the noun *λάθος* 'mistake' and attaches it as complement of the incomplete direct object of the verb *έκανε* 'made', the identification procedure considers iteratively all the governing nodes of the attached noun and checks whether the association of the lexical head of the governing node and the attached element constitutes an entry in the collocation database. The process stops at the first governing node of a major category (noun, verb or adjective). In our example, going up from *λάθος* 'mistake', the process stops at the verb *έκανε* 'made'. Since *έκανε* 'made' - *λάθος* 'mistake' is an entry in the collocation

database and its type (verb-object) corresponds to the syntactic configuration, the identification process succeeds.

As already pointed out, in several cases the two constituents of a collocation can be very far apart, or do not appear in the expected order. For instance, verb-object collocations may undergo syntactic processes such as passivization, relativization, interrogation and even pronominalization, which can leave the collocation constituents far away from each other and/or reverse their canonical order.

In passive constructions, the direct object is promoted to the subject position leaving a trace, i.e. an empty constituent in the direct object position. The detection of a verb-object collocation in a passive sentence is thus triggered by the insertion of the empty constituent in direct object position. The collocation identification procedure checks whether the antecedent of the (empty) direct object and the verb constitute a (verb-object) collocation. In the example Πάρθηκε η απόφαση ‘It was made the decision’, the noun απόφαση ‘decision’ of the collocation παίρνω μια απόφαση ‘to make a decision’ precedes the verb.

Another transformation that can affect some collocation types is pronominalization. In such cases, it is important to identify the antecedent of the pronoun which can be found either in the same sentence or in the context. The example cited below illustrates a sentence where the pronoun τις ‘them’ found in the second sentence refers to the noun ευθύνες ‘responsibilities’ found in the first sentence. Since the pronoun is the object of the verb αναλάβουν ‘take on’, it stands for an occurrence of the collocation to αναλαμβάνω ευθύνη ‘take on responsibility’: Ας αναλογιστούν τις ευθύνες τους. Να τις αναλάβουν. ‘Let them consider their responsibilities. Should they take them on.’

To handle them, the identification procedure sketched above must be slightly modified so that not only the attachment of a lexical item triggers the identification process, but also the attachment of the trace of a preposed lexical item. In such a case, the search will consider the antecedent of the trace. This shows, again, that the main advantage provided by a syntactic parser in such a task is its ability to identify collocations even when complex grammatical processes disturb the canonical order of constituents.

4 Setup for the Evaluation of the System

In this section, we are going to present the experiment that was performed for Greek and the modifications that were made to our parser in order to fulfill this task. The evaluation (see section 5) was made on the gold standard annotated corpus constructed in the framework of the PARSEME shared task on automatic identification of verbal multiword expressions (VMWEs) [16, 17]. We have focused on the annotation of Greek light verb (or support verb) constructions (παίρνω απόφαση ‘make a decision’). Based on the assumption that verbal collocations are formed by a light (or support) verb [18, 19, 20, 21], we first made a list of light verbs such as κάνω ‘make’, δίνω

‘give’, παίρνω ‘take’, έχω ‘have’ etc. Then, we proceeded to the necessary modifications to our system.

4.1 Implementation

As the parser already includes a collocation identification module and produces full syntactic trees for the constituents of the sentence, including the verbal constructions, we only had to develop a transformation code between the PARSEME and the parser’s input output formats. There were three kinds of transformation needed: (i) the reconstitution of the raw text from the tokenized one that was already provided (ii) the alignment of the provided tokens with the tokens generated by the parser and (iii) the copy of the parser detected VMWE to the tokenized parsemetsv file, i.e. the annotation of the identified VMWEs.

Raw Text. The parser requires raw text input. This led us to develop a pre-processor that reconstructs the original text from the tokenized data provided for the shared task. The pre-processor consisted in concatenating the tokens, taking into account the ns field indicating the presence or absence of a space character.

Tokens Alignment. The shared task evaluation measures being token-based, we had to produce the results using strictly the same tokenization as those given in the data sets. Although, the parsemetsv and the parser’s tokenization of words are identical in general, in numerous cases they differ. The trend in parsemetsv tokenization is to consider two words separated by a space as two different tokens. On the other hand, the parser’s tokenization procedure is based on linguistic criteria, i.e. a token is a significant lexical unit. Thus, the parser groups together two or more words if they form a complex lexical unit, for instance the Greek compound noun φακός επαφής ‘contact lens’, the preposition σύμφωνα με ‘according to’ or complex fixed adverbial phrases such as λίγο-πολύ ‘more or less’. The parsemetsv format exhibits some special treatment for some tokens, e.g. the contracted determiner στο ‘into’ in Greek that generates three lines of data.

Annotation of Collocations. The parser can produce several output formats: syntactic tree, tagger, XML/TEI, etc. We chose the tagger output because it gives all the necessary information for the annotation and, like in parsemetsv, it outputs one token per line. In short, each token is displayed on one line, divided in six columns: the token, the Universal POS tag, the richer tag, the lemma, the grammatical function / valency (if any), the collocation (if any). The annotation is processed sentence by sentence as follows: the parser’s output (aligned with the parsemetsv data file) is sequentially traversed line by line. For each verb token, the following tests are performed (in the following priority order). Note that in every case the annotations are fulfilled in the parsemetsv (aligned) data file:

- if the collocation is lemmatized in the lexicon and its main verb is listed as a light verb, it is flagged (353 constructions were annotated);
- if the verb is a light verb and the grammatical function displays a direct object or a prepositional object, it is flagged (427 constructions were annotated); the parser’s output is then traversed forward until the object is encountered; if the object is not encountered, a backward traversal is performed (in order to deal with the passive forms).

It should be noted that both the canonical and the inflected forms of the verbs and nouns are taken into account.

5 Evaluation and Results

In order to measure the performance of our parser, we used the test data file that contains the reference gold annotations against which system outputs were compared for evaluation and that was made available online after the evaluation phase was over. Evaluation metrics are precision, recall and F1, both strict (per VMWE) and fuzzy (per token, i.e. taking partial matches into account). The token-based F1 takes into account:

- discontinuities (λαμβάνω κάτι υπόψη ‘take something into account’);
- overlapping (έχω τη διαίσθηση και την ψευδαίσθηση ‘have the intuition and the illusion’);
- embeddings both at the syntactic level and at the level of lexicalized components.

However, VMWE categories (e.g., LVC, ID) were ignored by the evaluation metrics. In the evaluation per MWE, our system achieved 0.2913 precision (178/611) with a recall of 0.3560 (178/500) and F-measure of 0.3254.² In the evaluation per token, our system achieved 0.4341 (520/1198) precision with a recall of 0.4266 (520/1219) and F-measure of 0.4303.

6 Conclusion

The performance achieved by the system confirms that deep syntactic information helps to identify MWEs and especially VMWEs. Although the VMWE annotation would be more accurate if it was based on the syntactic tree, the “flat” rich tagger output chosen for the alignment ease with the required parsemetsv tokenization was a good solution. An enhancement of this output would be to implement a token identification scheme so as to establish explicit links between the verbs and their arguments (instead of sequentially traverse the sentence and rely on the orthographic form of the word).

² The system ranked first in the competition achieved 0.3612 precision, 0.4500 recall and 0.4007 F-measure.

References

1. Church K, Hanks P (1990) Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16/1: 22–29.
2. Brun C (1998) Terminology finite-state preprocessing for computational LFG. In: 17th International Conference on Computational Linguistics (COLING). Université de Montréal, Montréal, pp 196–200.
3. Zhang Y, Kordoni V (2006) Automated deep lexical acquisition for robust open texts processing. In: 5th International Conference on Language Resources and Evaluation (LREC). Genoa, pp 275–280.
4. Alegria I, Ansa O, Artola X, Ezeiza N, Gojenola K, Urizar R (2004) Representation and treatment of multiword expressions in Basque. In: Takaaki T, Villavicencio A, Bond F, Korhonen A (eds) Second ACL workshop on multiword expressions: integrating processing. Association for Computational Linguistics, Barcelona, pp 48–55.
5. Villavicencio A, Kordoni V, Zhang Y, Idiart M, Ramisch C (2007) Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, Prague, pp 1034–1043.
6. Heid U (1994) On ways words work together – topics in lexical combinatorics. In: Sixth Euralex International Congress. Amsterdam, pp 226–257.
7. Seretan V (2011) *Syntax-Based Collocation Extraction*. Springer, Berlin.
8. Nasr A, Ramisch C, Deulofeu J, Valli A (2015) Joint dependency parsing and multiword expression tokenization. In: 53rd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, Beijing, pp 1116–1126.
9. Constant M, Nivre J (2016) A transition-based system for joint lexical and syntactic analysis. In: 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Berlin, pp 161–171.
10. Frantzi K (2003) Updating LSP dictionaries with collocational information. In: Archer D, Rayson P, Wilson A, Mc Enery T (eds) UCREL Technical Papers. Special Issue: Corpus Linguistics 2003, 16. Lancaster University, UK, pp 219–226.
11. Fotopoulou A, Giannopoulos G, Zourari M, Mini M (2009) Automatic recognition and extraction of multiword nominal expressions from corpora. In: 28th Annual Meeting of the Department of Linguistics. Aristotle University of Thessaloniki, Thessaloniki, pp 620–633.
12. Linardaki E, Ramisch C, Villavicencio A, Fotopoulou A (2010) Towards the construction of language resources for Greek multiword expressions: extraction and evaluation. In: LREC Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages (ELRA). Valetta, pp 1–8.
13. Samaridi E-N, Markantonatou S (2014) Parsing Modern Greek verb MWEs with LFG/XLE grammars. In: 10th Workshop on Multiword Expressions (MWE 2014). Association for Computational Linguistics, Gothenburg, pp 33–37.
14. Wehrli E (2007) Fips, A “Deep” Linguistic Multilingual Parser. In: ACL 2007 Workshop on Deep Linguistic Processing. Association for Computational Linguistics, Prague, pp 120–127.

15. Wehrli E, Nerima L (2015) The Fips Multilingual Parser. In: Gala N, Rapp R, Bel-Enquix G (eds), *Language Production Cognition, and the Lexicon. Text, Speech and Language Technology*, vol 48. Springer, pp 473–489.
16. Savary A, Ramisch C, Cordeiro S, Sangati F, Vincze V, QasemiZadeh B, Candito M, Cap F, Giouli V, Stoyanova I, Doucet A (2017) The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In: *13th Workshop on Multiword Expressions*. Valencia, pp 31–47.
17. Savary A, Ramisch C, Cordeiro S-R, et al (2017) Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.0). LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, <http://hdl.handle.net/11372/LRT-2282>.
18. Stavrakaki S (1999) KANO: The case of a light verb in Modern Greek. In *12th International Symposium of Theoretical and Applied Linguistics*. Aristotle University of Thessaloniki, Thessaloniki, pp 1171–1185.
19. Kyriacopoulou T, Sfetsiou V (2003) Les constructions nominales à verbe support en grec moderne. *Linguistic Insights: Studies in Language and Communication* 5(1): 163–181.
20. Sfetsiou V (2007) Predicative nouns: methods of analysis for electronic applications (in Greek). PhD dissertation, Aristotle University of Thessaloniki.
21. Foufi V (2014) Morphological, semantic and syntactic description of multiword compound nouns in the form of adjective + noun: applications in teaching Greek as a foreign/second language (in Greek). PhD dissertation, Copy City, Thessaloniki.

Corpus Linguistic Exploration of Modern Proverb Use and Proverb Patterns

Kathrin Steyer

Institut für Deutsche Sprache, R5, 6-13, 68161 Mannheim, Germany
steyer@ids-mannheim.de

Abstract. In my talk, I present an empirical approach to detecting and describing proverbs as frozen sentences with specific functions in current language use. We have developed this approach in the EU project ‘SprichWort’ (based on the German Reference Corpus). The first chapter illustrates selected aspects of our complex, iterative procedure to validate proverb candidates. Based on our corpus-driven *lexpan* methodology of slot analysis I then discuss semantic restrictions of proverb patterns. Furthermore, I show different degrees of proverb quality ranging from genuine proverbs to non-proverb realizations of the same abstract pattern. On the one hand, the corpus validation reveals that proverbs are definitely perceived and used as relatively fixed entities and often as sentences. On the other hand, proverbs are not only interpreted as an interesting unique phenomenon but also as part of the whole lexicon, embedded in networks of different lexical items.

Keywords: Proverb; Patterns; Corpus Driven Approach

1 Introduction

The proverb lives on. A quick glance at corpora or the world of new media is enough to see the endless creativity of fixed sentences. New media revive one of the oldest (not old-fashioned) text types. These colloquial forms of communication resemble spoken language, and perhaps this is why they seem to inspire the speakers/writers to use them to reflect upon their messages and to comment on problems of everyday life, politics, culture and sports.

As a phenomenon of the mental lexicon, proverbs are also of interest to a branch of linguistics that focuses on language structures and functions, as a phenomenon of the mental lexicon, embedded in networks of constructions and relationships with other lexical items.

Language technology, large data collections and sophisticated automatic methods, allow the exploration of current proverb use based on authentic language mass data in a new dimension.

In this paper, I discuss an empirical approach to describing proverbs as frozen sentences with specific functions in current language use, namely corpus-based (see 2.) as well corpus-driven (see 3.) [1] [2]. Corpus analysis helps to conceive their meaning and usage more accurately and nuanced than by pure introspection. Below, I will focus on proverb status and patterns. First, I present our complex, iterative, corpus-based procedure to validate proverb candidates and in the second part, I show how the *lexpan* methodology enables scholars to detect functional restrictions of sentence patterns and the nature of slot fillers.

2 Iterative empirical approach for proverb validation

2.1 The project background

Our proverbial studies have arisen from the multilingual EU project ‘SprichWort. Eine Internetplattform für das Sprachenlernen’ (Proverb. An online platform for language learning)[3]¹. The goal was to examine and describe the similarities and differences of contemporary proverb use in different languages and cultures, based on exhaustive corpus studies.

The results are documented on the online platform *SprichWort* (since 2012 hosted at the Institute for the German Language (IDS) in Mannheim, Germany).² The platform contains of three sections:

- a) A multilingual lexicographic database of 300 proverbs in five languages
- b) A series of didactical exercises for teachers and students
- c) A proverb community – proverbs in social networks like Twitter and Facebook.

For the source language German, we had to run the empirical validation for a total of 2000 German proverbs based on the German Reference Corpus (DeReKo)³. The proverb candidates were extracted from dictionaries, textbooks and collections. As a result, about 900 proverbs could be proved in the corpus (this is 45%).

¹ This project was financed by the EU commission for two years. Aside our own IDS group, the partners came from Austria, Czech Republic, Hungary, Slovenia and Slovakia (2008-2010, 143376-LLP-1-2008-1-SI-KA2-KA2MP) [4] [5].

² The German Database is integrated in the OWID proverb online dictionary that we compile continuously [6].

³ The German Reference Corpus, located at the IDS is the world's largest linguistically motivated collection (31,68 billion words: March 2017) of electronic corpora with written German texts from today and the recent past [7]. For this paper, I used a sub corpus, the W archive (W) with a size of about 9 billion words. All queries (Q) are formulated in COSMAS II, the Corpus Search, Management and Analysis System in DeReKo. The frequencies are in brackets.

In the absence of a comprehensive corpus validation in the past, we first had to develop an iterative methodology. Searching for proverbs in a general language corpus is not a trivial task that could be allocated solely to a machine. Few if any assumptions about the surface form and behavior of a proverb can be made in advance (a priori), because again and again corpus evidence proved our intuition wrong.

Considering established proverb definitions [8] [9] we had to address the following questions:

- 1) Is a candidate indeed a lexicalized sentence or a non-finite clause or merely a phrase?
- 2) Are well-known quotations or new sentences, e.g. advertising slogans, also used frequently as wisdom sentence without the original context of creating?

2.2 Corpus-based validation of the proverb status

2.2.1 Frozen sentence or phrase?

As I mentioned in 2.1., the corpus-based validation is a complex quantitative-qualitative procedure. Each proverb candidate must be examined individually in an alternation of automatic analysis and formulation of hypotheses. If one searches for a fixed sentence structure in the corpus, one will only find this sentence. All possible variations, extensions and reductions will not be covered by this search. Therefore, the best strategy is to start with a wide search which is then gradually restricted.

The first step was to check whether the lexical components of the proverb candidate appear in the same sentence at all. If the search for the lexical components in the same sentence was successful, this can indicate a proverb. KWIC concordance lines help to quickly check the important proverb criterion: Is the form a sentence or a non-finite clause equivalent to a sentence [10]? In some cases, the pure search of two components in a sentence is sufficient to get a positive result like in the following example of the proverb candidate *Not macht erfinderisch*⁴ (Distress makes ingenious⁵).

- (1) P14 *Not macht erfinderisch.* Die Reform beginnt im Kopf.
 RHZ01 *Not macht erfinderisch:* Die Soldaten hatten Mangel an
 Schreibpapier.
 NUN06 Allein in Bayern fehlen zurzeit 30 000 Ausbildungsplätze.
 Die *Not macht erfinderisch*, auch auf Arbeitgeberseite.

In other cases, one has to reduce the query to a very narrow range, e.g. for the proverb candidate *Zeit ist Geld* (Time is money)⁶: In this case, wide queries are not useful because of ‘false hits’, e.g. *Die Fahrer hören nichts vom Lärm* (The drivers hear nothing of the noise) or *Zeit und Geld* (Time and money).

⁴ Q: „Not“ /s0 &erfinderisch (2.973)

⁵ The literal translations are bracketed; semantic equivalents are marked by ee.

⁶ Q: Zeit /+w1:1 ist /+w1:1 Geld (1.976)

There also were surprising observations, particularly in relation to those candidates that seemed to be common proverbs based on our intuition. The analysis, however, did not support this. One example is the candidate *Niemand ist ohne Fehl und Tadel* (Nobody is without faults and blame, ee: Nobody is perfect). *Fehl* is a Middle High German word (meaning: mistake, weakness).

- (2) F95 Wer ohne Fehl und Tadel ist, der werfe den ersten Stein.
 N92 löste diese Aufgabe ohne Fehl und Tadel.
 M98 selbst Heilige sind nicht frei von Fehl und Tadel.

(He who is without *faults and blame*, may throw the first stone
 ... did the task without *faults and blame*
 Even saints are not free from *faults and blame*)

In this case, only the word pair *Fehl und Tadel*, mostly in combination with the preposition *ohne* (without) is fixed, but the contexts vary. It is a very frequent multi-word expression (2.235), but not a proverb.

2.2.3 Quotation or proverb?

One of the common sources for the genesis of proverbs is quotations or citations. To be regarded as a real proverb, a sentence has to be used frequently in daily communication in several situations and also without reference to the original quotation context. Let us take a look of the proverb candidate *Viel Lärm um nichts* (Much noise about nothing, ee: Much ado about nothing). Many occurrences related to the comedy of William Shakespeare. We tried to find as many words as possible that indicated a Shakespearean context in any form. The final, very complex search query excluded such thematic words or tokens like *Shakespeare*, *Hollywood*, *Branagh* or German words for 'comedy', 'clock', 'movie', 'stage direction' and 'actor' or compound words with *theatre* etcetera.⁷ This search reduced the frequency from 4.455 to 2.681 really 'Shakespeare free' occurrences and reflect its usage as a proverb.

This double life as a quotation and a proverb is a very frequent phenomenon in the corpus.

Currently, we also use corpus-driven methods to discover new proverb candidates, among others collocation profiles of proverb introducers or labels like *proverb*, *wisdom* or *as the saying goes* and proverbial keywords, e.g. *world*, *time* or *you shouldn't*.

At the end of this chapter I will make a brief note about *proverb frequency*: Calculating proverb frequency is a complex problem which has no standard solution. There

⁷ Q: (&viel /s0 Lärm /s0 "nichts") %s1 (&Shakespeare oder &Komödie oder &Hollywood oder &Uhr oder &Stück oder &Schauspieler oder &Theater oder &Kino oder &Film oder &Verfilmung oder &Regisseur oder Theater* oder *theater oder Branagh* oder ado oder ZDF oder ORF oder 20.00 oder 20.15 oder *komödie oder Sommernachtstraum oder Sa oder CET oder Zitadelle) (2.681)

will be different results depending on the corpus and the search query that was used. Statements about frequency are only meaningful if one clearly outlines on which corpus basis and with which search queries the numbers have been obtained. It is also recommended that reference is made to proportional frequencies or frequency trends rather than absolute numbers.

3 Proverb patterns – corpus driven

3.1 Slot filler analysis with *lexpan*

The corpus-driven exploration of proverb schemes, so called ‘proverb patterns’ [11]⁸, is one of the most innovative areas in paremiology. The results can also contribute to new researches in pattern-based Phraseology, Construction Grammar and Cognitive Linguistics. Proverb patterns consist of fixed lexical components (‘lexical anchors’) as well as slots. These fillers indicate realizations of those patterns in specific communicative situations, both proverbial and non-proverbial. They can be determined by different degrees of typicality (frequencies). Fillers have similar semantic and/or pragmatic characteristics, but don’t necessarily belong to the same morpho-syntactic category. The nature of filler groups cannot be predicted a priori or by rules but only based on an inductive, bottom-up analysis. For this, we developed the language independent pattern matching tool *lexpan* [13] (free available since March this year). *lexpan* makes it possible to explore corpus data in its own working environment independent of a corpus platform. It can be used for restructuring and annotating the interpreted data and for visualizing it for new forms of lexicographic representation. Currently, we can work with collocation data and KWIC lines.

The KWIC lines have been captured by a search pattern with fixed lexical elements and slots (one or more), and we can observe the proportional relations of the fillers and of the underlying syntagmatic structures.

3.2 Semantic and functional restrictions of proverb patterns

I will demonstrate the approach using three examples. The first pattern is *Andere X, andere Y* (Other X, other Y)⁹. Table 1 shows a *lexpan* filler table of the X and Y slots (counted as bigrams):

⁸ In Folklore, the idea of proverb patterns and underlying abstract meanings and functions dates back to the 19th century. Overviews are given by Röhrich & Mieder [7] and Mac Coinnigh [12].

⁹ Q: \$andere /+w2:2 \$andere (5.715)

filler	frequency	%	comment
Länder ... Sitten	1089	21,24	(countries – customs)
Zeiten ... Sitten	110	2,15	(times – customs)
Stimmen ... Räume	54	1,05	(voices – places)
Räume ... Träume	22	0,43	(places – dreams)
Völker ... Sitten	20	0,39	(peoples – customs)
[...]			
Länder ... Kulturen	13	0,25	(countries – cultures)
Länder ... Regeln	12	0,23	(countries – rules)
Sprache ... Kultur	12	0,23	(language – culture)

Tab. 1. Automatic filler tables of *Andere X, andere Y* (lexpan snippet)

In 21.24 % of all occurrences the fillers are *Länder* (countries) and *Sitten* (customs) for the common German proverb *Andere Länder, andere Sitten* (Other countries, other customs). This result is typical for many patterns: Often proverbs are the prototypical realizations. Because of that, one can assume a single proverb entry in the mental lexicon. An interesting observation is that the majority of fillers in the X position refer to concepts of nationality in a broad sense, the fillers in the Y positions refer to norms of behavior (also in the range of low frequency). In about 30% the X filler is *Länder*. Therefore, this pattern is highly restricted by concepts like nationality and behavior.

My next example treats the pattern *Niemand ist X* (Nobody is X). The X slot is filled by a number of adjectives, but there are only two semantic groups that indicate a lexical pattern quality: PERFECTION and REPLACEABILITY. Table 2 illustrates these filler groups, qualitatively systematized by the *lexpan* feature for manual annotation:

filler	frequency	%	tag	comment
perfekt	171	15,23	[[PERFECTION]]	perfect
unfehlbar	52	4,63	[[PERFECTION]]	infallible
unschlagbar	27	2,40	[[PERFECTION]]	unbeatable
vollkommen	20	1,78	[[PERFECTION]]	perfect
fehlerfrei	10	0,89	[[PERFECTION]]	faultless
immun	10	0,89	[[PERFECTION]]	immune
unersetzlich	56	4,99	[[REPLACEABILITY]]	irreplaceable
unersetzbar	21	1,87	[[REPLACEABILITY]]	irreplaceable
unantastbar	8	0,71	[[REPLACEABILITY]]	untouchable
sakrosankt	5	0,45	[[REPLACEABILITY]]	sacrosanct

Tab. 2. Grouped filler tables of the pattern *Niemand ist X* (Nobody is X)

The prototypical realization of this pattern is again a proverb: *Niemand ist perfekt* (Nobody is perfect). All these fillers indicate expressions of worldly wisdom. These

distinct characteristics become clear when searching the past tense of the sentence: *Niemand war X* (Nobody was X). In this form, no German adjective refers to one of the two concepts (**Niemand war perfekt / unersetzlich*). The other realizations are not expressions of wisdom but regular declarative sentences, e.g. *Nobody was injured / responsible / sad*. Of course, these sentences can also be contextualized pragmatically but this is not incorporated in the meaning of adjectives like *injured* or *responsible*.

A further insight of our exploration is that there exist transition zones range from genuine proverbs to bogus proverbs to non-proverb realisations of the same pattern, e.g. *Wer X der Y* (*He who X Y*). The most frequent and prototypical realizations are also the proverbs *Wer rastet, der rostet* (*He who rests, grows rusty*), *Wer sucht, der findet* (*He who searches, finds*) or *Wer wagt, (der) gewinnt* (*He who takes risks, wins*). The second group are realizations that seem like proverbs because of their typical short proverb structure, but the filler of the X Y slots are strongly context-dependent and often rare: *Wer fastet, der friert* (*He who fasts, freezes*) or (*Wer kämpft, der tötet*) (*He who fights, kills*). One can also find non-proverb realizations which are regular uses of the pattern without the fixed proverb structure: *Wer 60 von 74 Punkten erreicht, der hat bestanden* (*He who reaches 60 of 74 points, has passed*), *Wer es etwas gemütlicher mag, der ist beim Freizeitrudern richtig* (*He who likes it more relaxed, is at the right place with recreational rowing*). Independent of the structure all these realizations still transport the same holistic meaning: ‘If certain facts are true for a person, in consequence the other fact must also be true for him/her’.

In future, proverbs patterns will feature as a new user approach in our dictionary.

4 Conclusion

Our corpus linguistic exploration of modern proverb use shows on the one hand that proverbs themselves can be realizations of more general patterns and schemes, furthermore, they share attributes and characteristics with non-proverb multi-word units and other lexical items. It is assumed that there are two lexicon entries: once as a lexicalized proverb and once as a pattern that can also be activated for non-proverb use.

On the other hand, our exploration proved that proverbs are definitely perceived and used as relatively fixed entities and often as sentences. Speakers seem to have a strong sentence-level knowledge, even though they do not distinguish proverbs from sayings, mottos etc. This sentence-level knowledge enables them to create analogies and to produce new realizations of the same proverb pattern. Proverbs are more salient in the mind of the speakers, while non-proverb units of the same schema tend to be subject to creative ad-hoc variations.

This raises the interesting question for future research why some proverbs have hardly any variants while others have many. As you can see, strictly corpus-based proverb studies can create a fresh impetus for a pattern-based theory of the lexicon and vice versa.

References

1. Sinclair, J.: Corpus, Concordance, Collocation. University Press, Oxford (1999).
2. Hanks, P.: Lexical Analysis. Norms and Exploitations. MIT Press, Cambridge (MA) (2013).
3. Sprichwortplattform, <http://www.sprichwort-plattform.org/>, last accessed 2017/08/27.
4. Steyer, K. (ed.): Sprichwörter multilingual. Theoretische, empirische und angewandte Aspekte der modernen Parömiologie. Narr, Tübingen (2012).
5. Ďurčo P., Steyer, K., Hein, K.: Sprichwörter im Gebrauch. Unveränderter Wiederabdruck der 2015 in Trnava erschienenen Erstausgabe. Institut für Deutsche Sprache, Mannheim (2017).
6. OWID Sprichwörterbuch, <http://www.owid.de/wb/sprw/start.html>, last accessed 2017/08/27.
7. Institut für Deutsche Sprache: *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2017-I* (Release vom 08.03.2017). Mannheim, www.ids-mannheim.de/DeReKo, last accessed 2017/08/27.
8. Röhrich, L., Mieder, W.: Sprichwort. Metzler, Stuttgart (1977).
9. Hrisztova-Gotthardt, H., Aleksa Varga, M.: Introduction to Paremiology: A Comprehensive Guide to Proverb Studies. Berlin, de Gruyter (2015).
10. Lüger, H.-H.: Satzwertige Phraseologismen. Eine pragmalinguistische Untersuchung. Praesens, Wien (1999).
11. Steyer, K.: Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht. Narr, Tübingen. (2013).
12. Mac Coinnigh, M.: Structural Aspects of Proverbs. In: Hrisztova-Gotthardt, H., Aleksa Varga, M. (eds.), pp. 112-132.
13. lexpan: Lexical Patterns Analyzer. Ein Analysewerkzeug zur Untersuchung syntagmatischer Strukturen auf der Basis von Korpusdaten – An tool for the exploration of syntagmatic structures based in corpus data <http://www1.ids-mannheim.de/lexik/uvw/lexpan.html>, last accessed 2017/08/27.

Life Values Reflection in Idioms: Corpus Approach

Seda Yusupova

Grozny State Oil Technical University, Zhigulevskaya Street 11-60, 364059 Grozny, Russia
seda_linguist@mail.ru

Abstract. The article deals with the analysis of semantics of the English, German, Russian and Chechen idioms representing such life values as trust, consent and success in globalization conditions. In different languages similarities are found in cognitive models of idioms' meanings, conceptualization of trust, consent, and success. The corpus approach has revealed distinctions in the actual meaning of quasi-equivalent idioms, the semantic properties and additional meanings which are not fixed in dictionaries.

Keywords: Idioms, Life values, Corpus approach.

1 Introduction

Life values take an important place in the value system in various cultures. The importance of life values is revealed at the level of a certain, specific person, and also in the context of society, group of people, culture, the country. The hierarchy of life values has a great influence on the psychological state and development of the personality, society.

1.1 Hypothesis

In the most general sense life values are universal for different cultures, but nevertheless there may be the peculiarities characteristic for this or that community, in languages both similarities, and non-trivial distinctions are found/evident.

The aim of this article consists in studying the semantics of the English, German, Russian and Chechen idioms representing life values: trust, consent and success. Cognitive and semantic analysis of idioms will allow to reveal cognitive models of meaning, similarities and distinctions in differently structured languages. Corpus approach will reflect the frequency and contexts of use of idioms, semantic and syntactic properties shown in real contexts; preferential spheres of use, stability or tendency to violation of phraseological integrity; influence of the inner form on actual meaning [2], ability of the compared idioms to appear as equivalents in translation – equivalence at the level of language and speech.

Under the influence of technical progress in the modern world there is such phenomenon as globalization, integration and convergence of different cultures and the people. Development of the international economy, distribution of democratic values

and institutes, migration turn out to be consequence of globalization. In all these processes coincidence of life values as a factor of integration, consolidation and unification, overcoming dissociation and disagreement is of great importance. Values of trust, consent and success are necessary for building relationships, successful communication, development, self-realization and adaptation. The study of the semantics of idioms will reveal the universal, cultural and specific in perception of these values that will promote a dialog of cultures, consent and peace.

2 Methodology

In the work a complex of methods and approaches has been used: cognitive, semantic, comparative, corpus analysis. The material of research includes the English, German, Russian and Chechen idioms collected from phraseological dictionaries of the English, German, Russian and Chechen languages and also contexts of use from The British National Corpus, the Corpus of the Institute of the German language (IDS) in Mannheim, and the National Corpus of Russian. On the example of these idioms it is possible to see the conceptualization of trust, consent and success in different languages. The non-equivalent idioms entering the same taxa have semantic and conceptual similarities, revealing the common cognitive models of meaning. Also such selection of idioms shows specifics of phraseological funds.

3 Content

3.1 Analysis of the idioms describing trust value

In cognitive linguistics the idioms are considered as cognitive structures, and their semantics, meaning and form as a result of mental transformations. Construction of cognitive models of meaning of idioms will allow to reveal various options of their functioning in speech, conceptual similarities and distinctions in different languages.

English. Pin your faith / hopes on sb/sth – "to put your trust in sb/sth; hope for sb/sth".

Trust – an attachment of belief, hopes to someone – making a connection, touching.

..... you really can't afford to pin your hopes on the rather uncertain prospects currently in the wind (BNC, 2009).

German. In German trust is described as a lack of legal confirmation.

Auf Treu und Glauben – "(without legal security), conscientiously, in good faith".

"Wir brauchen ein sicheres System und können künftig nicht mehr auf Treu und Glauben arbeiten". Eine Vertragskündigung schloss er nicht aus, er bezeichnete sie aber als "eher theoretische" Möglichkeit (Rhein-Zeitung, 14.02.2004). (We need a sure employment contract and cannot work *in good faith* in future anymore. He did

not exclude a termination of contract, he called it, however, "rather theoretical" possibility).

The context reflects the necessity of providing protection of the rights of workers.

Russian. *Войти в доверие* (lit. "to come into trust") – "to gain the trust". Trust as making good relationships.

Втираться в доверие (lit. "to creep into trust") – "to get trust by all means, to try to obtain someone's favor". To get trust in the unusual way that is expressed in the verb, with effort, using some means. In the context has a negative connotation, accentuating insincerity of motives.

Он удалял от себя людей порядочных и разумных, к нему стали втираться в доверие карьеристы, интриганы, проходимцы (Борис Ефимов. *Десять десятилетий* (2000)). (He deleted from himself decent and reasonable people, careerists, intriguers, rogues began to ingratiate with him).

Chechen. In the Chechen language the attempt to receive trust to what you are telling, convincing of or the information you report is expressed in the idioms denoting oath assurance, the appeal to oaths.

Лавттан буха зIойла со (lit. "may I go under the earth") – "readiness to die or go to another world as the proof of sincerity, truthfulness of words". Receiving trust – readiness to sacrifice life.

Делан возаллора (lit. "by greatness of God") – "expression of assurance, confirmation, etc. in something". The address to God, for the proof of sincerity and honesty. Trust – faith in God, feeling of reliability and trust at the mention of God.

Налха хьакха (lit. "to butter") – "with flattery, cunning obtain a favor, trust from someone". Gaining trust using extraordinary tools for making the relations.

Trust – belief in sincerity and truthfulness, overcoming distance, gaining the trust, using different means.

3.2 Analysis of the idioms describing consent value

English. *Speak/ talk the same/ a different language* – "to share / not share ideas, experiences, opinions, etc., that make real communication or understanding possible".

*As leaders, we share the same values, and as you said, on so many issues we see the world in the same way. And most of the time, we **speak the same language*** (BNC, 2015).

A gentleman's agreement (also gentlemen's agreement) – "an agreement, a contract, etc. in which nothing is written down because both people trust each other not to break it". Image of the gentleman as a guarantee of honesty, responsibility, reliability.

*For the most part, opponents are cooperative, but they are under no requirement to exchange films with a nonconference team. It is simply **a gentleman's agreement** when they do* (BNC, 2013).

Build bridges (between A and B) – "if you build bridges between people who disagree on sth or who do not like each other, you try to find ways to improve the relationship between them". Consent, friendly relations, peace – construction of bridges – connection, communication, contact.

The goal of the initiative, which was formed from a 2005 grant, is to build bridges with neighborhoods, law enforcement and faith-based groups to create meaningful and sustainable conversations (BNC, 2015).

Meet sb halfway – "to reach an agreement with sb by giving them part of what they want". Consent – meeting someone on the half way – making contact, though not full, to agree partly. In contexts the meaning is to achieve an agreement.

I think we need to compromise, though. If I'm going to see things the way I see them, then we just need to meet halfway (BNC, 2001).

Pour oil on troubled waters – "to try to settle a disagreement or dispute; take action which will calm a tense or dangerous situation". Reaching an agreement – pouring out oil on uneasy water – calming down.

My own temperament is one that is forever seeking to pour oil on troubled waters (BNC, 2000).

Thus, in idioms the spatial metaphor is mentioned. The consent is associated with closeness, disagreement with distance, overcoming disagreement is a construction and mitigation. National and cultural specificity is also expressed in the second idiom, the gentleman means the noble person from whom the high level of behavior is expected.

German. *Dieselbe /die gleiche Sprache sprechen* (lit. "to speak the same language") – "to have the same opinions, the same level and therefore get on well". In the contexts also *dieselbe Sprache reden, die gleiche Sprache haben, lernen*.

a. *Sie kommen aus diversen Ländern und Kulturen, lernen jedoch dieselbe Sprache: Insgesamt 15 Migrationskinder aus Thuis besuchen seit November den Deutschkurs des Pilotprojekts «sprachliche Frühförderung» (Die Südostschweiz, 18.03.2011).* (They come from various countries and cultures, nevertheless, learn the same language: A total of 15 migration children from Thuis visit since November the German course of the pilot project «early language support»).

b. *Das Theater war immer eine Basis, damit sich Menschen besser verstehen können. Es ist eine Kunst ohne irgendwelche Grenzen. Egal, ob man dieselbe Sprache spricht: Man kann Theater miteinander spielen (Die Presse, 08.06.2011).* (The theatre has always been a base, so that people can better understand each other. It is an art without any borders. It doesn't matter whether one speaks the same language or not: you can play theatre with each other).

Eine Brücke schlagen – "to make a connection". In the contexts also such variants as: *bauen, spannen, stellen*. The structure of the idiom can be violated as in the following context.

Doch meine Frau spricht Deutsch und ich Japanisch, das erleichtert vieles. Ausserdem liebe ich es, täglich eine Brücke zwischen Kulturen zu schlagen (St. Galler Tagblatt, 29.08.2014). (However, my wife speaks German and I Japanese, this makes easier a lot. Moreover, I love to build daily a bridge between cultures). The

importance of language in the process of integration, and also the knowledge of culture is emphasized. Art acts as a uniting factor.

Russian. *Находить/найти общий язык* (lit. "to find a common language") – "to try to obtain, reach full mutual understanding".

a. *Я мог разговаривать и **находить общий язык** не только с людьми, но с камнем, речкой, звездами, с любой травинкой* (Ирина Краева. *Тим и Дан, или Тайна «Разбитой коленки»: сказочная повесть* (2007)). (I could talk and *find a common language* not only with people, but with a stone, river, stars, with any grass).

b. *Мне бы именно хотелось научить его **находить общий язык** с теми, с кем интересы не совпадают* (коллективный. *Форум: Компьютерные игры* (2012)). (I would like to teach him to *find a common language* with those with whom interests do not coincide).

In many contexts it is said about talent, ability, importance to find a common language, a positive quality in economy, on service (political), at work.

Chechen. *Кубъга тIе кубг моха* (lit. "to strike a hand with a hand") – "to agree, make a decision with someone about something". Agreement, consent – close contact – overcoming distance.

Цхъа мотт каро (lit. "to find one language") – "to try to obtain, reach full mutual understanding, consent". Consent – common language – existence of means to report information, instrument of impact on the interlocutor, creation of communication, contact.

3.3 Analysis of the idioms describing success value

English. *Have come a long way* – "have made a lot of progress and achieved a lot". The meaning in contexts is "a progress, not material, not connected with wealth, public, scientific progress, improvement and coming to an understanding. Success is presented as an experience and achievement as a result of experience and a long way. Success – a way – to pass a long way from a starting point. Also *be crowned with success* for making progress.

a. *Do you think America is changing for the better? It's clear we **have come a long way*** (BNC, 2014).

b. *But, in terms of women in politics, there certainly are more women serving in the Congress, 20 percent of the Congress. It's not parity, but women **have come a long way*** (BNC, 2014).

German. *von Erfolg gekrönt wurden/sein* (lit. "to be crowned by success") – "to lead to success, finish successfully".

*Die Arbeiten des Studios **waren** über Jahrzehnte **von Erfolg gekrönt**, man hatte unzählige Weltstars unter Vertrag* (siehe Infobox) (*Luxemburger Tageblatt*,

06.11.2010). (The works of the studio were crowned for decades by success, there were countless world stars under contract (see info box)).

Also *weit gekommen* (lit. "gone far"). *"Ich muss ihnen eine Million Blumensträuße schicken – ohne die Kontroversen wäre ich nie so weit gekommen"* (Rhein-Zeitung, 15.12.2016). ("I must send them one million bouquets – without controversies I would never have come so far").

Russian. *Пробивать* (прокладывать, пролагать себе дорогу, путь) / *пробить себе дорогу* (lit. "to punch (to lay the road, way) / to carve one's way") – "to obtain a certain status, success in life, in any field". In contexts the inner form influences the actual meaning, close to direct, literal, to achieve something, to receive something with effort". Success – a road, way – passing a way, an opportunity to move forward.

В свою очередь, рост объемов добычи заставляет российские компании активнее "работать локтями" и **пробивать себе дорогу** на новые рынки сбыта (Василий Богачев. Врата в поднебесную. Достижение экономических договоренностей открывает политические перспективы (2001) // «Известия», 2001.07.23). (In turn, growth of volumes of production makes the Russian companies more actively "work with elbows" and *carve their way* to new sales markets).

Мы должны дать возможность **прокладывать дорогу** в неизведанное тем, кто способен добиваться выполнения общей задачи полета талантом, ответственностью, преодолением трудностей и своих слабостей ради согласия в экипаже и в интересах процветания жизни на Земле (Роль человека в космическом полете (2004) // «Жизнь национальностей», 2004.03.17). (We must give the chance to *carve the way* into unknown to those who are capable to ensure performing the general task of flight by talent, responsibility, overcoming difficulties and weaknesses for the sake of a consent in crew and for the benefit of prosperity of life on Earth).

Далеко пойти [*yiti*] (lit. "to go far") – "to achieve big success in life, big results in sth".

Можете **далеко пойти**. Вверх по служебной лестнице, разумеется (Сергей Романов. Парламент (2000). (You can *go far*. Up the career ladder, certainly).

Chechen. *Некъ баккха* (lit. "to open the way") – "to obtain the status, success in life, in some field". Also has the meaning "open the new horizons, roads", "acquire the right for something". Success – moving forward, far from a starting point on a horizontal scale.

Гена ваха (lit. "to go far") – "to achieve great success in life, big results in something (sometimes about someone not having respect)".

4 Conclusion

Corpus approach has shown that in contexts the idioms reveal new meanings, shades of meaning, significant for finding functional correspondences in different languages,

for the correct use of idioms in speech. At the same time, the analysis of semantics of idioms in different languages and cultures has shown conceptual similarities and distinctions. In all languages the trust is based on belief, lack of proofs and official written legal confirmation. Consent and reaching an agreement as finding a common language, tool for information report, explanation of the position. An important conceptual similarity is the spatial metaphor "trust and consent – being close", "mistrust, disagreement – being far". Success is associated with a road and moving forward on a horizontal scale. Distinctions are caused by national and cultural specifics. The trust, consent and success are significant values in societies that is expressed in the issues touched upon in modern contexts, migration, successful communication in the process of integration and coexistence.

References

1. Baysultanov, D., Baysultanov, D. Chechensko-Russkiy phraseologicheskiy slovar. Grozny: Kniga, 320s. (1992).
2. Baranov, A.N., Dobrovolsky, D. O. Aspekti teorii frazeologii. M.: Znak, 656s. (2008).
3. Frazeologicheskiy slovar russkogo yazika / Pod red. i s posl. A.I. Molotkova. – 7-e izd., ispr. M.: AST: Astrel, 524p. (2006).
4. Duden Redewendungen. Wörterbuch der deutschen Idiomatik 3., überarbeitete und aktualisierte Auflage. (Duden Band 11). Mannheim etc.: Dudenverlag (2008).
5. Oxford idioms Dictionary for learners of English. Oxford University Press. 470p. (2006).
6. <http://www.ids-mannheim.de/kl/.../korpora/>
7. <http://www.ruscorpora.ru/>
8. <http://corpus.byu.edu/bnc/>

Metaphors of Economy and Economy of Metaphors

Antonio Pamies^[0000-0001-8193-9359] and Ismael Ramos Ruiz^[0000-0002-5661-0460]

¹ University of Granada, Spain

² Palacio de las Columnas. Calle Puentezuelas, nº 55, 18071, Granada, Spain
apamies@ugr.es

Abstract: The economic discourse is essentially metaphorical, as it is observed in the analysis of its terminology, where economy is generally represented in terms of other domains. The aim of this study is to establish a relation between the metaphors found in economic discourse and the systemic economy of figurative language. A qualitative and quantitative analysis of the most frequent source domains of these constructions has been carried out. The most productive type of metaphor in the discourse on economy is the so-called *medical metaphor*, where economy is understood as a living organism. We have analyzed a corpus of economic texts from the Spanish press, in order to identify and quantify all the "diseases of the economy", be they terminological, phraseological or purely discursive. We find relevant regularities between the lexicalized metaphors of economic terminology and the internal economy of figurative semantics.

Keywords: metaphor, phraseology, Corpus linguistics.

1 Introduction

The terminology of economics is basically metaphorical, and -at the same time- economy serves as a model for conceptualizing other domains metaphorically, to the point that even linguistics explains the evolution of languages by means of the so-called *principle of linguistic economy* [17], which is a dynamic balance between the employed means (syntagmatic and paradigmatic) and the obtained communicative results¹.

A fundamental resource of this economy is figurative language, which, by structuring one reality in terms of another one, saves having to create and memorize new words, albeit in return for the effort of disambiguate very often. Such analogies are not as arbitrary or unpredictable as they seem, for recurrent associations of ideas have been observed, such as *conceptual metaphors* ([14]: 43); or *culturemes* [22], mechanisms considered as very *productive* (another economic image). This also applies for specialized terminology (cf. [37]). This systematic phenomenon explains why -in the lexicon of modern languages- there is an average of four figurative meanings for each

¹ Thus, for example, a language with few phonemes compensates it with polysyllabic words, and a language with monosyllabic words compensates it with a rich inventory of phonemes.

literal meaning, and that one word can accumulate dozens of meanings throughout its history.

2 Metaphors of economy

The economy of metaphor may be converted into a metaphor of economy, when this field is the target domain. An outstanding amount of research has been dedicated to the metaphors used in the media when speaking about the banking system or the stock market ([1, 2, 9, 13, 21, 30, 39, 40], among others). Since, according to Lakoff & Johnson, *metaphor is pervasive in everyday life, not just in language but in thought and action* (1980:4-6), its omnipresence in economic discourse *also influences individuals' decision making* ([4]: 1405). The perception of economic problems is mediated by these omnipresent metaphorical mappings. Stender [36], in a comparative study on Spanish and German with two similar corpora specialized in the recent economic crisis (*CrisCorp_DE* & *CrisCorp_ES*), concludes that HEALTH is the most abundant model (Spanish: 492 tokens; German: 438), followed by WAR (Sp.370; Grm.225), FLUID PHYSICS (Sp.134; Grm.147). Other important source domains are NATURAL CATASTROPHES; MOVEMENT; PRESSURE; BODY PARTS; COLORS; ANIMALS; PLANTS; NAVIGATION; CONSTRUCTION; TRANSPORT; RELIGION; SPORTS; GAMES; etc. The work of Charteris-Black and Ennis [2] about English and Spanish presents a similar ranking of source domains: WAR (English: 67 tokens, 27 types, 23%) (Spanish 90, 20, 25%), PHYSICAL HEALTH (Eng. 32 tok., 12 typ., 11%) (Sp. 32 tok., 7 typ., 9%), MENTAL HEALTH (Eng. 29 tok., 15 typ., 10%) (Sp. 78 tok., 26 typ., 22%); NATURAL DISASTERS (Eng. 49 tok., 12 typ., 17%) (Sp. 56 tok., 13 typ., 16%), etc. If we put together physical and mental diseases, the HEALTH model would outnumber WAR.

2.1 Economic terminology, and its source domains

Serón's detailed study of collocational metaphors in a corpus of specialized English texts on economy, compared with their Spanish translation by professionals [34], establishes for both languages a list of models, besides the "medical" one. The other iconic models which mainly dominate among the figurative phrasemes are:

ZOOLOGY: *condor position; shark watchers; monetary snake; bear market; galloping inflation; vulture fund; swipe fees...*

BOTANY: *green shoots; to ripe the fruits; economic blossoming; to reap the benefits; the roots of economy; the seeds of development; to harvest millions of dollars; the roots of poverty...*

NATURAL CATASTROPHE: *economic earthquake; seismic shift; M&A drought; to flood the market; economic plague; monetary storm...*

TRANSPORT: *financial highway; to brake consumption; taking its foot off the accelerator; put on the brakes; the train of progress; fuel of economy; to run out of steam; keep the market afloat; buoyant markets; sinking banks; economic taking-off; a hard landing...*

MECHANICS: *job machine; balance of trade; economy is overheating; to cool the economy; to give a quick start; to grease the wheels; monetary flexibility...*

WAR: *to fight inflation; to combat deflation; to beat the index; monetary offensive; to attack the euro; to defeat unemployment; to destroy concurrency; commercial war; to conquer the market; financial strategy; marketing tactics...*

Therefore, the models underlying the metaphors in the target domain of economy are the same as those that rule common phraseology in many languages: BODY, AGGRESSION, MOVEMENT, TEMPERATURE, ANIMALS, PLANTS, RELIGION, etc. [27]. A calculation on Slovak Phraseology by P. Ďurčo ([7]: 730), based on the 8000 Slovak idioms contained in Soták's dictionary finds out that, whatever their target domain, their most productive source domains are: ANATOMY; METEOROLOGY; PHYSIOLOGY; MATERIAL OBJECTS; ZOOLOGY; ABSTRACT CONCEPTS; FOOD & PLANTS; RELIGION; PERSONS; CONSTRUCTION; MATERIALS and CLOTHES.

English folk phraseology on economic affairs shows that, for the profane, this field also has abundant verbal images with the same source domains:

as poor as a church mouse; as poor as a dog; to be moth-eaten; to be vermin-eaten; to have only a shirt on one's back; to live from hand to mouth; put your money where your mouth is; to go through hard times; to fall in poverty; to go through hell; to be hard up; money does not grow on trees; to be wealthy; to be affluent; to be loaded; to be rolling in money; to be flush with money; to be on easy street; to be a sugar-daddy; to be filthy rich; to be stinking rich; to be lousy with money...

2.2 The medical model

Some types of metaphors are more productive than others, not only in frequency, but also in variety, as in the so-called *medical metaphor*, which reflects an organic view of the economy (e.g. *anemic economy; healthy economy; sick economy; financial cancer; to heal the finances; market ill; chronic shortfall; endemic poverty*), and which has been specifically investigated by [21, 30, 38] among others.

In previous studies we have created a Spanish corpus for specific purposes [31, 32, 33], composed of 2333 texts and 2.985.594 words, taken from two important national newspapers (*El País* and *El Mundo*) and 70 specialized economic journals (including *Expansión*, *El Economista*, *Cinco Días*, *Actualidad económica*, *elblogsalmon.com*, *finanzas.com*, *icnr.com*). The compilation included two steps. First, creating a list of potential medical terms (monolexical or multi-lexical) used in the metaphorical constructions on economy, according to the *International Classification of Diseases* (10th Sp. ed.). Second, the resulting list has been enriched with synonyms, and morphological variants. Once the definitive list is established, each term of the list is introduced into the *My News*² platform, which allows accessing both printed and digital publications, in order to extract we extract the occurrences of all the medical terms of our economic texts.

² As its website explains, “MyNews Hemeroteca is the only digital newspaper library of the modern Spanish press and it has become the most used tool in journalistic documentation among information professionals. With MyNews Hemeroteca you will be able to recover the news published in the Spanish press from 1996. It considers more than 190 titles (national, economic, regional, sport and free press)” (<http://hemeroteca.mynews.es/about/>).

Later, we have applied a simplified version of the Metaphor Identification Procedure (MIP) designed by the Pragglejaz group [5] in order to establish, for each lexical unit in a stretch of discourse, whether its use can be described as metaphorical in this particular context. This simplified system entails a selection process instead of an annotation procedure, beginning with the reading of the recovered text, in order to assure that its content really belongs to the economic field, as well as verifying the presence of (at least) the searched term and the existence of a literal medical meaning of this given term. Within the "medical" metaphor, we have found different sub-categories, depending on the kind of disease recalled by the source sub-domain. The following graphic shows their distribution.

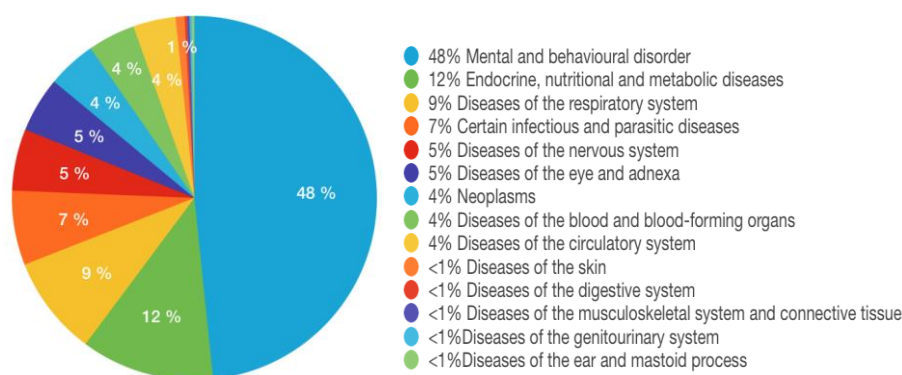


Fig. 1. Sector graph of the average of each disease in the corpus of texts.

To make the study achievable, we limit here to ten the number of recovered texts selected for each term, so we discard a high number of repeated tokens, but we keep all the types, even if they appear only once. Surprisingly, this test finds out that mental and behavioral disorders represent the great majority (48%) of the medical Spanish metaphors of economic discourse. But, of course, these numbers are representative of the variety of each model (types), not of its frequency (tokens). On the other hand, this calculation can include only specific medical terms, not general concepts such as *ill*, *disease*, *heal*, *health*, etc. which could not be classified in such a distribution.

3 Formal economy and semantic economy

3.1. Most of the involved examples are phraseologisms, whether they be collocations (*galloping inflation*), idioms (*to run out of steam*), phraseo-terms (*condor position*), onymic constructions (*International Monetary Fund*) or even proverbs (*don't sell the bear's skin until you hunt it*). From a formal point of view, all of them are characterized by their fixedness, a property which is not limited to the grammatical constraints between the components (e.g. *financial net* neither allows *<*this net is financial>* nor *<*how financial is that web>*). Fixedness has also been defined in terms of frequency (co-occurrence of lexical items higher than randomly expected). This brings us back to the linguistic economy, since frequency is a purely quantitative fact, used as a de-

fining feature of phrasemes. What Firth called *collocation* [in the broad sense] was exclusively limited to this criterion: *words in habitual company* ([8]: 14). Two factors should measure this *mutual expectancy*. For Sinclair & Jones [35], the global frequency of co-occurrence is compared to that of the components separately; whereas Halliday ([11]: 276) compares it to the normal probability of co-occurrence³. Hence, subsequent algorithms have been proposed and investigated for the automatic detection of phrasemes on a statistical basis in large electronic corpora (e.g., [6, 12, 29, 4]), taking for granted that *a phraseologism is defined as the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance* ([10]: 3).

However, the difficulties encountered in the automated detection experiments come to question the principle of co-occurrence frequency of the components, because the global frequency of some idioms is very low [4], or because the individual frequency of some components is very high, for ontological reasons. Mathematically, the number of combinations of several items is potentially higher than the number of items, although this theoretical possibility is not necessarily actualized. For example, Ďurčo ([7]: 730) calculates that the 8000 Slovak idioms included in Sotáks's dictionary comprise 17600 words, with an average of 2,26 "full words" per phraseme, and, although some components are highly recurrent, their average repetition is only 3,40. Besides, in speech, the individual frequency of a word depends on the ontological properties of its referent. Relative frequency is often useful for automatic processing purposes [3], but it is not a reliable criterion in order to define phrasemes from a theoretical point of view.

Therefore, from the point of view of linguistic economy, it is not the formal frequency itself that characterizes phraseologisms, but the multiplier effect of their combinatory nature, both in form and content. Following Martinet's reasoning [17], if natural languages may designate an infinite number of referents, it is thanks to their double articulation: a small number of meaningless units (phonemes) serves to form thousands of meaningful major units (lexemes). This is why Mejri, defines the phraseological level as *the third articulation of language* [18, 19, 20], which consists of combining lexemes within the language system, assigning to the members of this paradigm a global meaning that differs from the one they would have if joined syntagmatically, as they are in free speech (e.g. *to pull one's leg*)⁴. Polysemy and idiomaticity are two faces of the same coin, their economy does not only reduce the formal paradigms level but also the number of "directly accessed" meanings.

³ *collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at n removes (a distance of n lexical items) from an item x, the items a, b, c ...* ([11]: 276).

⁴ For this purpose, Mejri inverts the order of Martinet's articulations: the meaningless (phonemic) level would be the first one, instead of the second.

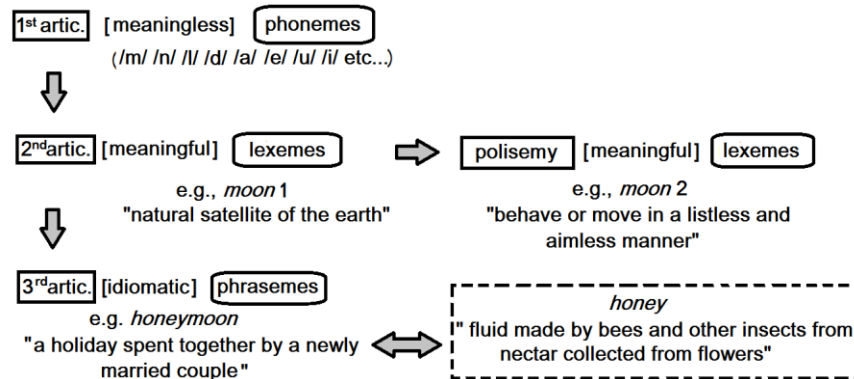


Fig. 2. Our interpretation of Mejri's *troisième articulation* in relation with polysemy.

3.2. Among the mental models underlying metaphors based on other target domains, the notional field of ECONOMY stood out -as a source domain- for long before linguists discovered the *principle of linguistic economy*. Natural languages took notions from economy to represent other concepts, which objectively little have to do with it. Thus, concepts such as POSSESSING, GIVING, LOSING, SELLING, BUYING, STEALING, INHERITING, etc., conceptualize the most diverse actions as if they really dealt with ownable and transferable goods (*to have sorrows, to pay attention, to lose one's nerves, sold out to the enemy, to steal one's heart, a bought off judge*, etc.). However, except as a humorous resource, we do not say <*he lost his nerves and the keys of his car>; <*he has many sorrows and a big house>; <*she stole my heart and my motorcycle>; <*they bought the referee and a new stadium>, because the semantic difference between meanings affects the syntactic behavior of these verbs. POSSESSION, DONATION, LOSS, RETURN, PURCHASE, etc., behave as what Lakoff calls *Idealized Cognitive Models* ([16]: 6), conceptualizing the maximum information with a minimum cost, thanks to the fact that information that is previously structured allows an easier access to the one that is not. Natural languages have even grammaticalized certain tools specialized in marking possession relationships, such as *genitives, datives, or possessive pronouns*. However, by means of *grammatical metaphors* [25], the function of these morphological markers has been also extended to express other relationships, which -in the "real world"- are neither possessive nor dative: *my faculty, my neighbors* (cf. [23, 24, 28]). This bi-directionality between source and target domains is also an essential factor for linguistic economy since it duplicates the productivity of figurative language.

4 Discursive economy and linguistic economy

The economy of metaphor is not limited to phraseologisms (those that are already lexicalized), it also affects free speech. Even in the domain of economics, "creative" metaphors of the speakers are usually understood by the receivers of the message, despite their novelty. E.g. DEATH metaphor, or NATURAL CATASTROPHE metaphor,

magnifying a loss frame, ([40]: 1408), share the same metaphoric macro-models as the figurative lexicalized phrasemes already mentioned. For example, sequences such as:

- a devastating deflation set in...
- after savaging the financial markets...
- deflation is a bigger threat [1].
- both private and institutional investors are still trying to recover from the wounds that the equity crash inflicted on their portfolios...
- higher productivity is the only way that Japan can defeat the twin demons of aging and fiscal deficits...
- there are signs that even moribund retail sales could find themselves resuscitated [34].
- In China, surplus capacity and sliding prices are sounding the death knell for half of the companies making light emitting diode (LED) chips...
- We now inhabit a world of the living dead: a eurozone that will not collapse but cannot be reformed [40].

neither contain idioms nor collocations, but they exploit in free discourse a model which rules idiomatic units, connecting the conceptual domain of aggression with economy. As a consequence of institutionalization, fixedness distinguishes a metaphoric phraseme from a novel metaphor, but it does not prevent the latter from being decoded according to the same semantic model than the former. The most evident variant are the de-automatized idioms, phrasemes that, although being manipulated by the speaker (in form and meaning), are understood by analogy with their model. For example, *familias con la deuda al cuello* (*families with the debt at the neck = "families with a high debt") ([39]: 303) is decoded by analogy with the idiomatic marine metaphor *con el agua al cuello* (*with the water at the neck = "up to one's neck"). This derivative mechanism -from more general iconic models to particular metaphors- is another fundamental pillar of linguistic economy.

We may conclude that there is a kind of qualitative and quantitative analogy and feedback between the systemic logic of linguistic economy (e.g. polysemy, *third articulation*) and the use of metaphors in economy, whose most productive models are systematically reused in figurative expressions mapped onto other target domains. At the same time, economy is the source domain of hundreds of metaphors where economic concepts (POSSESSION, DONATION, TRADE, THEFT, etc.) are conceptualizing the most diverse domains (FEELINGS, EMOTIONS, etc.). This bi-directionality of conceptual mappings is also a highly economic resource.

References

1. Cesiri, D. & Colaci, L.: Metaphors on the global crisis in economic discourse: A corpus-based comparison of The Economist, Der Spiegel and Il Sole 24 Ore. *Rassegna Italiana di Linguistica Applicata* 1(2), 201–224 (2011).
2. Charteris-Black, J., Ennis, T.: A comparative study of metaphor in English and Spanish financial reporting. *English for Specific Purposes Journal* 20(3), 249–266 (2001).

3. Colson, J.: Set phrases around globalization: an experiment in corpus-based computational phraseology. In: Alonso, F., Ortega, I., Quintana, E., Sanchez, M. E. (eds.) *Input a Word, Analyze the World. Selected Approaches to Corpus Linguistics*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 141–152 (2016).
4. Colson, J.: Where does phraseology actually begin? *Yearbook of Phraseology* 6, 1–2 (2015).
5. Crisp, P., Gibbs R., Deignan, A., et al.: MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol* 22 (2007).
6. Daille, B.: Combined Approach for Terminology Extraction: lexical statistics and linguistic filtering. *UCREL*, 5 (Univ. of Lancaster), apud. Adam Kilgariff (1995).
7. Ďurčo, P.: Slovak Phraseology. In: Burger, H., Dobrovol'skij, D., Kühn, P., Norrick, N. (eds.) *Phraseologie: Ein internationales Handbuch der zeitgenössischen Forschung*. Berlin/New York: De Gruyter: vol. 2, pp. 728–736 (2007).
8. Firth, J. R.: *Papers in Linguistics 1934–1951*. London: Oxford University Press (1957).
9. Gallego Hernández, D.: Estudio comparativo trilingüe de la traducción de la metáfora náutica en el lenguaje económico-financiero. In: Mejri, S., Mogorrón Huerta, P. (eds.) *Opacité, idiomaticité, traduction*. Alicante: Universidad de Alicante, pp.99-128 (2010).
10. Gries, S.: Phraseology and linguistic theory: a brief survey. In: Granger, S., Meunier, F. (eds.) *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamin, pp. 3–25 (2008).
11. Halliday, M.: Categories of the theory of grammar. *Word*, 17/3, 241–292 (1961).
12. Heid, U.: Computational linguistic aspects of phraseology. In: Burger, H., Dobrovol'skij, D., Kühn, P., Norrick, N.R. (eds.) *Phraseologie/Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung/An International Handbook of Contemporary Research*. Berlin/ New York: Walter de Gruyter, pp. 1036-1044 (2007).
13. Kelly, P. F.: Metaphors of meltdown: political representations of economic space in the Asian financial crisis. *Society and Space* 19 (6), 719–742 (2001).
14. Lakoff, G., Johnson, M.: *Metaphors We Live By*. Chicago: University of Chicago Press (1980).
15. Lakoff, G.: The contemporary theory of metaphor. In: Ortony, A. (ed.) *Metaphor and thought*, Second edition. Cambridge: Cambridge University Press, pp. 202–251 (1993).
16. Lakoff, G.: *Women, Fire and Dangerous Things: What Categories Reveal about Thought*. Chicago: University of Chicago (1987).
17. Martinet, J.: *Éléments de linguistique générale*. Paris: Armand Colin (1960 [reprint 1980]).
18. Mejri, S.: Délimitation des unités phraséologiques. In Ortiz A. M. L., Huelva Unternbäumen, E. (eds.) *Uma [re]visão da teoria e da pesquisa fraseológica*. Campinas, Pontes, pp. 139–156 (2012).

19. Mejri, S.: La mémoire des séquences figées: une troisième articulation, ou la réhabilitation du culturel dans le linguistique?. In: Cinquièmes journées scientifiques du réseau LTT (AUPELF-UREF), Tunis, CERES, pp. 3–11 (1998).
20. Mejri, S.: Polylexicalité, monolexicalité et double articulation. *Cahiers de lexicologie* 2, 209–221 (2006).
21. O'Mara-Shimek, M.: Metaphor and digital finance reporting of the stock market crash of 2008: Organicist vs. mechanistic visions. Tesis Doctoral, Universidad Católica de Valencia (2011).
22. Pamies A.: The concept of cultureme from a lexicographical point of view. *Open Linguistics* 3/1 (january), 100–114 (2017).
23. Pamies, A., Natale, D.: Observaciones contrastivas sobre las construcciones posesivas y pseudoposivas en español e italiano. *Beoiberística - Revista de Estudios Ibéricos, Latinoamericanos y Comparativos*, 1/1, 11–25 (2017).
24. Pamies, A.: Il concetto di dono nel linguaggio. Conferenza internazionale Da lontano: Dono, Istituzioni e Ospitalità. Napoli, Italia, 27–29 aprile 2016 (in press) (2016a).
25. Pamies, A.: Metafora grammaticale e metafora lessicale: implicazioni teoriche per la fraseologia. In: Dal Maso, E., Navarro, C. (ed.) *Gutta cavat lapidem. Indagini fraseologiche e paremiologiche*, Mantova: Universitas Studiorum, pp. 87–120 (2016b).
26. Pamies, A.: De la idiomatidad y sus paradojas. In: Conde Tarrío, G. (ed.) *Nouveaux apports à l'étude des expressions figées*, Cortil-Wodon (Belgique): Inter-Communications & E.M.E., pp. 173–204 (2007).
27. Pamies, A.: Modelos icónicos y archimetáforas: algunos problemas metalingüísticos en el ámbito de la fraseología. *Language Design*, 4, 9–20 (2002b).
28. Pamies, A.: Sémantique grammaticale de la possession dans les langues d'Europe. In: Castagne, E. (ed.) *Modélisation de l'apprentissage simultané de plusieurs langues apparentées*, Nice: Université Sophia-Antipolis: 67–98 (2002a).
29. Pazos, J. M., Pamies, A.: Combined statistical and grammatical criteria for the retrieval of phraseological units in an electronic corpus. In: Granger S., Meunier, F. (eds.) *Phraseology: an Interdisciplinary Perspective*. Amsterdam: John Benjamins, pp. 391–406 (2008).
30. Peckham, R.: Economies of contagion: financial crisis and pandemic. *Economic and Society*, 42 (2), 226–248 (2013).
31. Ramos Ruiz, I.: El cáncer de la economía: La fraseología de las metáforas médicas periodísticas. *Opción*, 31(76), 747–767 (2015).
32. Ramos Ruiz, I.: La metáfora en el periodismo económico: infecciones económicas. In: Mendieta, A., Santos, C.J. (eds.) *Líneas emergentes en la investigación de vanguardia*. Madrid: McGraw-Hill Interamericana, pp. 515–524 (2014).
33. Ramos Ruiz, I.: La metáfora médica como dominio fuente en los spots electorales de contenido económico. In: Padilla, G. (ed.) *Estrategias en comunicación y su evolución en los discursos*. Madrid: McGraw-Hill Interamericana (2016).
34. Serón Ordóñez, I.: La traducción de la metáfora en los textos financieros: estudio de caso. In: Torres, M. G. (ed.) *Traducción y cultura: el referente*

- cultural en la comunicación especializada, Málaga: ENCASA, pp. 205–250 (2005).
35. Sinclair, J. M., Jones, S.: English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie*, 24(2), 15–61 (1974).
 36. Stender, A.: El lenguaje económico alemán y español de la prensa especializada: análisis basado en un corpus de la crisis económica (CRISCORP). Tesis doctoral. Sevilla: Universidad Pablo de Olavide (2015).
 37. Tercedor, M.: Rutas de metaforización y traducción especializada: una aproximación cognitiva. *Sendebarr*, 10/11, 249–260 (2000).
 38. Vukićević-Đorđević, L.: On Biological Metaphors in Economic Discourse. *Journal of Teaching English for Specific and Academic Purposes*, 2(3), 429–443 (2014).
 39. White, M.: Metaphor and economics: the case of growth. *English for Specific Purposes*, 22, 131–151 (2003).
 40. Williams, A. E.: Metaphor, Media, and the Market. *International Journal of Communication*, 7, 1404–1417 (2013).

A Note on Controlled Composition in Japanese EFL Classes for Intermediate Learners: With a Focus on Reordering Questions

Shimpei HASHIO¹ and Nobuyuki YAMAUCHI²

¹ Doshisha University, Kyoto 610-0394, Japan
smp.hashio@gmail.com

² Doshisha University, Kyoto 610-0394, Japan
nyamauch@mail.doshisha.ac.jp

Abstract. Controlled composition can be an important and useful method for Japanese EFL learners preparing to study Japanese-to-English translation and paragraph writing. This paper focuses on reordering questions, one traditional type of controlled composition. It compares intermediate-level learners' English sentences in reordering questions with their counterparts in translation questions in the same Japanese sentences, and considers what reordering questions should be like and what intermediate-level learners should study in their composition classes. The results show that students are likely to make significantly more errors related to verbs, prepositions, articles, and pronouns in the translation questions than in the reordering questions. In conclusion, English classes for intermediate-level learners in Japan should make better use of reordering exercises as well as Japanese-to-English translation and paragraph writing by giving more explicit instruction designed to make learners more conscious of the uses of the word classes whose errors are likely to occur more frequently when learners both translate and deal with reordering questions.

Keywords: Japanese EFL learners, composition instruction, reordering

1 Introduction

Writing instruction in Japan has long depended on a method called controlled composition including such tasks as fill-in-the-blank, conversion, and reordering, some of which have been adopted in formal school education. Controlled composition are considered to work as a bridge between the basic-level instruction of grammar and vocabulary and advanced-level instruction of Japanese-to-English translation and paragraph writing. Kimura *et al.* [4] argued that the instruction in controlled composition facilitates acquisition of correct grammatical structures and reduces teachers' workload, but it also pointed out that it may degenerate into mere grammar learning or have the effect of demotivating learners so that such questions are prepared according to the intentions of teachers or teaching material designers.

This paper will examine how using reordering questions, the type of controlled composition most commonly adopted in writing instruction, can contribute to the development of learners' sentence production.

2 Previous Research

“Reordering questions” is an activity that requires examinees to reorder shuffled word chunks to make a coherent sentence frequently used in university entrance examinations and writing instruction in Japan. Fig. 1 is an actual question given at the National Center Test for University Admissions in 2006.

Q1 Taking a warm _____ better.
① may ② you ③ help ④ sleep ⑤ bath

Fig.1. A sample of reordering questions

Morishita & Yamamoto [6] reported that a four-month session of reordering training for elementary-level Japanese EFL learners improved their sentence production ability. The report pointed out that after the training, the learners’ scores of reordering questions improved with their increased consciousness of sentence structures. Sase [7] argued that learners with a high rate of correct answers in reordering questions were likely to produce more English sentences in correct word order in speaking, and that a test of storytelling and reordering questions suggested a correlation between them. Furthermore, it pointed out a problem in learners’ sentence production in which their answers in the storytelling test had a tendency to avoid and paraphrase complex structures, a learner behavior called “avoidance” (e.g. [1]).

In addition, in order to reveal the contribution of reordering questions to the learners’ sentence production, we must confirm the characteristic of learners’ English sentences they made from scratch without any hints. Tono & Mochizuki [8] analyzed the JEFLL Corpus, a learner corpus collecting essays written in English by students from Years 7 to 12, and classified the detected errors into the three types of errors proposed by James [3]: misinformation (use of incorrect word class or form), omission (lack of necessary words), and addition (use of unnecessary words). The analysis showed a high frequency of errors in nouns, adverbs and prepositions, and many omission errors in articles and prepositions.

3 Research

3.1 Research Purpose

Based on these previous studies, this paper aims to carefully examine the characteristics of the English sentences used as answers to reordering questions and translation questions made from the same Japanese sentences. It thus identifies the points of attention in the transitional stage between the instruction of controlled composition and that of Japanese-to-English translation and paragraph writing, and examines how reordering should be adopted as part of instruction at the intermediate level between the basic and advanced levels. This study conducts questionnaires of reordering questions and translation questions among Japanese learners of English. It then analyzes the English sentences in the answers.

3.2 Participants

The Japanese EFL learners participating in this study consisted of 20 students from private co-educational high school A in Osaka Prefecture and 20 students from private girls' high school B in Kyoto Prefecture. Many of the students in these two schools had obtained either EIKEN Grade Pre-2 or Grade 2 or the TOEIC scores within the test range between 300 and 500, which indicates that the students of these schools had intermediate-level English proficiency on average.

3.3 Preparing and Conducting the Questionnaires

Twenty reordering questions were selected from Ishiguro *et al.* [2], Kobayashi & Hayasaki [5], and Yamauchi *et al.* [9], which are writing textbooks including reordering questions, and were labeled Q1 to Q20. The degree of difficulty in the twenty questions was adjusted based on the results of the preliminary research. They were then organized into Questionnaire X and Questionnaire Y as shown in Table 1 below for the comparison of answers to different types of questions made from the same set of Japanese sentences. Questionnaire X was answered by participants from high school A. In this questionnaire, Q1 to Q10 were given as reordering questions as in the original text, while Q11 to Q20 were given as translation questions, deleting the chunks of reordering questions and leaving only the Japanese sentences. Meanwhile, participants from high school B answered Questionnaire Y, with Q 1 to Q10 as translation questions and Q 11 to Q20 as reordering questions. Preparing questionnaires in this way enabled the analysis of the answers to apply to both reordering questions and translation questions made from the same set of Japanese sentences. Fig. 3 shows a part of the Questionnaire X and Fig. 4 represents a part of each questionnaire.

Table 1. The detail of Questionnaire X and Y

	Group	Reordering	Translation
Questionnaire X	High school A	Q1-10	Q11-20
Questionnaire Y	High school B	Q11-20	Q1-10

1. 次の(1)～(10)の日本語文に合うように[]内の語句を並び替えよ。

(1) その歌手はたくさん子どもたちに取り囲まれた。[a / surrounded / of / by / kids / lot / singer / was / the].

(2) 部屋に入ると、見知らぬ人が僕を待っていた。[the / waiting / found / a / me / entering / room / stranger / for / I].

⋮

2. 次の(1)～(10)の日本語を英語に訳せ。

(1) こう私に言った男は、名前を名乗ることを拒否した。

(2) 彼は少年の頃から両親に虐待されてきた。

⋮

Fig. 3. A part of the Questionnaire X for high school A

1. 次の(1)～(10)の日本語文に合うように[]内の語句を並び替えよ。
 - (1) こう私に言った男は、名前を名乗ることを拒否した。[who / his / give me / man / told / this / refused / me / name / to / the].
 - (2) 彼は少年の頃から両親に虐待されてきた。[since / parents / been / badly / childhood / has / his / by / he / treated / his].
 - ⋮
2. 次の(1)～(10)の日本語を英語に訳せ。
 - (1) その歌手はたくさん子どもたちに取り囲まれた。
 - (2) 部屋に入ると、見知らぬ人が僕を待っていた。
 - ⋮

Fig. 4. A part of the Questionnaire Y for high school B

3.4 Statistical Analysis

In order to examine the differences in characteristics between the English sentences for the reordering exercises and those of the translation exercises, an analysis of how the three types of errors suggested by James [3] occur in reordering and translation was conducted for each word class. After collecting the sentences given by all participants separately for reordering and translation, word class tags are applied using CLAWS C7 tagset and the numbers of conjunctive (COJ), adjective (ADJ), adverb (ADV), preposition (PRP), article (AT), pronoun (PRN), noun (NOUN) and verb (VERB) are counted on AntConc. Then, the frequencies of the correct uses and errors of eight word classes are organized separately for reordering and translation. Errors are categorized into misformation, omission and addition. Each word class was labeled as “RO_ADJ” for adjectives in reordering and “TR_VERB” for verbs in translation, and the frequencies of the correct uses and three types of errors were calculated. A calculation on a 16x4 cross table and a 4-axe correspondence analysis revealed how word classes in reordering and translation relate to three error types, respectively.

4 Results

Table 2 shows the percentage of correct uses and three types of errors of the eight word classes for reordering and translation, respectively. With most word classes, the percentage of correct uses decreased when the question type changed from reordering to translation. Though this result looks natural based on the results of the correspondence analysis, we further examined what types of error in which word classes became prominent when the question type changed from reordering to translation.

Fig. 3 shows the results of the correspondence analysis of the relation between word classes and correct uses and errors in reordering questions and translation questions. The contribution ratio is 50.2% for the first component and 39.4% for the second component, and Fig. 3 is considered to have the interpretability of approximately 90%.

Table 2. The percentage of correct uses and errors of 8 word classes in questions

Word class	Questions	Misformation	Omission	Addition	Correct use
Verb	Translation	23.5%	3.9%	0.2%	72.4%
Verb	Reordering	3.5%	2.2%	3.0%	91.2%
Article	Translation	4.4%	18.2%	1.8%	75.6%
Article	Reordering	1.0%	5.1%	6.1%	87.8%
Pronoun	Translation	3.0%	12.8%	0.9%	83.3%
Pronoun	Reordering	0.2%	2.3%	4.4%	93.1%
Proposition	Translation	21.1%	16.7%	7.2%	55.0%
Proposition	Reordering	1.8%	0.6%	2.4%	95.2%
Noun	Translation	7.1%	2.6%	0.9%	89.4%
Noun	Reordering	1.1%	1.8%	1.4%	95.8%
Adverb	Translation	2.6%	15.9%	2.6%	78.8%
Adverb	Reordering	0.5%	18.8%	16.5%	64.2%
Adjective	Translation	8.4%	2.7%	0.8%	88.2%
Adjective	Reordering	0.0%	3.0%	3.0%	94.0%
Conjunctive	Translation	5.1%	3.8%	0.0%	91.1%
Conjunctive	Reordering	0.0%	0.0%	0.0%	100.0%

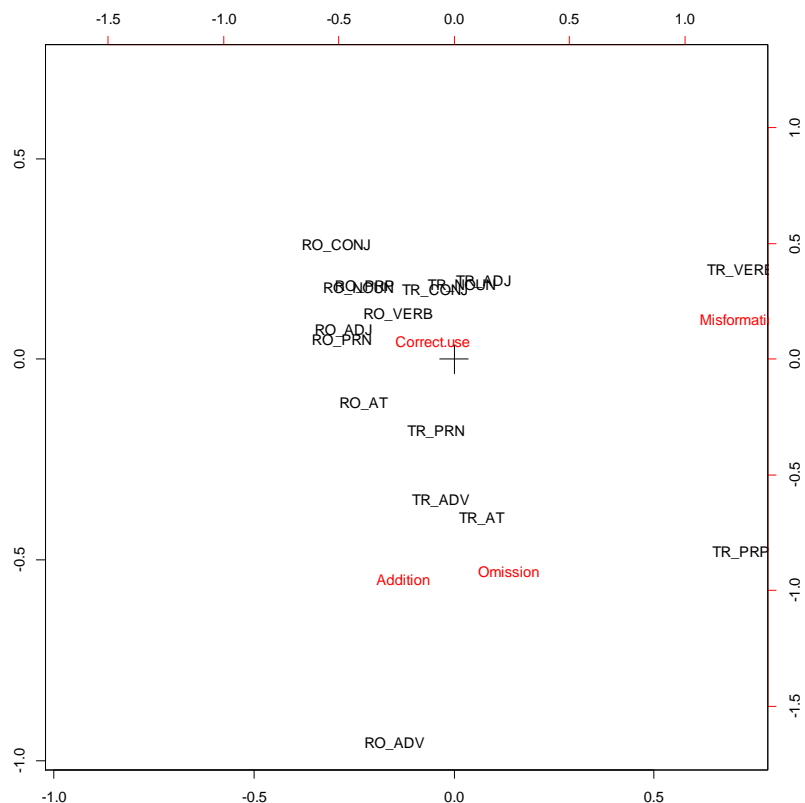


Fig. 3. The result of correspondence analysis between word classes and questions

As seen in Fig. 3, misinformation errors increase in verbs and prepositions in particular when the question type changed from reordering to translation, while omission errors increase in pronouns, prepositions and articles. Regarding verbs, singular/plural and tense errors were observed in the intermediate-level learners, as well. As for pronouns, omissions of second person subject and possessive form occur frequently, and the results also indicate that many students do not seem to understand the appropriate use of prepositions. Note that this paper will not discuss adverb errors as these omission errors occur equally frequently in both translation and reordering questions.

5 Discussion

In this chapter, based on the previous studies and the present results, we will examine the merits and demerits of the traditional instruction of reordering and discuss how we should make better use of reordering in writing instruction.

As Morishita & Yamamoto [6] argued, practicing reordering improves elementary-level learners' consciousness of word order and sentence structure, and as Sase [7] argued, it also leads to the development of Japanese paraphrasing ability so as to choose the most appropriate sentence patterns that learners feel confident in using. But, as indicated by the results in this paper, reordering exercises do not seem to be sufficient for learners to acquire the correct use of verbs, pronouns, prepositions and articles. The chunks in reordering can act as too broad hints and keep learners from finding the errors in these word classes in translation and paragraph writing. For intermediate level, a transition from the instruction of reordering to that of translation and paragraph writing would be ideal. Then, it is necessary to give them explicit instructions with an emphasis on the use of such frequently misused word classes.

Yet, the answers of translation and paragraph writing are not always fixed unlike reordering questions, and scoring learners' answer or giving feedback can be a burden on teachers. Also, as Sase [7] pointed out, since some grammatical items and expressions tend to be avoided by learners, it is also necessary to adopt motivated instruction to prevent the avoidance of complex sentence structures. In the instruction of translation and paragraph writing, however, if we attempt to give explicit instructions with an emphasis on the use of such frequently misused word classes based on the result in this paper, we can easily overcome these dilemmas.

Some strategies of instruction of translation and paragraph writing have been suggested by Kimura *et al.* [4] and others. One effective method is to intentionally choose the type and method of feedback according to the level of learners and specify a time for feedback in every class, so as to draw the attention of learners to potential errors in specific items. When we give the instructions of paragraph writing, aiming at the misuses of verbs, giving feedback to students by focusing on verb tenses and conjugation is as effective in terms of both explicit instruction to learners and reduction of the burden on teachers as we give reordering questions.

In addition, we introduce an example to stop learners' avoidance of complex items and expressions. If we want learners to produce relative pronouns that most learners find difficult to use, another possible way is, for example, to instruct learners to use,

for example, at least 15 words per sentence, which will create a situation where they need to use relative pronouns. Furthermore, a direct instruction like “use a relative pronoun in each sentence” is also instructive in acquiring the ability to use relative pronouns in sentence production. These are a few ways to maintain the advantages of controlled composition in instructing translation and paragraph writing.

Meanwhile, we suggest another way to utilize reordering questions for the acquisition of new grammatical items and vocabulary. The downside of reordering exercises mentioned in this paper is that word chunks give too many hints to the correct answer. In this regard, this paper turns out to be more persuasive in suggesting that the use of reordering questions focusing on frequently misused word classes with the chunks in reordering questions including an unnecessary or missing component in order to make learners more conscious of the usage of particular word classes.

6 Summary and Further Implications

In this paper, we have analyzed the characteristics of the English sentences in both reordering and translation exercises and also clarified how learners’ error patterns in reordering are different from those in translation. We conclude that there may be good uses for reordering questions in the beginning level required to acquire a basic knowledge of word order, but its effectiveness for intermediate students is limited, and reordering exercises do not follow that they can be expected to correctly use verbs, pronouns, prepositions and articles. With the limitation of the exercise of only reordering for intermediate-level learners, it is essential to increase opportunities for them to tackle Japanese-English translation and paragraph writing and be made explicitly conscious of the word classes whose errors are frequently observed.

On the other hand, instruction in controlled composition has the advantages of clarifying the important points and, most of all, reducing the workload of teachers. In instruction focusing on outputs, it is important to take into consideration the need to efficiently teach learners. It seems beneficial to both learners and teachers to give learners such kinds of writing instruction as Kimura *et al.* [4] suggested.

It is unrealistic to drastically decrease the practices in formal school education, because it is part of traditional writing instruction. However, teachers should reconsider how to use reordering in view of the fact that the course guidelines in Japan are trying to increase opportunities for learners to express themselves in English.

*We appreciate Prof. Neil Heffernan of Kurume University for proofreading the first version and Prof. Susan Pavloska of Doshisha University for checking the final version. Special thanks are due to the three anonymous reviewers and the editors of the EUROPHRAS 2017 Programme Committee for their comments and encouragement. All remaining errors and inadequacies are, of course, our own. A part of this paper is supported by the research grant from the Graduate School of Culture and Information Science, Doshisha University.

References

1. Ellis, R.: Understanding Second Language Acquisition. Oxford University Press, Oxford (1985).
2. Ishiguro, T., Yamauchi, N., Kitabayshi, T., Akaso, N.: ENGLISH COMPOSITION AT WORK: Hyougen no tameno Hasshingata Eisakubun [Productive English Composition for Expression]. Kinseido, Tokyo (1994).
3. James, C.: Errors in Language Learning and Use: Exploring Error Analysis. Longman, Harlow (1998).
4. Kimura, H., Kimura, T., Shiki, O.: Riidingu to Raitingu no Riron to Jissen [Theory and Principle in Reading and Writing: Nurturing Independent Learning]. Taishukan, Tokyo (2010).
5. Kobayashi, I., Hayasaki, Y.: Maakushiki Kiso Mondaishu (5) Eigo [Goku Seijo] Goteiban [Mark-sensing Fundamental Workbook: English Reordering Questions, 5th Edition]. Kawai-publishing, Tokyo (2005).
6. Morishita, M., Yamamoto, T.: How Syntactic Processing Training Affects Oral Production of Elementary Level of Japanese ESL Learners. *Linguistic Research*, 30(3), 435-452 (2013).
7. Sase, F.: Supiikingu Nouryoku to Gojun Chishiki no Kanrensei ni okeru Chosa [Survey on the Relationship between Speaking Abilities and Knowledge about Word Order]. *EIKEN BULLETIN*, 26, 31-49 (2014).
8. Tono, Y., Mochizuki, H.: Toward Automatic Error Identification in Learner Corpora: A DP Matching Approach. Paper presented at CL2009, UK, Liverpool (2009).
9. Yamauchi, N., Akaso, N., Kitabayashi, T.: Bunpou kara Semeru Eisakubun no tameno 15 Shou [15 Chapters for English Composition Focusing on Grammar]. Eihosha, Tokyo (2005).

How Does Data Driven Learning Affect the Production of Multi-Word Sequences in EAP Students' Academic Writing?

Melissa Larsen-Walker ¹

¹ University of South Florida, Tampa, FL 33620, USA
mlarsenw@mail.usf.edu

Abstract. Multi-word sequences (MWSs) have been found to occur with high frequency in academic writing [2, 25]. MWSs are recurrent expressions, which a writer retrieves from his/her long-term memory in order to construct utterances. In written discourse, such MWSs serve to refer the reader to previous research, organize the sections of texts and discourse within them and position the writer as knowledgeable [2, 3]. Previous research suggests that L2 writers frequently misuse these forms, resulting in disfluent written discourse [8, 20, 23]. Nevertheless, Hyland [14] suggests that use of appropriate and sophisticated MWSs helps to establish the writer as a member of an academic discourse community. The current, corpus-based and quasi-experimental study investigates the effectiveness of using Data Driven Learning (DDL) in conjunction with teaching MWSs. Key MWSs have been selected from Simpson-Vlach and Ellis' [25] Academic Formulas List (AFL), specifically from the Referential Function and two sub-categories of the Stance Function, Hedges and Epistemic Stance. The researcher used an objective pretest-posttest to ascertain how DDL affects students' receptive knowledge of AFL-MWSs and used the first and final drafts of an argumentative essay to assess students' ability to produce them. A statistically significant difference between pretest and posttest scores for the treatment group supports the assertion that DDL positively impacts students' receptive knowledge of AFL-MWSs. Discussion includes comparison between students' self-generated inductions regarding each AFL-MWSs and how they used them within their essays.

Keywords: Data Driven Learning, Multiword Sequences, Academic Formulae.

1 Introduction

L2 writers, endeavoring to craft appropriate essays and research papers, must conform to the academic expectations of colleges and universities in the U.S. Despite many strengths, one aspect of L2 undergraduate students' writing that appears to be a weakness is their use of multi-word sequences (MWSs). Research suggests that in many cases, they use more MWSs than native speakers, whether novice or expert [13, 19-20, 22, 24]. However, they tend to use MWSs that are considered to be too informal

and conversational for the essay task. Other aspects of writing, such as lucid content, vocabulary choice and organization may be relatively more important to the overall effectiveness of a student's written composition. Nevertheless, specific academically-appropriate MWSs influence the quality of the writing in terms of cohesiveness and positioning of the writer as an expert on the topic [14]. Such MWSs frequently occur in the writing of published academic authors [3, 21-22, 27]. Use of these MWSs has also been associated with higher grades on the Common European Framework of Reference [18]. In spite of numerous studies that document the importance of MWSs, no consensus exists as to how best to teach these formulas.

The current research investigates the possible effectiveness of using Data Driven Learning as a means to teach MWSs, specifically MWSs selected from Simpson-Vlach and Ellis' Academic Formulas List (AFL) which is described below. Data Driven Learning (DDL) is an inductive approach wherein learners search a corpus or corpora under instructor guidance, in hopes of finding a pattern in the data. It has been compared to discovery learning in that the instructor guides students and acquaints them with how to use the material, i.e. learner website and corpus, but refrains from pre-teaching the rule or generalization regarding the target form [15-16]. Based on previous research, the researcher infers that DDL could be a viable alternative to traditional direct instruction. Despite ample documentation for the importance of using appropriate MWSs to academic writing, most studies on MWSs do not contain any instructional component [2, 7, 11, 22, 24]. Even rarer are classroom based studies on MWSs that include DDL. The current research aims to address this gap.

Awareness of the tendency of L2 students to select inappropriate MWSs motivated Simpson-Vlach and Ellis [25] to compile a list of MWSs or *academic formulas* that frequently occur in academic writing and would be valuable pedagogically. The corpora used were: Hyland corpus of research articles; (spoken) academic portion of Michigan Corpus of Academic Spoken English (MICASE), British National Corpus (BNC) spoken academic corpus; and Lee's genre category of the BNC. For comparison, they consulted the FROWN corpus (informal American English speech). Simpson-Vlach and Ellis used multiple measures, qualitative and quantitative methods, combining information from log likelihood, mutual information score, and formula teaching worth (FTW). First, they measured simple frequency in the target and comparison corpora. They were concerned about the fact that lists of MWSs which have been compiled based solely on frequency, such as that composed by Biber, Johansson, Leech, Conrad and Finegan [4], tend to contain MWSs which from a pedagogical standpoint are unimportant [19-20, 26]. For example, *one of the* was found to be the most frequently occurring lexical bundle in the initial search, yet these individual words were extremely common throughout both the expert and comparison corpora. In order to correct for this tendency, without going to the other extreme of selecting MWS solely on the researchers' intuitions, Simpson-Vlach and Ellis used an alternative way to analyze MWSs, the mutual information (MI) statistic. The MI statistic, which originated in the field of information science, rates the likelihood that two (or more) words did not occur together by chance. As such, relatively low frequency collocations such as *blue moon* have a high MI score.

The current study applies the formulas from the AFL, i.e. AFL-MWSs to the community college, EAP setting, testing whether DDL can be effective in teaching these AFL-MWSs to L2 English under-graduates. Specifically, this study explores the following questions:

Research Question #1: To what extent does Data Driven Learning affect receptive knowledge of academically appropriate MWSs, operationalized as key formulas from Simpson-Vlach and Ellis' *Academic Formulas List* (AFL). Receptive knowledge would be measured by an objective test of MWSs (i.e. academic formulas).

Research Question #2: To what extent does DDL affect productive knowledge of academically appropriate MWSs? Productive knowledge would be measured by accurate usage of academic formulas within the argumentative/persuasive essays.

2 Literature Review

2.1 MWS Studies Including an Instructional Intervention

Cortes [8] endeavored to teach MWSs to L1 undergraduates in a history course. Direct instruction including presentation of the formulas was followed by practice activities throughout her five visits to the classroom. Results confirm the null hypothesis; i.e. the students did not produce more MWSs in their assigned writing after the Cortes' instruction. The small number of participants (n=12) may have influenced the results.

2.2 Review of Empirical Studies on DDL

Linking adverbial studies. Numerous studies have investigated the effectiveness of DDL in teaching linking adverbials (LAs), a construct related to MWS [5, 9-10, 12]. Linking adverbials connect sections of text, either within a sentence or across longer sections of text. These transitional devices may be single word (however) or multi-word (on the other hand). Boulton [5] explored paper-based DDL as a means to elicit appropriate use of LAs among L1 French undergraduates, writing in English. Results suggest that DDL was effective despite the fact that the participants were intermediate level. Garner [12] used both the COCA and the MICUSP to investigate DDL and LAs. His quasi-experimental study aimed to elicit use of academically appropriate MWSs and to elicit correct use of LAs. The measure of the treatment's effectiveness were the pretest and posttest essays written by the participants. Garner found a statistical difference between correct use of LAs in pretest and posttest essays, from which he infers the effectiveness of the treatment. Cotos [9] aimed to find out whether the inclusion of a learner corpus would impact the effectiveness of DDL in teaching LAs to L2 English undergraduates. She instructed two classrooms via DDL, but only the LLD group used a learner corpus of essays written by unidentified course students. All participants used a native speaker corpus comprised of articles from his/her discipline. Her instruments were both an objective LAs test and essays written pre- and

post-treatment. The NSC and LDD groups showed a statistically significant difference between the pre- and post-test scores on LAs recognition. As such, Cotos claims the DDL intervention including a learner corpus was effective, and the effects of instruction persisted beyond the course limits.

Tangible learning gains were made in studies wherein DDL was used to teach vocabulary and/or collocation of prepositions. Koosha and Jafarpur [17] study aimed to investigate the effectiveness of using DDL to teach collocation of prepositions to L1 Farsi undergraduates. Their quasi-experimental included 200 participants at two university campuses in Iran. Participants in the experimental group used the Brown corpus under instructor guidance, while the control group used the Longman Dictionary online. Koosha and Jarfarpur used an objective test of target collocations for their pretest-posttest. Results suggest that DDL was effective in eliciting knowledge of correct collocation of prepositions, based on both a statistical difference between pretest and posttest scores for the experimental DDL group and a difference between the experimental and control groups. Using a similar research design, Celik [6] investigated the effectiveness of using DDL to teach collocation of prepositions to L2 medical students in Ankara, Turkey. His results were less impressive than Koosha and Jafarpur's insofar as the difference between control and treatment groups on the immediate posttest. Nevertheless, on the delayed posttest there was a statistical difference between the mean scores for the DDL and comparison groups, with the former achieving the higher mean score. These vocabulary studies appeared to be effective although the outcome measure was an objective test rather than a measurement of students' use of MWSs within pretest and posttest writing assignments.

In a study by Park [21], the researcher explored the effects of corpus consultation on errors in essays and a microanalysis of the students' revision process. Analysis of results shows corpus consultation resulted in a correct revision in 56% of cases.

3 Methods

The design of the study is quasi-experimental and will include four classrooms across the course of three semesters. This short paper addresses the first phase of data collection, which included on treatment group only. The instructional sequence lasted two weeks, including the time the administration of the objective pretest and posttests. All instructional materials were included as part of an instructor designed website. <http://mws-afl-eap-bluelotusfeet.sitey.me/> As part of the process of guiding students to ascertain a generalization for how the phrase is used, the instructor modeled inductive reasoning, using the phrase, *claim that*. After this modeling, the students continued to work on induction independently, completing the second task associated with the website, which required them to induce a rule for how each of the key AFL-MWSs (shown in Table 1) should be used. To address RQ#1, the researcher designed and administered an objective test, described below, to measure knowledge of the target MWSs. To address RQ#2, the researcher analyzed the first and second drafts of

an argumentative essay, with a researched component, the first drafts of which each student wrote prior to treatment. Participants revised and re-submitted them as final drafts after the DDL treatment.

3.1 Target Formulas and Instructional Sequence

The target formulas in this study are MWSs from the AFL, the most frequently occurring from within the Referential functional category, such as *the fact that*. The rationale for this is the fact that previous studies have shown that academic writing by experts contains relatively more referential MWSs [2, 7, 19]. Along with referential, the proposed study will include two of the subcategories of stance, hedges, such as *are likely to be* and epistemic stance, such as *we assume that*. Hedges qualify or mitigate the writer's assertion, leaving them open to the reader's attempts to falsify such claims [25]. Use of MWSs from *AFL* contributes to the coherence of the text, positioning the writer as one who builds upon the scholarly research of others while qualifying the claims for his/her own possible contribution. For that reason, the researcher will provide instruction via the website in the Hedging subcategory of the Stance function prior to students' independent searches in the corpus. The table below displays the specific MWSs for which the participants will seek, classified within stance or referential functional categories.

Table 1. Target MWSs from the Academic Formulas List, Including Functional Category.

Major Functional Category	Sub-Category	Formula
Referential	Intangible Framing Attributes	<i>the fact that</i>
	Quantity Specification	<i>in some cases</i> <i>a series of</i>
	Contrast and Comparison	<i>on the other hand</i>
Stance	Hedges	<i>are likely to</i> <i>to some extent</i>
	Epistemic Stance	<i>we assume that</i>

3.2 Data Analysis

The researcher will tally the pretest scores on the receptive test (RQ#1) for the comparison group and the treatment group classrooms. The score will be based on a possible 35 points, one point for each correct response. After calculating the mean for each group, the researcher would subject these to an independent samples T-test. This statistic will enable the researcher to infer whether the two groups are equivalent prior to instruction. After treatment, students will take the posttest (i.e. retake the same test), with the items appearing in a different order. A paired T-test will be used to find

out if there is improvement in the students' receptive knowledge of academic formulas as evidenced by their posttest scores.

The researcher investigated RQ2 by coding the MWSs used by each student in his/her essay, first by classifying them according to the functional categories in the AFL and then assessing them for appropriateness.

4 Results

The results of the objective test suggest an improvement in the students' receptive knowledge of AFL-MWS. A paired T-Test was used to analyze the difference between pretest and posttest scores for the experimental group. With the p value set to $p = 0.05$, the obtained T value ($t = 5.299$) was found to be higher than the critical T ($t = 2.26$). Based on this analysis, the researcher rejects the null hypothesis. From this result, the researcher cautiously infers that DDL positively impacted students' receptive knowledge of AFL-MWSs.

Table 2 below categorizes the types of MWSs used in students' posttest essays by functional category. While the change in the number of AFL-MWSs between pre-and post-test essays was insignificant, the micro-analysis of the formulas used and the induction process provides ample opportunity for comparison. Additionally, misuse of the formulas was not observed. It appears that except for *the fact that* and *in some cases*, the students were unfamiliar with these phrases prior to treatment. The appropriacy of target AFL-MWS use within the essays and the inclusion of other relevant academic phrases such as *claim that* may suggest non-quantifiable learning gains.

Table 2. Usage of AFL-MWSs I Pretest and Posttest Essays Sorted by Functional Category

Category	Formula	Pretest Essay	Posttest Essay
Referential – Intangible Framing Attributes	<i>the fact that</i>	3	4
Referential – Quantity Specification	<i>in some cases</i>	2	3
	<i>a series of</i>	0	3
Referential – Contrast and Comparison	<i>on the other hand</i>	0	2
Stance – Hedges	<i>are likely to</i>	0	1
	<i>to some extent</i>	0	4
Stance – Epistemic Stance	<i>we assume that</i>	0	3

5 Discussion and Conclusion

Analysis of the essays revealed that while students used AFL-MWSs sparingly, they used them accurately. Detailed qualitative analysis of MWSs within student essays reveals a connection between the student's induction and how s/he used the AFL-

MWSs. In the examples shown in Table 3, the student’s induction, based on task 2 from the website, was entirely consistent with how s/he employed it in the essay. The greatest uptake was on the noun phrase MWS *the fact that*.

Writing by each of the participants seemed to follow the self-discovered rule for that particular AFL-MWS. If we view the essay excerpts beside the induction, we can see how closely they align, especially for participants 23, 29, 34. The research cited by participant 28 suggest “high probability” of the youth skipping classes and having low self-esteem. Participant 29 used induction to ascertain that the fact that begins a Noun Clause and introduces a significant point, both of which were reflected in the essay. The study by Common Sense Media, cited by participant 34, contradicts the argument that social media negatively impact youth, at least according the teens themselves. Some of their classmates used not only a few of the AFL-MWSs above but also academic formulas from the AFL which were not included on the instructional website. The appearance of non-target formulas, such as according to and as a result, in the students’ essays may be attributable in part to their textbook, *Final Draft 4* [1]. Every chapter had a section on academic formulas and included in these were the two academic phrases above.

Previous research suggests that students struggle with induction [5, 27]. However, it appears that if induction is modeled and practice activities reinforce it, then students can induce a rule for how an MWS should be used, as evidenced by their objective test scores. In some cases, students take the next step, applying the induction to their academic writing. If researchers conduct future studies that have a greater number of participants and use both treatment and control groups, they should be able to confirm or disconfirm the conclusions drawn from the small exploratory study.

Table 3. Comparing Student Inductions with Excerpts from Essay.

Induction	Short Context from his/her Essay
Part. 29 <the fact that> is the beginning of the noun clause. Most of the time it comes before the pronoun or noun. The author uses the MWS <i>the fact that</i> to put a significant point on the case.	Today over 60% of young people between 13-18 years old have a profile on social media. <i>The fact that</i> children spend too much time on social media make them vulnerable to predatory adults.
Participant 28 <are likely to be> after a noun, and before an infinitive verb. This phrase means having a high probability of occurring or being true.	The 2014–2015 School Crime Supplement indicates that, nationwide, about 21% of students ages 12-18 experienced bullying. When kids are cyberbullied, they <i>are likely to</i> have lower self-esteem, be unwilling to attend school.
Part. 33 <to some extent> is usually used as an object or a preposition phrase. A writer uses the MWS, <i>to some extent</i> which means partly.	Therefore, they do not have enough time to do other important tasks, such as studying, exercising, real communicating. In the long run, it has a negative effect on the physical development as well as the spirit of youth <i>to some extent</i> .

References

1. Asplin, W., Jacobe, M. & Kennedy, A. Final Draft 4. Cambridge University Press, New York. (2016).
2. Biber, D. A corpus driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3), 275-311 (2009).
3. Biber, D., Conrad, S., & Cortes, V. 'If you look at . . . ": Lexical bundles in university teaching and textbooks,' *Applied Linguistics*, 25, 371-405 (2004).
4. Biber, D., Johansson, G., Leech, Conrad, S. & Finegan, E. *Longman Grammar of Spoken and Written English*. Longman, Harlow, England. (1999).
5. Boulton, A. Testing the limits of data-driven learning: Language proficiency and training. *ReCALL*, 21(01), 37-54 (2009).
6. Celik, S. Developing collocational competence through web-based concordance activities. *Novitas Royal Research on Youth and Language*, 5(2), 273-286 (2011).
7. Chen, Y. & Baker, P. Lexical bundles in L1 and L2 academic contexts. *Language Learning and Technology*, 14(2), 30-49 (2010).
8. Cortes, V. Teaching lexical bundles in the disciplines: Examples from a history intensive writing class. *Linguistics and Education*, 17(4), 391-406 (2006). DOI: 10.1016/j.linged.2007.02.001
9. Cotos, E. Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26, Special Issue 02, 202 – 224 (2014). DOI: 10.1017/S0958344014000019.
10. Cresswell, A. Getting to 'know' connectors? Evaluating data-driven learning in a writing skills course. *Language and Computers*, 61(1), 267-287. (2007).
11. Fan, M. An exploratory study of collocational use by EFL students: A task-based approach. *System*, 37, 110-133 (2009).
12. Garner, J. The use of linking adverbials in academic essays by non-native writers: How data-driven learning can help. *CALICO Journal*, 30(3), 410-422 (2013) doi: 10.11139/cj.30.3.410-42.
13. Granger, S. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A.P. Cowrie (Ed.) *Phraseology: Theory, analysis, applications*. Oxford University Press, Oxford, UK. (1998).
14. Hyland, K. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27: 4-21. (2008).
15. Johns, T. Should you be persuaded: Two samples of data-driven learning materials. *ELR Journal*, 4, 1-16. (1991).
16. Johns, T. 'Whence and whither classroom concordancing?' in Bongaerts, Theo et al (eds.) *Computer applications in language learning*. Foris, Dordrecht, Holland (1988).
17. Koosha, M. & Jafarpour, A.A. Data-driven learning and teaching collocation of prepositions: the case of Iranian EFL adult learners. *Asian EFL Journal Quarterly* 8(4) 192-209 (2006).
18. Mukherjee, J. & Rohrbach, J. M. Rethinking applied corpus linguistics from a language pedagogical perspective: New departures in learner corpus research. *Applied Corpus Linguistics and Learner Corpus Research*, 205-232 (2006).
19. O'Donnell, M., Romer, U. & Ellis, N.C. The development of formulaic sequences in first and second language writing: Investigating effect of frequency, association and native norm. *International Journal of Corpus Linguistics*, 18(1), 83-108 (2013).

20. Paquot, M. & Granger, S. Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130-149 (2012).
21. Park, K. Learner-corpus interaction: A locus of microgenesis in corpus-assisted L2 writing. *Applied Linguistics (Advanced Access)*, 1-26 (2012). doi:10.1093/applin/ams012
22. Peromingo, J. Corpus analysis and phraseology: Transfer of multi-word units. *Linguistics and Human Sciences*, 6, 321-343 (2012).
23. Phoocharoensil, S. Exploring learners' developing L2 competence. *Theory and practice in language studies*, 4(12), 2533-2540. (2014). DOI: <http://dx.doi.org/10.4304/tpls.4>.
24. Ping, P. A study on the use of four-word lexical bundles in argumentative essays by Chinese English: A comparative study based on WECCL and LOCHNESS. *CELEA Journal*, 32(3), 25-45 (2009).
25. Simpson-Vlach, R. & Ellis, N.C. An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487-512 (2010).
26. Thomas, J. Stealing a march on collocations: Deriving extended collocations from full text for student analysis and synthesis. In A. Lenko-Szymanska and A. Boulton (eds.) *Multiple affordances of language corpora for DDL*, (pp. 85-108). Philadelphia, John Benjamins (2015).
27. Yeh, Y., Liou, H., & Li, Y. Online synonym materials and concordancing for EFL college writing. *CALL*, 20(2), 131-152. (2007)

Phraseology in Teaching and Learning Spanish as a Foreign Language in the USA

Victoria Llongo Lopez

¹ University of Alicante, crta. San Vicente del Rapeig, s/n, 03690, Alicante, Spain

Brigham Young University, 84602, Provo, Utah, United States of America
victoriallongo@gmail.com

Abstract. The aim of this study is to analyze the acquisition of phraseological units at the elementary, secondary, and university levels in order to develop a better understanding of the linguistic competence of American students in learning Spanish as a foreign language in Utah. After discussing the importance of phraseology as a fundamental part of the linguistic content in the L2 classroom, we analyze the most frequent phraseological units in L2 Spanish manuals and how students in the different academic contexts were able to identify their meaning in contextualized and non-contextualized settings. The results showed the importance of context in the identification of these phraseological units with a significantly higher percentage of accuracy in the contextualized setting. The study also addresses the important role that idiomatic expressions can play in teaching phraseology and culture thus enriching the sociolinguistic competence with diatopic variation as a core of phraseology.

Keywords: Phraseology, Didactics, Diatopic variation, Spanish, Teaching.

1. Didactics and Phraseology in Learning a Foreign Language

The large number of studies focused on different aspects of language teaching have also touched one of the most relatively new disciplines in Linguistics called Phraseology and this has become an independent field of study, known as Phraseodidactics. Thus, within the evolution of Phraseology research this branch was considered a loophole. Nowadays, this part of the didactics of languages has been of the utmost interest for European researchers [8]. Most of them are focused on new didactic materials, different approaches, and teaching methods in order to recognize these exclusive units and their recognition by also paying particular attention to learning the meaning and linguistic communication.

Native speakers acquire phraseology as they do with vocabulary, generally by “repeating, memorizing sequences and meanings of different structures” (Forment-Fernández, 1998, 341). However, the Common European Framework of Reference

for Languages reinforces not only the comprehension but also the use of most frequent familiar and everyday expressions for beginner students (A1 level). “Consequently, Phraseological units must be an integral and specific part of Teaching Plans for any foreign language (FL)” (Soslano Rodríguez, 2006).

“There is not just one tried and true ‘method’ for teaching phraseology” (Kennedy, 2008). There are some tentatives of including distributional information from a corpus in a language learner’s curriculum. This study has been divided in three different levels in education: elementary, secondary and college.

The Phraseological units selected are the ones from *Repase y Escribe: curso avanzado de gramática y composición*. (Maria Cantelli Dominicis. John J. Reynolds). This survey was taken by 80 elementary school students, 47 high school students and 138 university students.

This survey was divided in two parts. The first part was composed by 10 questions with phraseological units isolated where students had to choose their meaning, explain why they chose this option and say how often they heard it. The second part of the survey included 6 questions with phraseological units in context, in this case, students had to choose their meaning and the frequency of usage.

This survey took 20 minutes overall. It was presented to the students through Google Forms. In this case, we selected for ourselves the Phraseological units by importance on the manual and their inclusion in a lexical activity as parameters of frequency did not apply. These can lead to overlooking the kinds of semantic relationships associated with phraseology. As Wray & Fitzpatrick wrote ‘a learner-directed phraseology curriculum could be ideal, if it could relate the items to be learned and what the learner might be motivated to say or write’.

As Kisner (1994) and Ellis (1994) have argued, it seems that phraseology is learnt especially through implicit learning by unconsciously meeting multiword sequences repeatedly in context. The more we encounter these multiword units, the more fluent we become in retrieving and producing them (Byte & Hopper 2001). It is almost certainly worthwhile to use explicit instruction in teaching phraseology as vocabulary. This is the most common case scenario.

2. Phraseological unit case study

2.1. Elementary results

This survey was taken by 80 students from a Spanish Dual Immersion program from Second Grade and Fifth Grade (6-8 and 10-12 years old range). All of them are native English speakers, as an anecdote, some of them recognize Spanish as their native language for being in a Dual Immersion program and speak the L2 language all the time at school. Some of them recognized speaking to a limited extent in a foreign language at home and reading books. Most of them were studying Spanish for 1-2

years and more than 4 years. None lived in a Foreign country where the language was spoken.

According to the survey results, the phraseological units (without context) with the higher percentage of accuracy are: “Amasar una fortuna” (51.2%). “Dar la cara por alguien” (48.8%). “Poner los puntos sobre las íes” (46.3%).

The idiomatic expressions with less accuracy are: “Irse a la francesa” (20%). “Montar cachos” (31.3%). “A caballo regalado, no se le mira el diente” (32.5%).

The expressions in context with higher percentage of accuracy are: “Trabajar a destajo” (65%). “Ponerse blando” (58.8%). “Andarse con rodeos” (51.2%).

The idiomatic expressions in context with less accuracy are: “Casa que se blanquea, inquilinos quiere” (33.8%). “Ojos que no ven, corazón que no siente” (36.3%). “Perro que ladra, nunca muerde” (42.5%).

When we analyzed the results of the survey, we saw the idiomatic expressions such as “amasar una fortuna” had an equivalent in English “to amass a fortune”. In this case, equivalences will be one of the most appropriate phraseological units to start with teaching in elementary schools. Followed by units like “dar la cara por alguien” or “poner los puntos sobre las íes” because they can be easily exemplified in a school context and reinforced with an image.

In this context, cultural aspects play an important role and students in elementary school know that in some European countries you greet somebody with two kisses. In this case, students identify the fact of saying ‘goodbye as a French’ by kissing on the cheek. It is also known that introducing vocabulary and expressions in context will help the student to understand and acquire any sequence.

2.2. Secondary results

This survey was taken by 47 students from different levels in high school. Most of the students (83%) belong to 15-18 age range and only 14.9% of the age range between 18-21. The rest 2.1% from 12-15 years old. In this case, we counted on more Hispanic students than elementary and college, exactly 59.6% of the students. Overall, students did not take Spanish in elementary, but they did in middle school and high school.

Most of them took 2 to 3 classes of Spanish during their education.

According to the results of the survey, secondary students achieve a higher percentage in the following phraseological units: “Dar la cara por alguien” (74.5%). “A caballo regalado, no le mires el diente” (66%). “No decir ni jota” and “Amasar una fortuna” (63.8%).

The idiomatic expressions with a lower percentage of accuracy are: “Estar en el bote” (14.9%). “Hacerle la cruz a alguien” (23.4%). “Irse a la francesa” (42.6%).

The expressions in context with higher percentage of accuracy are: “Casa que se blanquea, inquilinos quiere” (63.8%). “Trabajar a destajo” (61.7%). “Perro que ladra, nunca muerde” (57.4%).

The expressions in context with lower percentage of accuracy are: “Ojos que no ven, corazón que no siente” (34%). “Ponerse blando” (48.9%). “Andarse con rodeos” (55.3%).

The results in high school were very interesting because there were more hispanic students. Equivalents still played an important role in unit recognition and common sequences. Nowadays in Spanish as “dar la cara for alguien” or “a caballo regalado, no le mires el diente” will be the ones we want to focus on teaching in our classrooms due to their frequency and popularity.

Some expressions as ‘estar en el bote’, ‘hacerle la cruz a alguien’ or ‘irse a la francesa’ are easy to perform in class or reinforce with a picture.

2.3. University results

In college, this survey was taken by 138 students at BYU. The two highest age ranges were students from 18-21 years old (43.4%) and students aged from 25-30 years old (47.8%). In this case, a minority of approximately 8% of the students are from a Spanish speaking country.

That means that a 73.2% of the students speak only English at home. The majority did not take Spanish in elementary school. Only a 32.4% had from 1-2 years in middle school and a 37% took 3-4 years at University. At least all of these interviewed had taken one class. As cultural data, most of the students serve their missions in Spanish speaking countries, which means that they lived there at least for a couple of months in contact with the language and culture.

Phraseological expressions (without context) with higher percentage in college students are: “No decir ni jota” (92%). “A caballo regalado, no le mires el diente” (87%). “Poner los puntos sobre las íes” (82.6%).

Units (without context) with less percentage of accuracy are: “Hacerle la cruz a alguien” (26.1%). “Estar en el bote” (37%). “Montar cachos” or “Irse a la francesa” (56.5%).

Idiomatic expressions in context with a higher percentage of accuracy are: “Trabajar a destajo” (92.7%). “Perro que ladra, nunca muerde” (90.6%). “Ponerse blando” (89.8%).

Phraseological expressions in context with a lower percentage of accuracy are: “Casa que se blanquea, inquilinos quiere” (79.7%). “Ojos que no ven, corazón que no siente” (83.3%). “Andarse con rodeos” (88.3%).

In college, we can see an evolution in comprehension of the Phraseological units. The most accurate units are the hardest ones in other levels in education. Though simple

units, much more representative and easy to perform as ‘hacerle la cruz a alguien’, “estar en el bote” or “montar cachos” seem to be much more difficult to recognize.

3. Diatopic variation

This case study show us much more than how to teach idiomatic expressions, we can also take a step further to research the diatopic variation. The results of the survey show us that one of the hardest idiomatic expressions for students to recognize is ‘montar cachos’, which is an expression very used in Mexico. Also, the expression from the Spanish peninsula which is “irse a la francesa”. Or, easily to recognize for Hispanic students in high school the South American and Central American expression ‘casa que se blanquea, inquilinos quiere’.

Phraseology and culture are completely related. This should be an advantage for educators and learners. When assessing communication skills, the recognition and used of these native-speaker units are automatically considered as an indicator of fluency in foreign language learners. Phraseology research has specially challenged language educators to work out to maximize the exposure needed for learners to acquire phraseological units that cannot be taught explicitly. Thus, encourage students to autonomous language learning through watching movies, reading, listening to the radio, music, online, every possible way to maximize the exposure to language in use.

PU	Answer 1	%	Answer 2	%	Answer 3	%
'Poner los puntos sobre las íes'	Poner los puntos en todas	38.8%	Poner las cosas claras o	46.3%	Dejar de hacer algo y	15%
	las íes del texto.	42.6%	concretar alguna	46.8%	empezar a conversar sobre	12.8%
		10.9%	cuestión.	82.6%	un tema.	8.7%
'No decir ni jota'		33.8%		28.7%		37.5%
	No decir la verdad.	25.5%	Hablar demasiado.	10.6%	Permanecer en silencio.	63.8%
		5.8%		2.2%		92%
'De pe a pa'		42.5%	Saber algo de memoria	33.8%	Saber cosas básicas y	23.8%
	Saber algo desde el	57.4%	sin entenderlo.	17%	esenciales.	25.5%
	principio hasta el final.	78.3%		5.1%		16.7%
'Irse a la francesa'		40%		20%		40%
	Irse de un lugar y darle un	23.4%	Irse de un lugar sin	42.6%	Irse de un lugar y decir un	34%
	beso a todo el mundo.	21.7%	despedirse.	56.5%	adiós generalizado.	21.7%





'A caballo regalado no se mira el diente'	No me gusta el regalo que me han dado pero no me puedo quejar.	32.5% 66% 87%	Me gusta el regalo que me han dado tanto como a los caballos	41.3% 21.3% 4.3%	El regalo es tan feo como un caballo con dientes amarillentos.	28.2% 12.8% 8.7%
'Hacerle la cruz a alguien'	Bendecir a alguien.	26.3% 51.1% 54.3%	Marcar a alguien	30% 25.5% 19.6%	Odiar a alguien	43.8% 23.4% 26.1%
'Estar en el bote'	Estar a punto de obtener, convencer o conquistar algo o a alguien	40% 14.9% 37%	Estar a punto de atrapar algo o a alguien.	27.5% 31.9% 15.9%	Estar encerrado en algún lugar sin acceso al exterior.	32.5% 53.2% 47.1%
'Montar cachos'	Ponerse a llorar desesperadamente.	30% 19.1% 16.7%	Estar destrozado, con el corazón partido.	38.8% 27.7% 26.8%	Engañar o serle infiel a alguien.	31.3% 53.2% 56.5%
'Amasar una fortuna'	Contar mucho dinero.	23.8% 17% 10.9%	Producir mucho dinero.	51.2% 63.8% 74.6%	Usar el dinero que no es tuyo para tu propio beneficio.	25% 19.1% 14.5%
'Dar la cara por alguien'	Hacer el trabajo de otra persona.	33.8% 8.5% 9.4%	No tener vergüenza y obtener todo gratis.	17.5% 17% 11.6%	Defender o responder por alguien.	48.8% 74.5% 79%

Table 1. Table of idiomatic expressions without context, % of accuracy.

PU	Answer 1	%	Answer 2	%	Answer 3	%
Cuanto más lo pienso, más me enojo. Si vuestra relación está basada en la distancia, "ojos que no ven, corazón que no siente", ni te preocupes por lo que pase allá donde sea.						
	Me duele el corazón y los ojos de sufrimiento.	31.3% 25.5% 8.7%	No me creo lo que ven mis ojos, es mentira.	40% 44.7% 8.7%	Mejor ser desconocedor de algo para no sufrir.	36.3% 34% 83.3%
No me puedo creer que estén sacando todos esos casos en el Gobierno, al final todos los políticos acabarán en la cárcel. Sí, "casa que se blanquea, inquilinos quiere".	Todos los políticos han robado y como el dinero llama al dinero quieren más.	33.8% 63.8% 79.7%	Todos los políticos buscan nuevos inquilinos para sus casas.		Todos los políticos están en la cárcel por blanqueo de dinero y alquiler de casas.	31.3% 8.5% 9.4%
				38.8% 34% 11.6%		

Ernesto se puso muy furioso durante la reunión, empezó a discutir con todos sus compañeros de trabajo. Imagino que, "perro que ladra, nunca muerde" con lo cual podemos estar tranquilos.						
Ernesto se puso a	26.3%	Ernesto habla mucho,	42.5%	Ernesto nunca	36.3%	
ladrar y a morder	23.4%	discute mucho, pero	57.4%	hace nada malo, es	19.1%	
a todos.	2.9%	nunca hace nada.	90.6%	buena persona.	7.2%	
Miriam "no se anda con rodeos" si tiene que decirte que algo va mal, te lo diré. Me encanta su personalidad.						
Miriam no es muy	26.3%	Miriam es muy fuerte y	28.7%	Miriam es muy	51.2%	
sincera, le gusta	25.5%	rápida, siempre anda	21.3%	clara y sincera.	55.3%	
inventarse	5.8%	por los rodeos.	7.3%		88.3%	
historias.						
Mi padre siempre llega a casa el último, el pobre hombre, 'trabaja a destajo' por su familia. Le quiero un montón.						
Mi padre trabaja	65%	Mi padre no trabaja	18.8%	Mi padre ni trabaja	21.3%	
muchísimo y se	61.7%	mucho y llega a casa	17%	ni llega a casa a	21.3%	
sacrifica por su	92.7%	tarde.	4.4%	tiempo para estar	3.6%	
familia.				con su familia.		
Cuando David vio la aguja que el doctor tenía en la mano, "se puso blando" y le entró una sensación de hormigueo en las piernas.						
A David le gusta	20%	A David no le gusta ir	26.3%	A David no le		
ir al médico y	12.8%	al médico, no obstante,	40.4%	gusta ir al médico,	58.8%	
soporta cualquier	3.6%	es un valiente.	9.5%	se pone nervioso,	48.9%	
situación.				se mareo y	89.8%	
				palidece cada vez		
				que acude.		

Table 2. Table of idiomatic expressions in context, % of accuracy.

	University %		Elementary %
	High school %		Correct answer

4. Conclusion

These results belong to an ambitious postgraduate research where more than one manual was analyzed. The student progress in learning Phraseological units in Spanish L2 was tracked and analyzed. Also, this study shows the importance of context when teaching these idiomatic expressions and the importance of helping students develop habits for forming educated guesses. Burke (1998) claims that “knowledge of slang and idiomatic expressions is fundamental to nonnative speakers' understanding of the language that native speakers actually use” (p.5).

The study also addresses the important role that idiomatic expressions can play in teaching a foreign language and culture, thus enriching the sociolinguistic competence with diatopic variation as a core of phraseology.

5. References

1. Bybee, J. 1998. The evolution of grammar. Paper prepared for the symposium Darwinian perspectives on the origins of language. AAAS, Philadelphia.
2. Ellis, N.C., & Sinclair, S. (in press). Working Memory in the Acquisition of Vocabulary and Syntax: Putting Language in Good Order. *Quarterly Journal of Experimental Psychology*. Special Issue on Working Memory.
3. Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
4. Hopper, P. & Traugott, E. 2003. *Grammaticalization*. 2nd ed. Cambridge: Cambridge University Press.
5. Forment-Fernández, M del Mar. (2001). "Hacer novillos, hacer campana o hacer la vaca: ¿qué fraseología enseñar?", en M^a A. Martín Zorraquino y C. Díez Pelegrín (eds.) *¿Qué Español Enseñar? Norma y Variación Lingüísticas en la Enseñanza del Español a Extranjeros*. Actas del XI Congreso de la ASELE [en línea].
6. Kennedy, G.; Meunier, F; Granger, S. (2008): *Phraseology in Foreign Language Learning and Teaching*. In: John Benjamins.
7. Kirsner, K. (1994). Implicit processes in second language learning. In N. Ellis (Ed.), *Implicit and Explicit Learning of Languages*. London: Academic Press.
8. Kühn (1987), Larger (1997), Ettinger (1998), González-Rey (2012), Corpas Pastor (1996), Penadés Martínez (1999), Ruiz Gurillo (2000), Mugrón Huerta (2015), among others.
9. Luna, C.J and Ortiz Rodríguez, C.: *La semántica cognitiva en la enseñanza-aprendizaje de las unidades fraseológicas en ELE: el ejemplo de los somatismos*. In: Universidad Autónoma de Barcelona).
10. Nuñez-Román, F.: *Enseñar fraseología: consideraciones sobre la fraseodidáctica del español*. In: Universidad de Sevilla.
11. Solano-Rodríguez, C. (1993): "Las unidades fraseológicas del francés y del español: tipología y clasificación". In: *Paremia*, 21. 117-128.
12. Wray, A & Fitzpatrick, T (2008). Why can't you just leave it alone? Deviations from memorized language as a gauge of nativelike competence... In Meunier, F. & Granger, S. (eds.) *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins.

Extracting Formulaic Expressions and Grammar and Edit Patterns to Assist Academic Writing

Jhih-Jie Chen¹, Jim Chang¹, Ching-Yu Yang¹, Mei-Hua Chen², Jason S. Chang¹

¹Department of Computer Science

National Tsing Hua University

²Department of Foreign Languages and Literature

Tunghai University

{jjc, jim, chingyu, jason}@nlpplab.cc,
chen.meihua@gmail.com

Abstract

We present a method for extracting formulaic expressions, grammar patterns, and editing rules from a given corpus to assist learners in learning to write at the level required in English for Academic Purposes. In our method, sentences in a given corpus are parsed into chunks of base phrases, with the arguments sense disambiguated to derive syntactic and semantic grammar patterns. The method involves executing shallow parsing, transforming phrases into grammar patterns, and filtering and ranking grammar patterns for each headword. We applied the proposed method to a corpus annotated with writing errors and their corrections to derive editing rules. Experiments based on a large-scale academic English corpus and *WikEd Error Corpus* showed that the proposed method produces reasonable correct grammar patterns as well as edit rules. Thus, the method has the potential to assist learners in writing and self-editing.

1 Introduction

Approximately 1 billion people are learning and using English worldwide [10], mostly as a second language (L2). Specifically [3], there are approximately 375 million native speakers of English, and 750 million people use English as a second language.

This large number of L2 learners has motivated the research and development of computer-assisted language learning (CALL) systems. *Cambridge English Write & Improve* (writeandimprove.com) helps learners by assessing a language level and providing corrective feedback for a submitted essay. *Pigaiwang*

(www.pigai.org) assigns a score along with suggested edits based on collocation information. However, the suggestions provided often lack information such as the grammatical structures or collocations related to the errors. The aforementioned systems simply deal with basic edits (replace, insert, and delete) and lack comprehensive information that currently only a human editor can provide. Thus, the corrective feedback is either limited or overwhelming, in that it provides little help for learners to improve their writing.

Specifically, the consideration of taking grammar and collocation into account stems from the analysis of the common sources of learners' writing errors in the writing section of the *Macmillan English Dictionary 2nd Edition* [4]. This analysis attributes the common errors to the use of the incorrect complementation patterns after a verb, noun, or adjective. The reason for focusing on verbs, nouns, and adjectives is that they exhibit regularity in local syntax [11]. By contrast, the *Longman Dictionary of Common Errors* [19] attributes preposition errors and verb form errors to the headword preceding the error (e.g., *have difficulty breathing* instead of *have difficulty to breath*)

Quality feedback should provide information such as that presented in the recently published *Oxford Learners' Dictionary of Academic English*, which describes how people use feedback. A local grammar is limited in scope, with a group of forms for each verb, noun, and adjective. Each rule consists of a handful of elements including syntactic symbols (n, v, adj, adv, clause), semantic categories (amount, color, ord), and function words (e.g., "wh-" words, prepositions such as in, on, at). Moreover, *Patterns Dictionary of English Verbs* (PDEV) website takes one step further toward local grammar (e.g., abandon: V n) with coarse-grained semantic labels (e.g., *abandon* ACTIVITY or PLAN; PERSON; CONCEPT; LOCATION; ARTIFACT).

Although parsing complete sentences accurately remains a difficult problem, identifying local grammar patterns of verbs, nouns, and adjectives is relatively simple, yet pedagogically useful for learners. All these aforementioned elements can be easily and accurately identified by a shallow parser (e.g., GENIA Tagger). By shallow parsing sentences in the given corpus, converting phrases to the linguistic form described in [12], and discarding less frequent patterns, we can induce high-quality grammar patterns.

Subsequently, we can disambiguate the arguments (e.g., car, house, lawyer, attorney) in a grammar pattern (e.g., *afford n*) and assign each a semantic label (e.g., ARTIFACT and PERSON). Based on the observation noted by PDEV, the classes of argument tend to obey the power law that the most frequent class dominates the less frequent ones. By disambiguating the group of nouns to *WordNet* [15] senses in order to maximize semantic closeness, we can obtain a reasonably accurate semantic annotation such as the following example:

- e.g., V n : *afford a car/house* ARTIFACT
- e.g., V n : *afford a lawyer/an attorney* PERSON

We also apply our method to an error-annotated corpus along with the corrections to derive edit patterns. For example, from the annotated sentence “*My father couldn’t afford [-paying-] {+to pay+} for my education.*”, we learn the pattern: afford -ing → afford to-inf.

The rest of the paper is organized as follows. First, we review the related work in Section 2. Next, we elaborate our method for extracting grammar patterns, assigning semantic labels to arguments, and deriving edit patterns in Section 3. In Section 4, the performance of the proposed method is evaluated by verifying the accuracy of the derived grammar patterns, semantic labels, and edit patterns over a set of noun, verb, and adjective headwords. Finally, we describe the high potential of the grammar patterns and envision the future of writing.

2 Related Work

Statistical analysis of corpus data has been an active research area for corpus linguistics, computational lexicography, and CALL.

In the area of corpus linguistics, considerable work has been conducted to identify formulaic expressions (i.e., lexical bundles), or “extended collocations” that shape the structure of text (e.g., [1]). A study [11] presented *Pattern Grammar*, a comprehensive description of the local regularity of English verbs, nouns, and adjectives. Our work addresses the computational aspect of inducing *Pattern Grammar* from raw web-scale ngrams without the benefit of full sentences, annotation, and human judgment. For example, we automatically derive the pattern “HARMFUL to n” based on the ngrams such as “*harmful to my computer/minors/the environment.*”

In the area of computational lexicography, a study [16] described the method and implemented *XTract* for extracting collocations from ngrams. Another study [13] described the lexicographic tool *Sketch Engine*, which uses a set of hand-written patterns to extract collocation from a part-of-speech (POS)-tagged corpus. More recently, a study [6] proposed a method based on rank ratio for extracting collocations and multiword expressions. In contrast to *Sketch Engine*, our method automatically extracts grammar patterns from raw ngrams based on an extended notion of rank ratio.

In the area of automatic grammatical error correction, considerable work has been conducted using rule-based and statistical approaches. Many rule-based systems have been developed for detecting and correcting article and preposition errors in non-native texts (e.g., [7]). Recently, researchers have begun to use sta-

tistical methods to correct writing errors related to articles and prepositions (e.g., [5], [17]). More recently, a study [9] described and evaluates a proofing system, English as Second Language Assistant, based on ngrams and language models. An evaluation based on an online user log showed that 36% of suggested edits were accepted. Additionally, researchers have proposed methods for handling open-class verb errors ([14]; [20]).

In the area of automated essay scoring, a study [2] described the implementation of *Criterion*, which consists of a grammar checker, *Critique*, and an automated essay scoring system, *e-rater*. This scoring system assesses a score for a student’s essay and agrees with human judges in the evaluation as often as two human judges agree with each other. More recently, a study [8] described the technique used in *Write & Improve*, which assesses goodness to each sentence of an essay and marks some obvious grammatical errors.

In contrast to previous research in corpus-based CALL, we present a system that automatically induces grammar patterns from a given web-scale corpus. Moreover, the system interacts with the user intensively and guides the writing and editing process with grammatical suggestions. We exploit *Pattern Grammar* and the inherent regularity of natural language by filtering and transforming fragmented ngrams into a set of comprehensive grammar patterns.

3 Extracting Grammar and Edit Patterns

Grammar patterns compiled for dictionaries provide much help to language learners with continuous writing patterns. However, the process is labor intensive and time consuming; thus, customizing it for any given corpus is difficult. Additionally, these patterns tend to lack semantic labels, and the general idea of usage is difficult for a learner to grasp.

We attempt to induce the patterns from a large-scale reference corpus and an annotated learner corpus to provide concise, top-down information of word usage and misuses in order to assist learners in academic writing.

3.1 Extracting Semantic Grammar Patterns

In the first stage of the extraction process, we parse the sentences, identify possible grammar patterns, and group the grammar patterns by their content words (or headwords). The extent of the patterns described here (e.g., V -ing from *afford to pay*) consists of several base phrases produced by a shallow parser (e.g., a VP phrase followed by an infinite VP). The result of this stage is grammar patterns and phrase instances tagged with coarse-grained semantic labels.

The extraction process is as follows. First, we shallow parse all sentences and produce the results in the form of a sequence of base phrase chunks (Step 1). Next, for each sentence and for each sub-sequence of phrase chunk up to a certain length (*MaxNumberOfPhrases*), we convert the phrase sequence into a sequence of elements called grammar pattern candidates (Step 2).

Subsequently, the sequence of elements is filtered by matching it against a list of legitimate grammar pattern templates and paired along with the headword and phrase instance (Step 3).

After all the sentences are processed, for each headword, we compute the instance counts C of all related patterns and the instance mean μ , as well as the standard deviation δ of C (Step 4). With the statistics, we then retain the high-frequency patterns with $C > \mu + k\delta$ (Step 5). A sample of extracted patterns with phrase instances is presented as follows.

- *afford*: V to-inf (*afford to pay/miss*)
- *afford*: V n (*afford a lawyer/attorney/car/house*)
- *afford* V n n (*afford them the opportunities/protection*)

After we have extracted patterns with their phrase instances, we continue to disambiguate (to *WordNet* senses) and label the noun arguments n (or complements) in the phrases with a semantic category (e.g., *WordNet*'s lexical information or supersenses).

Next, for each pattern and each of the nouns found in related phrases, we refer to the *WordNet* database and find the most frequent senses with their SuperSense tags—ignoring senses with less than a 5% (MiniFreqPercent)—and sum up the occurrence counts of each SuperSense tag S (Step 6).

We then estimate the sense probability of each noun $P(S|n, pat)$ according to the maximum likelihood estimation of $P(S|pat)$ (Step 7) and recalculate the estimate of $P(S|pat)$ by summing $P(S|n, pat)$ for all n (Step 8). We repeat Steps 7 and 8 until $P(S)$ converges and stabilizes. Finally, we disambiguate each noun n to an S that maximizes $P(S|n, pat)$ and produces (pat, ph, S) as the output, where S represents the semantic labels for the argument n in the phrase instance ph . A sample of the extracted patterns with the phrase instances is presented as follows.

- *abandon*: V ACTIVITY (*abandon search/effort*)
- *abandon*: V CONCEPT (*abandon plan/project*)
- *abandon* V LOCATION (*abandon car/ship*)
- *abandon* V PERSON (*abandon child/wife*)

3.2 Extracting Edit Patterns

In the second stage of the extraction process, we apply the method described in Section 3.1 to a given corpus annotated with error and correction tags. The output of this stage is a set of grammar pattern pairs (e.g., <discuss about n, discuss n>), representing the process and linguistic explanation of common edits. Our extraction process is as follows. First, we convert all sentences with a least one edit into <*SentWrong*, *SentRight*> pairs of erroneous and corrected sentences. Then, for all sentence pairs, we apply the extraction process (Steps 1 to 3 in Section 3.1) to extract the patterns *PatWrong* from *SentWrong*, where each instance of *PatWrong* spans over at least one error position (Step 1).

Subsequently, we execute the same process for *SentRight* to extract *PatRight* (Step 2). We then pair up *PatWrong* with *PatRight* to produce the pattern pairs, <*pat-wrong*, *pat-right*>, where *pat-wrong* and *pat-right* are either headed or ended by the same word and preferably of the same length (Step 3).

Finally, we compute and organize the editing pattern pairs with the corresponding counts. A sample of extracted patterns with phrase instances is presented in Table 1.

4 Experiment and Evaluation Results

WriteAhead was designed to provide general and broad-coverage writing suggestions for L2 learners. As such, *WriteAhead* was trained using a publicly available sample of 1,244 exam scripts written by learners who sat for the Cambridge ESOL First Certificate in English examination in 2000.

In this section, we first present the details of training *WriteAhead* (Section 4.1). Then, in Section 4.2, we describe the evaluation and results.

4.1 Experimental Setting

We used the *CiteseerX* corpus (460 million words in 20 million sentences) and British National Corpus (100 million words in 5 million sentences), with the combined corpus comprising approximately 500 million words in 20 million sentences. We used GENIA Tagger [18] to tag the sentences.

4.2 Evaluation and Results

After training *WriteAhead*, as described in the preceding section, we conducted a preliminary evaluation to assess its performance by using sentences containing one of the most common types of errors in the publicly available Cambridge Learner

Test Sentence	Edit Pattern & Relevant Instances	ErrType
How to make better hairstyle?	make hairstyle \rightarrow design hairstyle <i>hairstyle</i> : v N \rightarrow v N	RV
She explained him that she could not help.	explain him \rightarrow explained to him <i>explain</i> : V n \rightarrow V to n	MT
I have the chance for giving my opinion.	chance for giving \rightarrow chance to give <i>chance</i> : N for -ing \rightarrow N to v	UT+VT

Table 1: Sample test sentences and results

Corpus of First Certificate in English (CLC-FCE) [21]. After excluding punctuation, verb tense, and word order errors, we had eight types of error: Replace a Verb (RV), Preposition (RT), Noun (RN), or others (R), Missing Determinant (MD), Preposition (MT), Unnecessary Determinant (UD), Preposition (UT).

The set of test sentences was generated by sampling the *CLC-FCE* for 10 sentences for each of the 8 error types. The extracted patterns and examples for the words triggering the error were examined to determine whether the information was sufficient to make the correction.

For example, for the top writing error of replacing verbs, we selected the annotated sentence “*How to [-make-]{+create+} better hairstyle?*”. We looked for instances of v HAIRSTYLE and found the pattern instances that included create a hairstyle; we then determined that *WriteAhead* was successful in this case. Examples of the edit patterns for some of the test sentences are shown in Table 1.

5 Conclusion and Future Work

In summary, we propose a method for providing writing suggestions and editing feedback while a person is typing or hovering their mouse over some words. The method involves extracting, retrieving, and ranking grammar patterns and examples. We implemented and evaluated the proposed method as applied to a scholarly corpus with promising results.

Many avenues exist for future research on and improvement of *WriteAhead*. For example, natural language processing and machine learning techniques could be used to improve the run-time ranking of writing suggestions or corrective feedback for editing. Furthermore, an appealing direction for exploration is systematically training a classifier to predict the POS or phrase boundary of an ngram entry by using predecessor or successor characteristics extracted from ngrams. Another direction of research would be to translate patterns and instances into the learners’ first language and derive synchronous pattern grammar to support the learners.

References

- [1] Biber, D., Conrad, S.: Lexical bundles in conversation and academic prose. *Language and Computers* (1999)
- [2] Burstein, J., Chodorow, M., Leacock, C.: Automated essay evaluation: The criterion online writing service. *AI Magazine* (2004)
- [3] Crystal, D.: English as a global language. Cambridge University Press (1997)
- [4] De Cock, S., Gilquin, G., Granger, S., Lefer, M.A., Paquot, M., Ricketts, S.: Improve your writing skills. M. Rundell (editor in chief) *Macmillan English Dictionary for Advanced Learners* (2007)
- [5] De Felice, R., Pulman, S.G.: Automatically acquiring models of preposition use. In: *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions* (2007)
- [6] Deane, P.: A nonparametric method for extraction of candidate phrasal terms. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (2005)
- [7] Eeg-Olofsson, J., Knutsson, O.: Automatic grammar checking for second language learners-the use of prepositions (2003)
- [8] Felice, M., Yuan, Z., Andersen, Ø.E., Yannakoudakis, H., Kochmar, E.: Grammatical error correction using hybrid systems and type filtering. *CoNLL-2014* (2014)
- [9] Gamon, M., Leacock, C., Brockett, C., Dolan, W.B., Gao, J., Belenko, D., Klementiev, A.: Using statistical techniques and web search to correct esl errors. *Calico Journal* (2009)
- [10] Graddol, D.: The decline of the native speaker. *Translation today: trends and perspectives*. Clevedon: Multilingual Matters (2003)
- [11] Hunston, S., Francis, G.: *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. MIT Press (2000)
- [12] Hunston, S., Francis, G., Manning, E.: *Collins cobuild grammar patterns 1: verbs* (1996)
- [13] Kilgariff, A., Rychly, P., Smrz, P., Tugwell, D.: Itri-04-08 the sketch engine. *Information Technology* (2004)

- [14] Lee, J., Seneff, S.: Correcting misuse of verb forms. In: ACL (2008)
- [15] Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM (1995)
- [16] Smadja, F.: Retrieving collocations from text: Xtract. Computational linguistics (1993)
- [17] Tetreault, J.R., Chodorow, M.: The ups and downs of preposition error detection in esl writing. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 (2008)
- [18] Tsuruoka, Y., Tateishi, Y., Kim, J.D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a robust part-of-speech tagger for biomedical text. Advances in informatics (2005)
- [19] Turton, N.D.: Longman dictionary of common errors, vol. 1. Pearson Education India (1987)
- [20] Wu, J.C., Chang, Y.C., Mitamura, T., Chang, J.S.: Automatic collocation suggestion in academic writing. In: Proceedings of the ACL 2010 Conference Short Papers (2010)
- [21] Yannakoudakis, H., Briscoe, T., Medlock, B.: A new dataset and method for automatically grading esol texts. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (2011)

Improving Requirement Boilerplates Using Sequential Pattern Mining

Maxime Warnier^{1,2} and Anne Condamines¹

¹ CLLE-ERSS (Université Toulouse – Jean Jaurès & CNRS), Toulouse, France

² Centre National d’Études Spatiales, France

{maxime.warnier, anne.condamines}@univ-tlse2.fr

Abstract. In the field of requirements engineering, the use of the so-called *boilerplates* (i.e. standard phrases and sentences containing gaps to be filled in) is a popular solution to reduce variation among requirements and writers, and thus to improve the clarity of technical specifications. However, the examples of boilerplates found in the literature are often very general, as they need to be applicable to projects as varied as computer software and aircraft or industrial machines. As a result, they only partially fulfill their role, leaving a lot of freedom to the writers in charge of filling in the gaps. Instead, we would like to propose a bottom-up approach for discovering more specific sequences that could constitute either boilerplates or elements to instantiate these boilerplates. To this end, we investigate whether sequential data mining techniques can be used on a small corpus of genuine requirements written in French at CNES (Centre National d’Études Spatiales), the French Space Agency.

Keywords: pattern mining, boilerplates, requirements, corpus linguistics, phra-seology.

1 Introduction

This study is part of a research project that aims at improving requirements writing at CNES (Centre National d’Études Spatiales) – the French Space Agency –, where technical specifications are currently written mostly in French. Unfortunately, natural languages (such as French or English) are known to be inherently ambiguous [1] and vague [2]. While these properties rarely have adverse consequences in everyday communication, they become serious threats when writing and reading requirements¹, as comprehension is crucial; indeed, a misunderstanding may lead to non-compliance with the requirements, which in turn can result in delays, additional costs, litigation (because requirements represent a legal obligation for stakeholders) or in the worst case accidents. This explains why requirements engineering has become so important

¹ *requirement*: “a condition or capability that must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed documents.” [3] A *specification* is a collection of requirements.

in the last decades [4] and why so many solutions have been proposed to prevent or at least to limit this risk.

One of those solutions is what is now commonly called a Controlled Natural Language (CNL), that is, “a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax, and/or semantics, while preserving most of its natural properties” [5]. There is a wide variety of CNLs, used in many different fields; some of them remain quite similar to their “base” language, while others are close to formal notations (such as programming languages or even mathematical notations [6]), each with advantages and disadvantages. Our goal is to propose a CNL perfectly adapted to requirements writing (in French) at CNES. In other words, we want it to be as close as possible to what engineers are already used to write and read, because we are convinced that if we want our solution to be applied and to be efficient, it should be tailored and not too disruptive.

To achieve this, we propose a methodology for creating CNLs for technical writing that takes into account (1) existing CNL rules and principles – as they may be viewed as the state of the art in the field – and (2) actual samples of writing used by the members of the discourse community [7] (in our case, requirements written in unconstrained natural language for past projects) that will serve as a basis for a corpus analysis [8]. This methodology combines a corpus-based and a corpus-driven approach [9-10]: in the former case, our linguistic hypotheses (based in particular on recurring prescriptions found in other CNLs) are tested on the corpus, whereas in the latter case, linguistic regularities themselves emerge from the corpus.

These regularities (which constitute part of the grammar of the textual genre [11]) – if they exist – can then be considered when defining (or adapting) rules and when proposing standard requirements. Therefore, although the CNL we wish to create is (necessarily) normative, in our opinion it is essential that it is preceded by a descriptive analysis of the final users’ language practices; this should insure that it really meets their needs – which would not be the case if it was too general.

Such an analysis looks at recurring sequences of multiple words (or categories of words) – and at their meaning – that can be given different names, depending on their status but also on the point of view (and goals) of the researcher. Here, we will simply call them (*textual*) *patterns* [12-13-14] and emphasize that they should be specific to a (sub)genre.

A feature of our corpus is that it includes terminological and phraseological units that may belong either to the textual genre (in this case, requirements) or to the technical domain (in this case, space projects); this may be a challenge if one tries to distinguish between the two using automated tools. Since we are interested mostly in the former (we want to describe requirements writing at CNES, regardless of the project), we suggested a solution to filter the results of a terminological extraction [15].

In this paper, we focus on the extraction of frequent patterns that could then be recommended to the engineers who are in charge of writing requirements. In Section 2, we introduce the notion of boilerplate in requirements engineering and we describe our method for suggesting new ones; in Section 3, we briefly comment on some of the results we obtain when we apply the method on our corpus of requirements; finally, we conclude in Section 4.

2 Aim and Method

2.1 Requirement boilerplates

The aim of this study (which is part of the corpus-driven approach of our global methodology) is to analyze our corpus of requirements in order to semi-automatically identify frequent patterns that are possible candidates for boilerplates or fixed elements to be used in future CNES projects (currently, engineers do not use boilerplates).

The term *boilerplate* was first used fifteen years ago in requirements engineering² to refer to a textual requirement template. A very simple example of boilerplate is given by “<system> shall <action>”, where “shall” is a fixed element and “<system>” and “<action>” are attributes that will be replaced by textual values in actual requirements [16] (e.g. “The vehicle shall respect the speed limit”). They are not necessarily full sentences: some boilerplates are clauses or even phrases that can be added at the beginning or at the end of other boilerplates; all of them are basically more or less complex grammatical (non-discursive) structures.

The concept of requirement boilerplate (or requirement template) is appealing because it is very simple to understand (fill-in-the-gaps texts, frames and similar concepts existed long before requirements³) and to use, and yet it is definitely useful when the same kind of needs must be expressed repeatedly, as is often the case in large projects with thousands of requirements (sometimes with only minor variations); this is especially true when using requirements authoring and management tools (IBM Rational DOORS, for instance) that allow to handle boilerplates easily.

It can be seen as a semi-formal solution for writing requirements: texts written according to a collection of boilerplates look exactly like those written in natural language (since in theory, all sentences are grammatically correct and can be understood by people familiar with the terms without any prior training) – except that it is obviously much more repetitive –, but at the same time, they are more formal because they allow less expressivity, which is exactly why they are thought to reduce ambiguity. At the very least, it can reasonably be assumed that they should increase consistency among specifications: although variation is unavoidable in oral and written natural language (because the same need can be expressed by different words and sentences), requirements are usually quite short written texts that should be readable independently and where repetition is not only acceptable, but recommended.

Nevertheless, even though existing boilerplates may be considered convenient and easy to use, the crucial difficulty remains of course to define a set of boilerplates that will be proposed (or possibly imposed) to the writers and that preferably best suit their needs. It appears that the boilerplates⁴ suggested by the CESAR (Cost-efficient methods and processes for safety relevant embedded systems) project [17], which are one of the semi-formal “Requirement Specification Languages”, are often regarded as

² It is also used with a similar meaning in computer science (“boilerplate code”) and contractual law (“boilerplate language”: parts of a contract that are considered standard) [16].

³ In a sense, it is close to the copy-paste function of all modern computers.

⁴ Defined as “semi-complete requirements that are parameterized to suit a particular context”.

a reference. (Jeremy Dick’s online list of boilerplates is often cited too – including by CESAR’s authors – but it is no longer available.)

Examples given by CESAR include: (ex. 1) “The <user> shall be able to <capability>”; (ex. 2) “...at a minimum rate of <number> times per <unit>” and (ex. 3) “While <operational condition>”. These can be combined if necessary (e.g. “While <operational condition>, the <user> shall be able to <capability> at a minimum rate of <number> times per <unit>”).

They constitute a useful guideline for expressing capabilities and are rather flexible. However, if example 2 requires only to indicate a (positive) number and a unit (of time) and is therefore subject to only slight variations, this is not really the case for examples 1 and 3, where “user”, “capability” and “operational condition” could be replaced by almost an infinity of very different values (in the document, “order entry clerk”, “raise an invoice” and “(While) disconnected from a source of power” are given as examples, respectively). Although the attribute “user” is rather restrictive (especially if these boilerplates are used in conjunction with an ontology of the domain, where the class “user” is properly defined), “operational condition” is a very large category and the writer has no real clues on how to fill in this gap (which means s/he is free to do as s/he wishes). For instance, one could have preferred a more explicit version (where the subject is mentioned) instead of the participle clause: “While *the photographer* is disconnected from a source of power”. On the other hand, as it is, this boilerplate does not allow the writer to replace “While” by “Even when” or another conjunction that might be preferable in the context. In addition to these limitations, some of the boilerplates they propose seem very close, such as example 2 and “... at least <number> times per <unit>”, where the only difference is “at least” vs. “at a minimum rate of”.

Thus, these boilerplates are probably good examples, but they also show the main limitations of such templates: it may be difficult for the user to choose an expression over another; they are sometimes too restrictive; and most of all, they are often too permissive, because they are too generic (and their number must remain rather low). They do help limit variation, but they could probably be refined by taking better into account the environment where they will be used, so that they are more specific and adapted. According to us, this can only be done by first analyzing genuine samples of written productions thanks to corpus processing tools.

Our idea is to move beyond the canonical concept of boilerplate and to combine a few generic structures (such as the ones given above) with more specific fixed phrases (such as “...by taking into account...”) that can fill in the gaps in a flexible way. Of course, these phrases will be much more numerous, and it would not be possible for a writer to memorize them all; but they could easily be added to requirements authoring software with two major benefits. First, when a user is typing the first word(s) of a frequent phrase, the remaining words could be automatically suggested (similarly to the word completion feature proposed by some software or search engines); this would guide him/her in the difficult writing process and also save him/her time. Second, if an expression is recognized, but another one should be preferred (e.g. if “equal to or greater than” is considered clearer than “not lower than”), it could be automatically highlighted and the preferred expression could be suggested (similarly to the

spell check function of text processors) – it is then up to the writer to accept it or not. This would be a further step in reducing variation (without reducing expressiveness).

2.2 Corpus, tool and method

In order to automatically identify these phrases, we constituted a corpus made of requirements extracted from several specifications written (in French) at CNES for two past projects, namely Pleiades (two very-high-resolution satellites for Earth observation) and Microscope (a microsatellite, whose main objective is to verify a physical principle). About 2,500 requirements were extracted for Pleiades and approximately 1,000 for Microscope (a smaller project). After all tables and figures were removed (as an automatic analysis would be much more difficult), the Pleiades subcorpus was composed of slightly more than 118,000 words, and the Microscope subcorpus was composed of more than 43,000 words, for a total of 161,403 words in the corpus. (Confidentiality reasons explain why we could not get access to more data.)

We assume that this corpus is representative because it includes authentic requirements from two different projects (15 specifications for each). Although these specifications were written under similar circumstances and represent all the levels of the so-called “product trees” (system, subsystem, interface) for both projects, Pleiades and Microscope have totally different scales and purposes (as mentioned above) and the writers were not the same engineers. Consequently, the corpus as a whole is consistent and the two subcorpora (which share essential similarities but also differ in some respects) are comparable (despite the difference in size).

To automatically extract the patterns from this corpus, we used the sequential data mining tool SDMC⁵ (Sequential Data Mining under Constraints) [18]: given an input text file and several parameters (threshold, minimum and maximum length of the patterns, etc.), it returns a comprehensive list of patterns found in this file. The major drawback of this method is that, depending on the size of the text and on the parameters (in particular threshold and type of pattern: frequent, closed, maximal), the number of results may be extremely large, and a manual revision would obviously be too time-consuming if it exceeds a few thousands.

To reduce the number of results and keep only those that could be relevant for our study, we extracted frequent patterns from the two subcorpora independently⁶. The minimal length was 2 items and the threshold was 2 utterances (i.e. the minimal number for the pattern to be considered “recurring” (as noted by Sinclair and Tognini-Bonelli, quoted in [19])). Given that both subcorpora are rather small, we were forced to perform the extractions on lemmas⁷, not on word forms – therefore this is not a strict corpus-driven approach, according to Biber [19].

⁵ SDMC can be used online for free at <<https://sdmc.greyc.fr/>>. However, we had to use an offline version instead – again for confidentiality reasons.

⁶ All named entities were previously replaced by the same tag (NAM), so that they are considered the same item by SDMC. For example, thanks to this replacement, the sentences “The system must provide ORAMIC with all the data...” and “The system must provide GAZMIC with all the data...” would be considered the same pattern.

⁷ SDMC relies on the results of TreeTagger [20] for lemmatization.

We obtained more than 63,000 patterns for Microscope and almost 100,000 patterns for Pleiades, but we kept only the 3,854 patterns that were common to both lists. This is a much more reasonable number and insures that they are recurring in different projects (and not specific to only one⁸).

To further lower it, we performed a third extraction, this time on a collection of newspaper articles (from *Le Monde*) composed of exactly ten times as many words as the requirements corpus; all the patterns that were present in these results were then removed from our previous list, since they are common also in another genre and thus cannot be considered specific to requirements (SDMC does not use a reference corpus for comparison). Since the minimum number of utterances in the whole requirements corpus was 4 (2+2; a relative frequency of 1 per 40,000 words approximately) and the newspaper corpus is ten times bigger, the threshold for patterns with the same relative frequency would be 40 (4×10). Instead, we decided to set it to 10, meaning that we kept only patterns that were at least four times more frequent in requirements than in newspaper articles. Thanks to this final step, an additional 1,413 patterns⁹ were eliminated.

The resulting list is composed of 2,441 patterns (the longest one is composed of 9 items and the average length is 2.8 items) that were manually reviewed. Some of them are commented in the next section.

3 Examples

Unlike those proposed by CESAR, very few of the patterns that were extracted can be used directly as full sentences or even as clauses. Among them are “Il est à noter que...” (“It should be noted that...”) – where “Il est” is actually often omitted – and “Il doit être possible de...” (“It must be possible to...”).

Unsurprisingly, “être” (“to be”) and “devoir” (“must”) are by far the most common verbs found in the patterns. The copula “être” may be followed by an adjective: “actif” (“active”), “autonome” (“autonomous”), “cohérent” (“consistent”), “compatible”, “responsable” (“responsible”), etc. From these patterns, it may be possible to propose a very generic boilerplate, e.g. *<subject> must be <adjective>*, or somewhat more restrictive ones, such as *<system> must be <adjective>*, where *<adjective>* could be any of the above mentioned values. However, “compatible” would not be an acceptable parameter if the subject was *<human operator>*.

As the auxiliary of the passive voice, “être” is also often followed by another verb (a past participle): “activé” (“activated”), “calculé” (“calculated”), “sauvegardé” (“saved”), “validé” (“validated”), etc. The modal “devoir” too is followed by several verbs (infinitives): “exécuter” (“to execute”), “générer” (“to generate”), “fournir” (“to provide”). Interestingly, some of these verbs can be found after “être” and after “devoir”, which proves that a same need is sometimes expressed in the active voice (“Le

⁸ For instance, “différer l’envoi” (“defer sending”) appears twice in Pleiades, but never in Microscope, while “être calculé à bord” (“to be computed on board”) has three utterances in Microscope, but none in Pleiades.

⁹ Such as “par rapport à” (“in relation to”) and “la position de” (“the position of”).

CECT doit conserver une copie locale”: “*must keep* a local copy”) and sometimes in the passive voice (“la donnée est conservée en ligne”: “*is kept* online”)¹⁰. Some verbs are peculiar, because they are found after the negation of “devoir”¹¹: “entraîner” (“to cause”), “dépasser” and “excéder” (both meaning “to exceed”).

Some expressions are very specialized such as “gérer en configuration” (used only in software configuration management), while others exist in general language but seem to be specific collocations in the corpus, such as “en phase de [routine/qualification]” (“in routine/qualification phase”), or even “en phase de recette en vol”, which seems to be used only at CNES.

Finally, we would like to show that several expressions may be used with the same purpose. This is the case, among many other examples, with the prepositional phrases “à l’aide de [l’IHM]”/“grâce à [la fonction]”/“au moyen de [l’outil]”/“au travers de [cette fonction]”, that can be all translated by “through”, “thanks to” or “by means of”; the past participles “[être] décrit/défini/spécifié/détaillé [dans le document...]” (“[be] described/defined/specified/detailed [in document...]”); the verbs “avoir en charge [la réception]”/“être responsable de [la gestion]”/“gérer [la liste des clés]” (“to be in charge of”/“to be responsible for”/“to manage”); the patterns “être conforme à [la spécification]”/“conformément à [la définition]”/“dans le respect de [la spécification]”/“doit respecter [les exigences de sécurité]” (“in compliance with”/“in keeping with”/“must respect”); or “La durée maximale/maximum [avant délivrance] est de <nombre> <unité>”/“La durée [de stockage] est limitée à/ne doit pas dépasser/ne doit pas excéder/doit être inférieure ou égale à <nombre> <unité>” (“The maximum duration is <number> <unit>”/“The duration is limited to/must not exceed/must be less than or equal to <number> <unit>”). This last case is very frequent and, as can be seen, there are plenty of possible variations; a simple boilerplate such as *The <maximum/minimum> <duration/length/...> <complement> is <number> <unit>* might be useful to maintain consistency among requirements. In the other cases, one could simply recommend to use only one of the phrases whenever possible.

4 Conclusion

In this paper, we evaluated the feasibility and interest of a simple method (which can still be refined) for extracting frequent patterns in a corpus with the aim of reducing linguistic variation in requirements, either by suggesting boilerplates or by identifying “competing” phrases expressing the same need. Of course, for each of them, the experts will be in charge of choosing the most relevant one before the list can be added to requirements authoring tools for automatic suggestion. This decision could be made through a survey where variants are presented in context and must be rated [21].

We believe that in specific contexts, such as technical writing, rules and norms may be very useful to users, as long as they are based on existing regularities, and not on mere intuitions or value judgments, as is unfortunately too often the case.

¹⁰ Of course, a combination is possible as well: “must be kept”.

¹¹ “ne pas devoir” is sometimes ambiguous in French: “must not”/“does not have to”. Here, the first translation is very likely to be preferred.

Acknowledgment

We would like to thank Daniel Galarreta, Nicolas Deslandres and Jean-François Gory for their strong involvement in this doctoral research, which was granted by CNES and the Regional Council of Midi-Pyrénées (France). We also wish to thank the developers of SDMC, especially because this study would not have been possible without the offline version.

References

1. Kamsties, E., Peach, B.: Taming ambiguity in natural language requirements. In: Proceedings of the Thirteenth International Conference on Software and Systems Engineering and Applications (2000).
2. Zhang, Q.: Fuzziness - vagueness - generality - ambiguity. *Journal of Pragmatics*. 29, 13–31 (1998).
3. IEEE Standard Glossary of Software Engineering Terminology. IEEE Std 610.12-1990. 1–84 (1990).
4. Nuseibeh, B., Easterbrook, S.: Requirements Engineering: A Roadmap. In: Proceedings of the Conference on the Future of Software Engineering. pp. 35–46. ACM, New York, NY, USA (2000).
5. Kuhn, T.: A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*. 40, 121–170 (2014).
6. Meyer, B.: On Formalism in Specifications. *IEEE Softw.* 2, 6–26 (1985).
7. Swales, J.M.: Reflections on the concept of discourse community. *ASp. La revue du GERAS*. 7–19 (2016).
8. Condamines, A., Warnier, M.: Towards the creation of a CNL adapted to requirements writing by combining writing recommendations and spontaneous regularities: example in a space project. *Lang Resources & Evaluation*. 51, 221–247 (2017).
9. Tognini-Bonelli, E.: *Corpus Linguistics at Work*. John Benjamins Publishing (2001).
10. Biber, D.: *Corpus-Based and Corpus-driven Analyses of Language Variation and Use*. In: Heine, B. and Narrog, H. (eds.) *The Oxford Handbook of Linguistic Analysis*. Oxford University Press (2009).
11. Bhatia, V.K.: *Analysing genre: Language use in professional settings*. Longman, London (1993).
12. Quiniou, S., Cellier, P., Charnois, T., Legallois, D.: Fouille de données pour la stylistique: cas des motifs séquentiels émergents. In: *Journées Internationales d’Analyse Statistique des Données Textuelles (JADT’12)*. pp. 821–833 (2012).
13. Sitri, F., Tutin, A.: Présentation. *Lidil. Revue de linguistique et de didactique des langues*. 5–18 (2016).
14. Legallois, D., Charnois, T., Poibeau, T.: Repérer les clichés dans les romans sentimentaux grâce à la méthode des « motifs ». *Lidil. Revue de linguistique et de didactique des langues*. 95–117 (2016).
15. Warnier, M., Condamines, A.: A Methodology for Identifying Terms and Patterns Specific to Requirements as a Textual Genre Using Automated Tools. Presented at the Terminology and Artificial Intelligence (TIA’2015) November 4 (2015).
16. Tommila, T., Pakonen, A.: Controlled natural language requirements in the design and analysis of safety critical I&C systems. SAFIR2014 Reference group. 2, (2014).

17. Rajan, A., Wahl, T. eds: CESAR - Cost-efficient Methods and Processes for Safety-relevant Embedded Systems. Springer Vienna, Vienna (2013).
18. Quiniou, S., Cellier, P., Charnois, T., Legallois, D.: What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics? In: International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'12). pp. 166–177. New Delhi, India (2012).
19. Biber, D.: A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International journal of corpus linguistics*. 14, 275–311 (2009).
20. Schmid, H.: Treetagger: a language independent part-of-speech tagger.
21. Warnier, M., Condamines, A.: A Case Study on Evaluating the Relevance of Some Rules for Writing Requirements through an Online Survey. In: 25th IEEE International Requirements Engineering Conference. To appear.

Intonational PEriods (IPE) and Formulaic Language: A Genre-based Analysis of a French Speech Database

Maria Zimina¹ and Nicolas Ballier¹

¹ Univ Paris Diderot, Sorbonne Paris Cité, CLILLAC-ARP, EA3967, 75013, Paris, France
mzimina@eila.univ-paris-diderot.fr
nicolas.ballier@univ-paris-diderot.fr

Abstract. This paper addresses prosodic aspects of phraseology from the point of view of the ‘lexicogrammar’ approach in *Rhapsodie*, a richly annotated corpus of spoken French. Textometric methods enable the selection of statistically relevant phenomena both in terms of marked prosodic salience and recurrent lexicogrammatical features. Among the possible prosodic characteristics of phraseology, salient initial prominences of prosodic macro-units are considered in this paper. Within the *Rhapsodie* annotation framework, these macro-units correspond to the largest prosodic constituents, the Intonational PEriods (IPE). We have analysed the IPEs bearing the strongest level of initial prosodic salience. The prosodic properties of the units so delineated are discussed, as well as the discourse type of these phenomena. Some recurrent patterns of oratory and procedural speech genres are described using prosody and Part-of-Speech (POS) annotations. We suggest that the role of initial salience of specific prosodic patterns is to facilitate perception and interaction in human speech and to establish the structure of speakers’ turns.

Keywords: Lexicogrammar, Prosodic Constituents, Salience, Textometrics

1 Introduction

1.1 Phraseology and Prosody

Our research examines the notions of phraseology and formulaic language in the process of speech production. The terms “phraseology” and “formulaic language” will be discussed in this paper from the point of view of the ‘lexicogrammar’ approach [5]. From this perspective, our objects of study are predictable and productive sequences of signs called lexicogrammatical patterns (lexical signs, grammatical constructions). Composed of permanent ‘pivotal’ signs and a more productive ‘paradigm’, these patterns may be discontinuous and may or may not be syntactic constituents. In contrast with the lexicological approach to phraseology, which studies phraseological phenomena in terms of a continuum ranging from ‘free combinations’ to ‘fixed phrases’, the ‘lexicogrammar’ approach is particularly suited to identifying extended patterns within a particular register or genre [ibid.]. We aim to explore the ways in which prosodic features may correlate with extended lexical patterns, as well as the extent to

which prosody corresponds to patterns which have a particular register or discourse function.

Recent studies have shown that prosodic features extracted from speech databases can be successfully integrated into the investigation of phraseology [16] and its pedagogical applications [1]. Prosody is key to fluency in speech production and reception [ibid.]. Most transcriptions of prosody indicate specific events in speech: boundary tones, pitch accents, disfluent segments, etc. [21]. These speech events coded in spoken corpora are possible candidates for identifying the prosodic characteristics of formulaic language [7] [20].

In this respect, a quantitative analysis of finely annotated spoken corpora facilitates research on the prosodic aspects of phraseology: *“If most of the formulaic expressions we know have been acquired from and are used in speech, the phonological representation of formulaic expressions should, in theory, play a fundamental role in the lexical storage and retrieval.”* [16]. Unfortunately, large spoken corpora are rarely distributed with a fine-grained prosodic annotation.

For French, a free reference corpus, the *Rhapsodie* speech database (ANR Rhapsodie 07 Corp-030-01), is now available [17]. This syntactic and prosodic treebank is composed of 57 short samples of spoken French (approximately 5 minutes long), orthographically and phonetically transcribed (approximately 33,000 words). This corpus was designed to investigate the prosody/syntax/discourse interface across several discourse types and speaking styles (oratory, narrative, description, argumentation, procedural; interactive, public and private; semi-interactive and non-interactive; planned, spontaneous and semi-spontaneous, etc.) [11]. The resource can be downloaded from www.projet-rhapsodie.fr.

1.2 A Summary of the Rhapsodie Methodology for Prosodic Annotation

The *Rhapsodie* project follows a bottom-up approach driven by the data [10]. The transcriptions and the annotations are aligned on the speech signal and *Praat* Text-grids [2] are available online [17]. The prosodic annotation is based on the assumption that, out of the total acoustic signal, only certain perceptual cues selected by the listener are relevant for linguistic communication [10]. Following this assumption, 10 research teams collaborated on the following workflow:¹ (1) Manual annotation of relevant perceptual prosodic events. (2) Automatic characterization of the prosodic constituents based on this manual annotation. (3) Automatic stylization of melodic contours and annotation of tones associated with the prosodic constituents.

This combination of manual and automated annotations allowed a segmentation of speech into prosodic periods [12], which relies on the initial characterization of two types of speech events retained from the manual annotation: prosodic prominence and disfluencies. For illustration purposes, Fig. 1 gives a summary of this prosodic structure [10]. It is organized around rhythmic and melodic components. The hierarchy of constituents includes:

¹ <http://www.projet-rhapsodie.fr/laboratoires>, last accessed 2017/09/01.

1. Intonational PERiods (IPE)
2. Intonational PACKages (IPA): sub-constituents internal to periods
3. Rhythmic Groups (RG): sub-constituents internal to intonational packages
4. Metrical Feet (MF): sub-constituents inside rhythmic groups
5. Syllables, with Prominence levels, including: 0 (non-prominent), W (weak) and S (strong).

IPE	que vous soyez devenue une vedette vous étiez normalement entraînée															
IPA	que vous soyez devenue une vedette vous étiez normalement entraînée															
RG	que vous soyez devenue				une vedette				vous étiez				normalement		entraînée	
MF	kvuswajədɔvny				ynvədɛt				vuzɛtjɛ				nɔr	malmã	ãtrene	
syllable	kvu	swa	je	dɔv	ny	yn	və	det	vu	ze	tje	nɔr	mal	mã	ã	tre ne
Prom	0	0	0	0	W	0	0	W	0	0	W	S	0	0	0	0 S

Fig. 1. Prosodic structure of the *Rhapsodie* speech database corpus [10]

The speech dataset annotated within this perception-driven prosodic annotation opens up new possibilities for the investigation of phraseology. As the link between the “marked status” as a +phrase/expression/formulaic expression etc. and prosodic constituents is still to be revealed, some of our research questions are of an exploratory nature and more than 60 layers of morpho-syntactic, syntactic, macro-syntactic and prosodic annotation in *Rhapsodie* necessarily open new perspectives for the exploration of the prosodic dimension of phraseology. We have decided to focus on the initial structure of IPE macro-units and on the contribution of recurrent patterns of initial prosodic salience to the perception of formulaic language. Previous research has considered various candidates for specific prosodic contours (sometimes called ‘melodic clichés’ [9]) in the phraseology of spoken discourse [3] [16]. Our quantitative approach focuses on the recurrent prominences observable after speech breaks.

2 Exploring Linguistic Fixedness in French on the Basis of Prosodic Features

2.1 Textometric Procedures

The segmentation of the *Rhapsodie* speech data into IPE considers melodic variations in time and silent pauses used, regardless of segmental and syntactic constraints [10]. Following our work on the analysis of lexicogrammatical patterns in written texts [6], our study explores the prosody/lexicogrammar interface. In order to discover regular patterns, we conduct a textometric analysis of repeated POS segments in relation to IPE boundaries. At this stage, the goal is to isolate a set of linguistic regularities associated with what is commonly perceived as a strong prosodic boundary [10]. Computation of **characteristic elements** [14], which comes after the first stage, aims to describe perceived regularities with respect to genre and speaking styles, categorized in *Rhapsodie* as ‘subgenres’ (see Figure 2 below) [17] [10].

Typically, a **hypergeometric model** [15] is the statistical rationale for the computation of indices signaling characteristic POS or POS repetitions within each of the

corpus parts (interactive/non-interactive/semi-interactive speech, dialogue / monologue, subgenre, etc.). The computation adapts classical statistical tests [14] that can detect, within each of the parts of a corpus, which elements are used frequently as well as the ones which tend to be rarely used. As a consequence, characteristic elements discovered in this second stage allow for the investigation of candidate formulae in the breaking of the speech flow across different social contexts. Different variables can be analyzed within this approach using textometric software.

2.2 Le Trameur Software

The software used in our study is *Le Trameur* [13]. It allows for the intersecting analysis of multiple speech and text annotation layers in various forms of textometric analysis. For example, more than 60 annotations are used in *Rhapsodie*. They are all displayed and processed in a single graphical user interface [4]. *Le Trameur* can be used to automatically re-annotate the dataset and potentially add new annotation layers. Moreover, the researcher can select and manually correct any occurrence of a given tag of the dataset. Various textometric analyses available using this software have allowed us to compare the specific lexicogrammatical regularities of IPE characteristics in several communicative situations. We used frequent characteristic patterns detected as “starters” of the main prosodic constituents to detect formulaic expressions in different speech contexts. The following section illustrates some of the first findings resulting from our experiments with the *Rhapsodie* corpus.

3 Characteristic POS Patterns at the Beginning of the IPE in Several Speech Genres of Rhapsodie

3.1 Detection of Salient POS and POS Patterns

Table 1. The most frequent POS in the IPE initial position of strongest salience (2, 609 occ.)

POS list	Strongest initial prosodic salience	Total of the POS (any position)
1. Cl (Clitic pronoun)	511 occ.	4, 179 occ.
2. J (Coordinating conjunction)	443 occ.	1, 142 occ.
3. I (Interjection)	439 occ.	1, 984 occ.
4. Adv (Adverb)	287 occ.	2, 789 occ.
5. Pre (Preposition)	238 occ.	3, 443 occ.
6. D (Determiner)	209 occ.	4, 080 occ.
7. V (Verb)	112 occ.	5, 994 occ.
8. Qu (Relative pronoun)	97 occ.	799 occ.
9. CS (Subordinating conjunction)	74 occ.	729 occ.
10. N (Noun)	65 occ.	6, 317 occ.

In decreasing order of frequency, clitic pronouns, coordinating conjunctions, interjections, adverbs, prepositions, determiners, verbs, relative pronouns, subordinating conjunctions and nouns are the most frequent POS categories (resulting from morpho-syntactic tagging with *SEM* [19]) that occur at the beginning of the IPE (Freq>50), see Table 1 above.

Repeated segments computation [18] is further used to study the specific attractions of these POS categories at the beginning of the IPE contexts. Table 2 presents the most frequent POS recurrences (Freq>50) in the IPE initial position.

Table 2. The most frequent POS recurrences (POS N-grams) in the IPE initial position

POS repeated segment N-gram list	Strongest initial prosodic salience	Total of the N-gram (any position)
1. CL + V	257 occ.	2, 223 occ.
2. D + N	129 occ.	2, 919 occ.
3. Pre + D	90 occ.	1, 112 occ.
4. J + Cl	77 occ.	164 occ.
5. Cl + Cl	76 occ.	525 occ.
6. J + Adv	70 occ.	150 occ.
7. Cl + Cl + V	69 occ.	479 occ.
8. J + I	67 occ.	107 occ.
9. I + I	60 occ.	258 occ.
10. Pre + D + N	55 occ.	939 occ.

3.2 Characteristic Elements in Different Speech Contexts

Characteristic elements [15] describe parts of the corpus displaying these POS repetitions that are significantly more salient or a great deal less salient in a given part of the corpus than in the overall corpus. For example, it is clear from the graphs on Fig. 2 that Cl + V is a positive characteristic element at the beginning of the IPE in the speech contexts of oratory genre (specificity indice: +10). The following examples reveal some lexicogrammatical realizations of this productive pattern in *Rhapsodie* (categories corresponding to **CL + V** are in bold in the following examples):

- Oratory genre: IPE starting with **Cl + V** (rhetorical function: performatives)
 - # **je suis** heureux de me retrouver ce soir #
 - # **elle salue** la loyauté #
 - # **il faut** les faire grandir #
 - # **je souhaite** que l'Europe #

Another strong characteristic pattern of this genre is D + N (specificity: +26):

- Oratory genre: IPE starting with **D + N** (rhetorical function: theme-selection)

- # **la démocratie** politique et sociale #
- # **la France** sera ce que nous voudrions qu'elle soit # une nation unie #
- # **le droit** de grève # le droit à l'instruction #
- # **un moment** fort #
- # **l'exigence** de solidarité #

CI + V is also revealed as a positive characteristic element of initial prosodic salience (specificity: +6) in the procedural speech contexts:

- Procedural genre: IPE starting with **CI + V** (rhetorical function: instructions)

- # **on passe** devant le le kiosque à journaux #
- # **tu vas** tout droit #
- # **vous continuez** # vous prenez le rond-point tout droit #
- # **on traverse** la rue #
- # **tu descends** toute la pente #

Coordinating conjunction (J) is a characteristic element of initial salience in procedural speech. It is present in three characteristic patterns: J + I (specificity: +5), J + CL (specificity: +3), J + Adv (specificity: +3):

- Procedural genre: IPE starting with **J + I/CL/Adv** (instructions, hesitation)

- # **et vous** allez toujours tout droit #
- # **et vous** suivez toujours la ligne du tram #
- # **et euh** donc je vais jusqu'au & jusqu'à la place Victor Hugo #
- # **et là** je me retrouve euh en effet euh s~ près des rails du tram #
- # **et euh** et ben voilà j'arrive au niveau de la grande place de la gare où ... #

4 The Role of Formulaic Expressions in the Organization of Speech

Our findings can be summed up as a set of observational statements:

1. Speech boundaries of intonational periods (IPE) can be easily related to the characteristic repetitions of lexico-grammatical patterns revealed by POS recurrences at the beginning of the IPE. This prosodic salience of recurrent initial left-aligned patterns is a valuable property which can be used to explore how prosody is related to formulaic language. For example, in the French reference corpus, intonational periods tend to begin with specific POS categories, the most frequent ones being clitic pronouns, coordinating conjunctions, interjections, adverbs, prepositions, determiners and verbs. These results are to be compared with other available speech corpora data in French.

2. The initial IPEs vary in different social contexts and reflect specific communicative needs of the speakers. However, the pivotal elements of these productive patterns, such as the expression of predicates CL+V, have stable lexico-grammatical realizations in the *Rhapsodie* speech dataset (“*je salue*”, “*elle souhaite*”, “*il faut*”, “*on continue*”, etc.). These elements are regularly reproduced in specific linguistic contexts and reflect regular rhetorical units (performatives, theme-selection, instructions etc.) with predictable/definable discourse functions.
3. The linguistic characteristics of these salient contexts and their degree of semantic unity are quite different. For example, the illocutionary unit § # *on passe devant le kiosque à journaux* # shows a possible extension of the pivotal element “*on passe*” in a specific communicative situation, while # *un moment fort* # has a greater semantic unity. More experiments with the *Rhapsodie* dataset are necessary to explore the precise nature of these phenomena using other annotation levels.

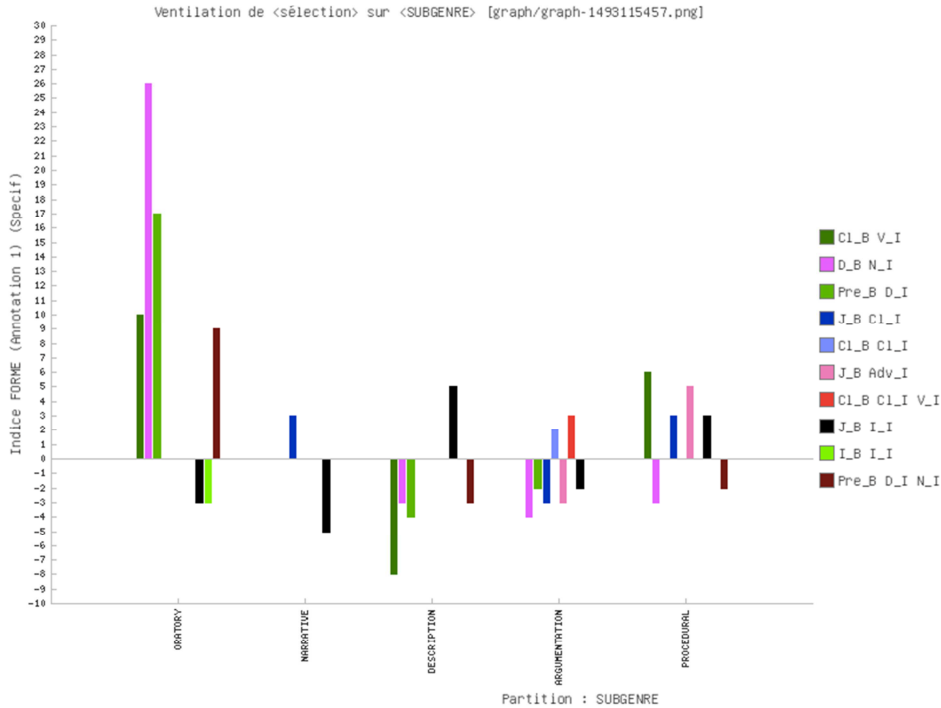


Fig. 2. Specificity of POS N-grams at the beginning of the IPE (with positional properties labeled as _B: Beginning, _I: Inside) in several speech genres of *Rhapsodie*

5 Future Work

One of the interesting outputs of this study is that it paves the way for a new investigation on inter-annotator agreement in prosodic processing. As manual annotation is

part of prosodic segmentation in *Rhapsodie*, future experiments could consider constraints of memory [8] and processing capabilities of the annotators in determining IPE boundaries when it comes to formulaic language. However, we are convinced that for the present study of the phraseology/prosody interactions, the specificities of the *Rhapsodie* speech database are valuable inputs.

6 Conclusion

Prosodic segmentation into intonational periods offers new insights for the observation of the functions of formulaic expressions in speech. A possible place to start with is the identification of annotated IPE boundaries and their correspondence with frequent morpho-syntactic repetitions. In this respect, characteristic POS repetitions at the beginning of intonation periods are more than simply recurrent groups of linguistic units. In our opinion, they represent an observational tool that can be used to investigate how prosodic variations depending upon several factors (interactional need, social context, genres, etc.) are related to formulaic language.

The experiments proposed in our study represent an attempt to account for the uses of these specific patterns after prosodic breaks where the speakers are likely to rely upon the formulaic language for specific communication purposes. A possible interpretation of these prosodic features naturally comes from the analysis of different speech contexts where the initial salience of specific prosodic phraseology facilitates speech perception. In this respect, recurrent patterns are likely to reflect strong speech signals to which speakers and listeners respond in a distinct way, showing an important influence of intrinsic experience of language acquisition in the structuring of speaker and listener interaction, speakers' turns, etc. This type of analysis can be further extended to include all the other prosodic characteristics (tone, pause length, etc.) available in the *Rhapsodie* speech dataset.

References

1. Aston, G.: Learning phraseology from speech corpora. In: Leńko-Szymańska, A., Boulton, A. (eds.) *Multiple Affordances of Language Corpora for Data-driven Learning* (Studies in Corpus Linguistics 69), pp. 63–84. John Benjamins, Amsterdam-Philadelphia (2015).
2. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [Computer program]. Version 6.0.29, retrieved 2017/09/01 from <http://www.praat.org/>
3. Cheng, W., Greaves, C., Warren, M.: *A corpus-driven study of discourse intonation: the Hong Kong corpus of spoken English (prosodic)*. John Benjamins: Amsterdam-Philadelphia (2008).
4. Fleury, S., Zimina, M.: Trameur: A Framework for Annotated Text Corpora Exploration. In: *Proceedings of 25th International Conference on Computational Linguistics (COLING 2014)*, August 2014, Dublin, Ireland. *Proceedings of COLING 2014 the 25th International Conference on Computational Linguistics: System Demonstrations*, August 2014, Dublin, Ireland, pp.57–61 (2014), <http://www.aclweb.org/anthology/C14-2013.pdf>, last accessed 2017/09/01.

5. Gledhill, C.: The 'lexicogrammar' approach to analysing phraseology and collocation in ESP texts. *ASp (Anglais de Spécialité)* 59, 05–23 (2011).
6. Gledhill C., Patin S., Zimina M.: Identification et visualisation de schémas lexicogrammaticaux caractéristiques dans deux corpus juridiques comparables en français. *CORPUS* 17 (2017), <https://corpus.revues.org/>, last accessed 2017/09/01.
7. Granger, S.: Pushing back the limits of phraseology. How far can we go? In: Cosme, C., Gouverneur, C., Meunier, F., Paquot, M. (eds.): *Proceedings of PHRASEOLOGY 2005. An Interdisciplinary Conference*, Université Catholique de Louvain, Louvain-la-Neuve, pp. 165–168 (2005).
8. Gurevich, O., Johnson, M. A., Goldberg, A. E.: Incidental verbatim memory for language. *Language and Cognition* 2 (1), 45–78 (2010).
9. Kawaguchi, Y., Fonagy, I., Moriguchi, T.: *Prosody and Syntax: Cross-linguistic Perspectives*, John Benjamins, Amsterdam-Philadelphia (2006).
10. Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J-P., Obin, N., Pietrandrea, P., Tchobanov, A.: *Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French*. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014, <http://www.lrec-conf.org/proceedings/lrec2014/index.html>, last accessed 2017/09/01.
11. Lacheret, A., Kahane, S., Pietrandrea, P. (eds.): *Rhapsodie: a prosodic and syntactic treebank of spoken French*, John Benjamins, Amsterdam-Philadelphia (2017).
12. Lacheret, A., Victorri, B.: La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques. *Verbum* 24 (1-2), 55–73 (2002).
13. LE TRAMEUR Homepage, <http://www.tal.univ-paris3.fr/trameur/>, last accessed 2017/09/01.
14. Lebart, L., Salem, A., Berry, L.: Recent developments in the statistical processing of textual data. *Applied Stochastic Models and Data Analysis* 7 (1), 47–62 (1991).
15. Lebart, L., Salem, A., Berry, L.: *Exploring Textual Data*. Kluwer Academic Publishers, Dordrecht, Boston (1998).
16. Lin, Ph. M.S.: The prosody of formulaic expression in the IBM/Lancaster Spoken English Corpus International Journal of Corpus Linguistics. *International Journal of Corpus Linguistics* 18 (4), 561–588 (2013).
17. RHAPSODIE Homepage, <http://www.projet-rhapsodie.fr/>, last accessed 2017/09/01.
18. Salem, A.: *Pratique des segments répétés. Essai de statistique textuelle*, Klincksieck, Paris (1987).
19. SEM Homepage, <http://www.lattice.cnrs.fr/sites/itellier/SEM.html>, last accessed 2017/09/01.
20. Wray, A.: *Formulaic language: Pushing the boundaries*, Oxford University Press, Oxford (2008).
21. Yoo, H-Y, Delais-Roussarie, E. (eds.): *Actes d'IDP 2009*, Paris, France, September (2009), http://makino.linguist.jussieu.fr/idp09/actes_fr.html, last accessed 2017/09/01.

Google N-grams Viewer and Food Idioms

Sarah V. C. Ribeiro¹ and Paula L. C. Lima²

¹ Instituto Federal do Ceará (IFCE) and Universidade Estadual do Ceará (UECE), Fortaleza-CE, Brazil – sarah.virginia@aluno.uece.br

² Universidade Estadual do Ceará (UECE), Fortaleza-CE, Brazil – paula.lenz@uece.br

Abstract. The purpose of this study is to use the Google Books N-gram Viewer as a tool of investigation in order to show the frequency of use and possible obsolescence of 16 food idioms in English. We analysed the percentage of use for the first and latest records and tallest spike of each idiom, as well as the period of time they occurred. We found evidence that some of them are in very little use and with a frequency of use in decrease, while others follow the opposite direction. We also compared these results with Webcorp occurrences of the same idioms and the findings were similar for most of them. The Google N-gram Viewer was found to be an appropriate tool to analyse the frequency of use of idioms.

Keywords: Frequency of Use, Obsolescence, *Corpus* Linguistics.

1 A first view

Idioms are part of our every-day language and, as such, they are an important topic that relates to different fields of study, such as machine translation, lexicography and second language acquisition, among others. They belong to figurative language and, for a long time, were traditionally considered as frozen constructions, but “new theories on metaphor comprehension have shed lights upon idiom studies, encouraging different perspectives [6]. These gave more emphasis to their cognitive essence rather than their semantic origins. Scholars such as Lakoff, Gibbs and Giora, among others, have brought important insights on the mechanisms of idiom comprehension. The number of studies on idioms has constantly increased in the last 5 decades [6]. Aspects like transparency, decomposability, salience and conventionality play an important role in order to determine idiom comprehension. Familiarity is another aspect, which is directly related to the frequency of use of this type of language.

Despite their importance, it is sometimes hard to know whether some idioms are still in use or have become obsolete. Many times, they are only seen in dictionaries, as a record of an expression that was highly used for some time, but has somehow fallen out of interest. The purpose of this study is to investigate the appropriateness of using the Google Books N-gram Viewer (GBNV, hereafter) to verify the frequency of use and possible obsolescence of 16 English idioms that have food names in their composition. This computer tool has more than 5 million books published from 1500 to 2008, contains 500 billion words from various monograph/book materials found in the Google Books collection as its *corpora*, and shows the occurrence of words (n-grams) or short phrases (up to 5 words) in the form of a plotted line chart.

2 A better view

Since GBNV's first release in 2009, many of its positive and negative aspects have been discussed. Some of the negative critics concerned the quality of the optical character recognition (OCR) software and other conditions that reduced digital image quality [8], or the overabundance of scientific literature, or yet, the messy metadata [9]. One of the positive aspects was the size of the *corpora* compared to other *corpora* available at that time. Although some scholars were excited about the possibilities of such large *corpora*, several others were sceptical about its dependability [3]. Another positive aspect was that Google gave the possibility of freely downloading the raw data available. According to Davies [4], one thing the GBNV 2009 version did well was "to show the frequency of a given word or exact phrase over time, which provides insight into lexical shifts in the language". Cohen [3] states that the best possibilities of using GBNV might be for longer grams "since they begin to provide some context." Their vision endorse the appropriateness of using this tool to achieve the objective of this study.

The GBNV 2012 release brought advances, such as the improvement of the OCR system and the inclusion of wildcards and other features, bringing more functionality to the searches [5]. Thus, this is the version we used for this analysis.

The 16 idioms analyzed here are licensed by the DIFFICULTY/EASINESS IS A FOOD DIFFICULT/EASY TO HANDLE/DIGEST metaphor, and were among those taken from two dictionaries of idioms [1][2] which make up the *corpus* of our broader study on food-idiom machine translation and conceptual metaphors. The obsolescence or frequency of use of the idioms may influence the quality of human or machine translation, more so for machine translators that use statistical paradigms.

For this study, we used the GBNV filters: time span from 1800 to 2008, with 0 smoothing, and with the case insensitive box activated (although it was not always possible to use this function, e.g. with wildcards, a limitation of the tool itself).

We searched each idiom individually and, for a few, we searched more than once since there was the possibility of different spellings (e.g. **sell like hot cakes/hotcakes**) and other variations (e.g. **get/got out of a jam**). All the graphs were analyzed and their percentages taken notes. We checked all the sentences (books) given for each idiom to confirm their idiomatic use. In order to validate the results, we crossed them with the number of occurrences of the same 16 idioms generated by the Webcorp whose idiomatic use we have previously confirmed. The Webcorp [7] is an online search engine, which allows access to the World Wide Web as a *corpus*, making it possible to extract concordances of the word(s) searched and generating much updated results.

3 A detailed view

An example of the charts plotted for the searches is presented in Fig. 1, the idiom **not cut the mustard anymore**. As shown, its first record occurred in 1968, its tallest spike was in 1981, with a percentage of use of 0.000001200%¹. The chart also shows other spikes during the period of use searched, and the percentage of use of 0% in 2008, which indicates that this idiom might be obsolete.

¹ The frequency is calculated according to the number of words in the GBNV *corpus*.

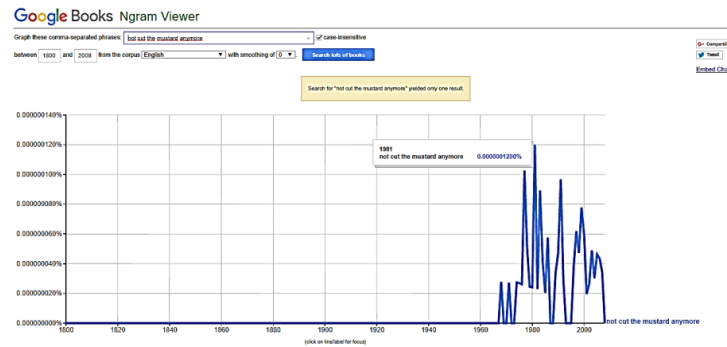


Fig. 1. Chart plotted for the idiom **not cut the mustard anymore**.

It is important to mention that GBNV only considers n-grams that occur in, at least, 40 books; otherwise, it plots a flat line [5]. Table 1 shows the percentage charted for the first and latest records (2008 for all), and the tallest spike (the highest percentage of frequency). It also brings the number of occurrences generated by the Webcorp.

Table 1. Frequency of use percentages of first and latest record, tallest spike in the GBNV and number of occurrences in the Webcorp²

Food idiom/Expressions ³	First record (%)	Tallest Spike (%)	Latest record (%)	Webcorp
sell like hotcakes/hot cakes	0.0000005996	0.0000023032	0.0000009362	83
walk on eggs/eggshells	0.0000005732	0.0000043996	0.0000016117	69
upset the apple-cart/apple cart	0.0000006810	0.0000044341	0.0000017583	60
a/no piece of cake	0.0000015499	0.0000296017	0.0000212620	58
a hard nut to crack	0.0000006048	0.0000062144	0.0000026118	44
a (pretty) kettle of fish	0.0000015515	0.0000086885	0.0000022535	43
a cake-eater/cake eater	0.0000002858	0.0000010682	0.0000000257	40
get out of a jam	0.0000003597	0.0000003996	0.0000002390	29
handle the hot potato	0.0000004702	0.0000004702	0.0000000216	15
not cut the mustard anymore	<u>0.0000000278</u>	<u>0.0000001200</u>	<u>0.0000000000</u>	14
butterfingers	0.0000002028	0.0000016333	0.0000004055	9
have a hot potato	0.0000001782	0.0000001941	0.0000000108	<u>3</u>

The analysis of the data revealed that, from the 16 idioms searched, 3 did not show any results (**left with * hot potato; have a lemon on your hands; and give * the/a hot potato**), a result similar to the number of occurrences generated by the Webcorp (4; 4; and 1, respectively). That does not necessarily mean they were not used at all, but that their frequency of use may have been lower than the 40 records necessary to be charted

² The highest results are in **bold**, and the lowest, underlined.

³ The frequency of use percentage from the GBNV includes non-idiomatic expressions.

by the tool. Nevertheless, the lower frequency can be a sign that these idioms are on the process of becoming obsolete. One of the idioms analysed, **a small beer**, showed a high frequency of use, but, after checking the sentences in which it appeared, we noticed that its use was not idiomatic in any (*e.g.* a small beer garden), so it was not included in the table, along with the 3 others that generated no results, afore mentioned.

The highest first record percentage found was for the idiom **a (pretty) kettle of fish**. The lowest first record was for the idiom **not cut the mustard anymore**. This idiom also had the lowest latest record and lowest tallest spike. The highest latest record and tallest spike were, by far, for **a piece of cake**, but that included a large percentage of non-idiomatic sentences (31.7%). On the other hand, **no piece of cake** had only 8% of non-idiomatic use. Some idioms had all, or nearly all, of the sentences in which they appeared with idiomatic use. A possible explanation for that may be the level of idiomaticity. In total, we analysed 1,517 sentences/books. From these, 75.4% were idiomatic, 22.3% were non-idiomatic, and 2.2% could not be accessed. The results from GBNV, for both the highest and lowest percentages, seem to be corroborated by the number of occurrences generated by the Webcorp, taking into consideration that these include only the occurrences where we identified idiomatic use. Although we can identify the (non) idiomatic use of each idiom, we cannot subtract it from the graphs.

Concerning the years, the idiom with the oldest first record was **a pretty kettle of fish** (1806). The one with the most recent first record was **not cut the mustard anymore** (1968). **Walk on eggshells** was the idiom with the most recent tallest spike (2007), so still probably highly used; while **walk on eggs** had its tallest spike much earlier (1843). The idiom with the oldest tallest spike was **a kettle of fish** (1824). The idioms whose frequency of use was falling in 2008 were **sell like hotcakes**, **a/no piece of cake**, **walk on eggshells**, **get out of a jam**, **butterfingers**, **a cake-eater/cake eater** and **upset the apple cart**. The idioms that showed a tendency to rise in frequency in 2008 were **sell like hot cakes**, **walk on eggs**, **have a hot potato**, **a hard nut to crack**, **a (pretty) kettle of fish**, **handle the hot potato** and **upset the apple-cart**.

4 A final view

The data from the GBNV show results similar to those generated by the Webcorp, as far as the frequency of use of the 16 idioms is concerned: 3 idioms did not show any results, 1 showed only non-idiomatic results, and 1 had 0% of frequency in 2008, showing that they might be obsolete or in the process of becoming so. The other idioms presented different percentages of use, half of them plotted a decrease of use in the last years, and half plotted an increase.

Similarly to the Webcorp, the limitations concerning the use of the GBNV are that the results include sentences where the n-grams searched are not used idiomatically, making it necessary to check each sentence/book; and many of the examples come from dictionaries - therefore not necessarily an example of the idiom in use, but its explanation. In addition, if idioms are larger than 5 words, the search can become more complex. Nevertheless, GBNV was found to be an appropriate tool to analyse the frequency of use of idioms and to identify a possible process of obsolescence.

References

1. Camargo, S. e Steinberg, M. Dictionary of Metaphoric Idioms English-Portuguese. E.P.U., (1990).
2. _____. Dicionário de Expressões Idiomáticas Metafóricas Português-Inglês. E.P.U., (1989).
3. Cohen, D.: Initial thoughts on the Google Books N-gram Viewer and datasets. <http://www.dancohen.org/2010/12/19/initial-thoughts-on-the-google-books-ngram-viewer-and-datasets/#comment-67769>, last accessed on 2017/06/10.
4. Davies, M.: Making Google Books n-grams useful for a wide range of research on language change. John Benjamins Publishing Company (2014).
5. Google Books Homepage, <https://books.google.com/ngrams/info>, last accessed on 2017/06/15.
6. Liu, W. and Shen, H.: CiteSpace II: Idiom Studies Development Trends. Journal of Arts and Humanities (JAH), 2 (2), March, 85-97 (2013).
7. Webcorp Homepage, <http://www.webcorp.org.uk/live>, last accessed 2017/08/10.
8. Weiss, A.: Google Ngram Viewer. California State University, Northridge (2015). <http://www.scholarworks.csun.edu/bitstream/handle/10211.3/173281/NGram-v2-1.pdf;sequence=1>, last accessed on 2017/06/04.
9. Zhang, S.: The pitfalls of using Google ngram to study language. <https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/>, last accessed on 2017/05/29.

The Effect of Learner Variables on Phraseological Proficiency

Kathrin Kircili

Justus Liebig University, Giessen, Germany

Kathrin.Kircili@anglistik.uni-giessen.de

Abstract. This paper is based on an empirical study which investigated the effects of learner variables on phraseological proficiency, a topic which has so far not received a lot of attention in phraseological research. By means of a questionnaire enquiring a variety of personal details as well as testing the participants' theoretical as well as their productive and receptive knowledge with regard to phraseology in general and collocations and phrasal verbs in particular, it was found that both longer periods abroad and a regular exposure to the English language have a very positive effect on a learner's practical abilities in this field, although, in direct comparison, the former is still the most effective way to improve a learner's phraseological proficiency – particularly when it comes to phraseological units that are highly idiomatic or easily confusable with a learner's native language.

Keywords: phraseology, EFL, language proficiency

1 Theoretical Background

When acquiring a foreign language, it is the ultimate aim of most students to develop a native-like knowledge of the respective L2. This encompasses proficiency not only with regard to grammar but also concerning vocabulary, since a rich vocabulary knowledge is the most important basis to enable communication. However, the mere acquisition of individual words does often not suffice, firstly, because the context in which a word is used has an influence on its meaning, and secondly, because “much of communication makes use of fixed expressions memorized as formulaic chunks” [2]. These so-called *phraseological units*, which comprise more than 55 percent of spoken and written English [3], often give away non-native speakers and pose a major difficulty – even for advanced learners of English. So far phraseological research in learner languages has mainly concentrated on the connection between a learner's L1 and his or her use of phraseological phenomena such as phrasal verbs or collocations in the English language. As far as the former are concerned, it has been revealed that it is particularly learners whose L1 does not contain phrasal verbs who either refrain from using them and prefer one-word synonyms instead [1] or show a higher frequency of erroneous uses than those learners in whose L1 the phenomenon is actually existent in a similar fashion [9]. With regard to collocations, learners are particularly likely to be influenced by their L1, which has, in previous research, been observed in

up to 53% of the occurrences [7]. This either results in word-for-word translations from the L1 into the L2 or in an attempt to adapt a word combination to the L2 to a certain extent [7; 8]. The effect of learning context variables on phraseological proficiency, however, is a topic that has mainly been neglected so far [5]. Although, in second language acquisition, it is a known fact that learner variables have an influence on a learner's overall performance, they "suddenly become a homogeneous group" [10] when it comes to the investigation of their phraseological proficiency. The empirical study this paper is based on broke with this habit and aimed at the investigation of learner variables and their effects on a student's proficiency in a linguistic field that is particularly advantageous for learners of English because it increases their overall fluency [5] due to a decrease in processing efforts [4; 7] and a more efficient retrieval of word combinations [6]. Consequently, the aim of this study was to find out which personal factors actually influence the productive and receptive knowledge of certain phraseological phenomena.

2 Database

The questionnaire used was comprised of three parts. The first section served to obtain personal information, enquiring the participants' age, gender, field of studies, experience abroad as well as the frequency of exposure to and use of the English language outside the university context. In the second part, the theoretical knowledge of phraseology was tested by asking for definitions and examples of the phenomena relevant for the study. The final part consisted of 40 sentences in which the productive and receptive knowledge of verb/noun as well as adjective/noun collocations and phrasal verbs was tested, all of which were made up of high-frequency adjectives or verbs such as DO, HAVE or MAKE. In the production tasks, participants were asked to fill in the blanks while in the reception task it had to be decided whether the underlined part in a given sentence was correct or incorrect. Examples included:

1. Verb/noun collocations

- a. Production: It won't _____ you any harm if you help her. (*do*)
- b. Reception: She got a baby when she was sixteen years old. (*incorrect*)

2. Adjective/noun collocations

- a. Production: She knows a lot about animals and the environment, but she only has _____ knowledge of computers. (*little*)
- b. Reception: He is a strong smoker. He smokes at least 25 cigarettes a day.
(*incorrect*)

3. Phrasal verbs

- a. Production: Most people would love to do _____ with Mondays.
(*away*)
- b. Reception: Little Jason was very proud when his teeth finally fell out.
(*correct*)

A total number of 161 EFL learners participated in the survey. Some of the results yielded will be presented in the following.

3 Findings

Fig. 1. illustrates the performances of students who had spent a period of more than six months in an English-speaking country compared to those without any experience abroad.

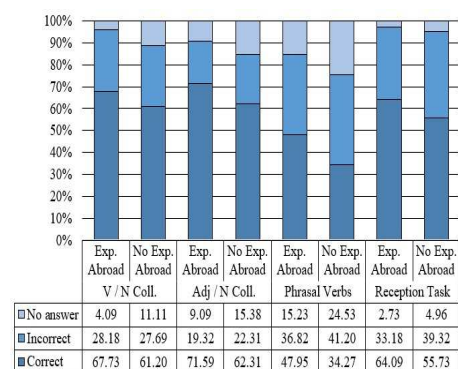


Fig. 1. Results of 44 students with and 117 without experience abroad

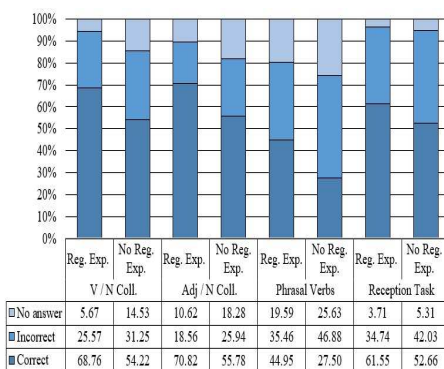


Fig. 2. Results of 97 students with and 64 students without regular exposure to L2

In fact, the differences in each of the four question groups were highly significant ($p < 0.001$),¹ with the most striking result having been determined in the phrasal verb exercises. This is particularly interesting since this kind of PU is mostly used in informal contexts and is therefore more frequent in speech than in writing. Hence, the results prove that longer periods abroad, during which EFL learners have the opportunity to talk to native speakers on a daily basis, make a considerable difference in an area that is otherwise so troublesome for learners of English. Likewise, students with experience abroad were much less likely to fall prey to false friends such as *sleep in* or to confuse common German collocations with English ones (e.g. *black* vs. *blue eye*; *have* vs. *get a baby*). As Fig. 2 above illustrates, similarly clear results could be determined for students who indicated a regular exposure to the English language outside university by reading and writing English texts, watching TV or talking to native speakers on a regular basis. In total, 97 participants met the stipulated requirements of at least two activities being performed daily or weekly. Again, the differences between the two groups of participants were highly significant in all of the four question groups (for all $p < 0.001$). Although it seems natural for an EFL learner to improve his or her language skills when regularly in contact with a language, it is particularly interesting here that the p-values actually indicate that, even though all of them are highly significant, the measured differences between these two groups of participants were even bigger than those determined between students with and those without experience abroad. This holds particularly true for the tasks on V/N

¹ The p-values were determined by means of a t-test.

collocations as well as the phrasal verb exercises. Now, due to the clear results of the first two comparisons, the final analysis sought to establish a connection between them by taking a look at the students who had not been abroad but indicated a regular exposure to the L2 and comparing them with those who had been in an English-speaking country for a longer period of time.

As Fig. 3 indicates, it was found that a regular exposure can result in a comparable proficiency, at least as far as the collocation tasks were concerned in case of which there was no longer a significant difference with p-values above the significance threshold of 0.05 for both V/N and A/N collocations.

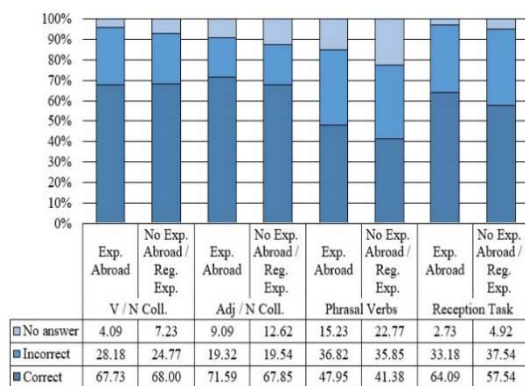


Fig. 3. Results of 44 students with experience abroad and 65 students without but with a regular exposure to the L2

With regard to phrasal verbs and the reception tasks, however, the results were again significant ($p < 0.001$ and $p < 0.05$, respectively) particularly with regard to idiomatic PUs as well as false friends, such as *take after*, *sleep in* or *get a baby*.

4 Conclusion and Outlook

This survey revealed that a longer period spent in an English-speaking country is indeed the most effective way to improve one's phraseological proficiency. However, even though the acquisition of a language in its native environment is inevitable if students aim for a native-like proficiency and wish to master even highly idiomatic structures and reduce the influence of their L1, a regular spare time exposure to the language also proved to be a valid way to enhance one's phraseological competence. Moreover, the results also showed that the generally high frequency of a certain verb does not necessarily imply that their various functions in a collocation or phrasal verb are equally known to EFL learners. It becomes clear that the acquisition of the phenomena is strongly connected to language contact although their complexity would justify an emphasis in school curricula as well. Since this paper only focused on small number of factors that influence advanced learners' phraseological proficiency, future analyses that focus on effects of other learning context variables such as age, gender or the field of studies might also yield interesting results. Additionally, a comparable study focusing on students' phraseological abilities in speech would help to uncover

register differences in learners' phraseological proficiency. Since the questionnaire did not have a time limit and the participants were free to take as much time as they needed, it does not necessarily shed light on the EFL learners' spontaneous reaction. Spontaneity and the ability to use language in speech, however, tells a lot about a speaker's true proficiency and should therefore be included in further investigations.

References

- [1] Dagut, M., Laufer B.: Avoidance of phrasal verbs: A case for contrastive analysis. *Studies in Second Language Acquisition* 7(1), 73-79 (1985).
- [2] Ellis, N.: Phraseology: The periphery and the heart of language. In: Meunier, F., Granger, S. (eds.) *Phraseology in Foreign Language Learning and Teaching*, pp. 1-13. John Benjamins, Amsterdam / Philadelphia (2008).
- [3] Erman, B. Warren B.: The idiom principle and the open choice principle. *Text and Talk – An Interdisciplinary Journal of Language, Discourse & Communication Studies*, 20(1), 29-62 (2000).
- [4] Fiedler, S.: *English Phraseology: A Coursebook*. Narr Francke Attempto Verlag GmbH, Tübingen (2007).
- [5] Granger, S.: From phraseology to pedagogy: challenges and prospects. In: Herbst, T., Faulhaber, S., Uhrig, P. (eds.) *The Phraseological View of Language: A Tribute to John Sinclair*, pp. 123-146. Walter de Gruyter GmbH, Berlin (2011).
- [6] Nattinger, J., DeCarrico J.: *Lexical Phrases and Language Teaching*. 2nd edn. Oxford University Press, Oxford (2001).
- [7] Nesselhauf, N.: *Collocations in a Learner Corpus*. John Benjamins, Amsterdam / Philadelphia (2005).
- [8] Paquot, M.: Exemplification in learner writing: A cross - linguistic perspective. In: Meunier, F., Granger, S. (eds.) *Phraseology in Foreign Language Learning and Teaching*, 101-119. John Benjamins, Amsterdam / Philadelphia (2008).
- [9] Sjöholm, K. *The Influence of Cosslinguistic, Semantic, and Input Factors on the Acquisition of English Phrasal Verbs: A Comparison between Finnish and Swedish Learners at an Intermediate and Advanced Level*. Åbo Akademi University Press, Åbo (1995).
- [10] Wray, A.: *Formulaic Language and the Lexicon*. Cambridge University Press, Cambridge: Cambridge (2002).

Synonymy Between Theory and Practice: The Corpus-based Approach to Determining Synonymy in Lexicographic Descriptions in Croatian

Goranka Blagus Bartolec^[0000-0002-3577-7026]

Institute of Croatian Language and Linguistics, 10 000 Zagreb, Croatia
gblagus@ihjj.hr

Abstract. The aim of this research is to determine the degree of synonymy among words based on a corpus search in hrWaC 2.0 (Croatian Web Corpus, Version 2). First, we analyze the lexicographic determination of absolute synonymy at the level of the dictionary definition (where the independent meaning of the words is extracted from usage context). The major part of the research focuses on the usage of synonyms based on corpus data. The analysis includes two pairs of synonymous nouns in Croatian: *krevet* and *postelja* ‘bed’ and *bijes* and *gnjev* ‘rage’. The ten most common multiword units (MWUs) of these synonyms were selected as attested in the corpus according to raw frequency. According to the obtained results, we determined how many synonyms are distanced from each other when compared to other words, whereby these words obtain the attributes of partial synonyms instead of absolute synonyms. Despite the fact that the context permits the use of both synonyms, it was confirmed that only one is significantly prevalent in some MWUs. The corpus-based approach imposes a different manner of defining absolute synonyms in dictionaries – apart from describing the basic semantic features of the word, the definition must consider the contextual representation and the collocational perspective of synonymous words. Corpus results can improve synonymous relationships among words in lexicographic descriptions.

Keywords: Corpus-based Approach to Synonymy, Croatian, Lexicographic Synonymy, Multiword Units

1 Synonymy as a lexicographic feature

Determining synonymous relationships in dictionary descriptions is a fundamental task of every lexicographer, one which contributes to both the semantic features of words and connecting words that have a different signifier, but the same signified. At the level of the dictionary definition many words can be described with the same definition as absolute synonyms, which presupposes that they should be interchangeable in all meanings and usage contexts [1: 268], [2: 107]. Three contemporary dictionaries of Croatian [3], [4], [5] show many synonyms, i.e. words that are equal or similarly defined. At the cognitive level [1: 270], if we only consider certain semantic values, we can agree that some Croatian words behave as absolute synonyms (eg. *bedro* and *natkoljenica* ‘thigh’). When words from the dictionary are put into a real context

and co-occur with other words in different multiword units (MWUs), there is a gap between absolute synonyms and their behaviour in practice. Many synonyms, even if the context allows, cannot be replaced by a synonymous mate with some collocates. At the level of contextual representation, it is more precise to discuss partially synonymous relationships between words [1: 285], [2: 107], [6: 60–61] because, although they have the same meaning, they are not absolutely substitutable in all contexts. Corpus tools [7], [8] enable exhaustive searches that provide statistical and other data about the usage potential of synonyms in the same context. The following analysis focuses on this aspect.

2 The corpus-based approach to synonymy on the collocational level

According to [9: 122], corpus searches allow synonyms to be observed from two perspectives: (i) the co-occurrence approach, which implies the collocation potential of individual words at the MWU level, (ii) the substitution approach, which determines the degree of synonymy depending on the substitutability of a word in a specific context. These two approaches actually include frequency, collocability, and preferences by style and type of text [10: 20]. Two synonym pairs were selected for analysis: the concrete (countable) nouns *krevet* and *postelja* ‘bed’ and the abstract (uncountable) nouns *bijes* i *gnjev* ‘rage’. The analysis includes the 10 most common MWUs selected by raw frequency. As open class words, the most frequent verbs, adjectives, and nouns were chosen as collocation candidates, while conjunctions, prepositions, and pronouns, as closed class words, are not taken into account because they are not considered semantic constituents of MWUs.

2.1 The synonymous pair *krevet/postelja* ‘bed’

The synonyms *krevet* and *postelja* usually collocate with adjectival collocates on the left. The ratio of the lemma *krevet* to the lemma *postelja* in hrWaC is 85,994 to 9,179, which indicates that the lemma *krevet* appears about 9.4 times more often than the lemma *postelja*. This result clearly shows that the noun *krevet* prevails in everyday language use in many different contexts in Croatian. From the aspect of substitutability, the noun *postelja* is not used in the same context as the noun *krevet*. A search for stable MWUs with adjective collocation candidates was carried out with CQL using regular expressions [tag="A.*"] [lemma="krevet/postelja"]. We chose the ten most frequent adjectives that classify the noun by type or purpose. Adjectives that describe the quality of the noun or possessive adjectives (eg. *velik* ‘large’, *nov* ‘new’, *isti* ‘same’), *tuđ* ‘foreign’, *vlastit* ‘one’s own’, *cijeli* ‘whole’, *topao* ‘warm’, *prazan* ‘empty’, *mek* ‘soft’, *zajednički* ‘shared’, *udoban* ‘comfortable’) were excluded from the analysis, as they are components of free combinations, not of MWUs [11, 12].

Table 1. Frequency distribution of the lemmas *krevet* and *postelja* and their first ten adjective collocation candidates in hrWaC. Common MWUs are in bold.

Query A. *, <i>krevet</i>	total: 12,039	Query A. *, <i>postelja</i>	total: 3,111
bračni krevet 'double bed'	1,596	bolesnička postelja 'sick bed'	492
bolnički krevet 'hospital bed'	950	samrtna postelja 'deathbed'	377
bolesnički krevet 'sick bed'	692	bračna postelja 'double bed'	273
vodeni krevet 'water bed'	223	bolnička postelja 'hospital bed'	204
pomoćni krevet 'additional bed'	165	samrtnička postelja 'deathbed'	101
slobodan krevet 'free bed'	154	hotelska postelja 'hotel bed'	98
hotelski krevet 'hotel bed'	153	turistička postelja 'tourist bed'	63
dječji krevet 'cot'	133	slobodna postelja 'free bed'	37
dodatni krevet 'extra bed'	102	privatna postelja 'private bed'	17
francuski krevet 'french bed'	84	registrirana postelja 'registered bed'	15

The search for the lemmas *krevet/postelja* showed the following: the lemma *krevet* appears with adjectives almost four times more often than lemma *postelja* [Table 1]. The ratio of matching and non-matching of adjective collocates for both lemmas is 50%. Common adjectives with lemmas *krevet* and *postelja* are *double* (lit. 'marriage'), *hospital*, *sick*, *free*, and *hotel*, and no others match. The substitutability criterion for the nouns *krevet* and *postelja* is not applicable with the adjectives *watery* and *French*. The corpus search showed that only *vodeni krevet* 'water bed', lit. 'watery bed' and *francuski krevet* 'French bed' appear as MWUs, and there are no matches for the MWUs *vodena postelja* i *francuska postelja*. It should be noted here that we specifically searched for the MWU *krevet/postelja na kat* 'bunk bed', lit. 'multi-storey bed'. The MWU *krevet na kat* appears 553 times, and the MWU *postelja na kat* appears 7 times. The MWU *postelja na kat*, according to the corpus results, appears in regionally marked texts (which offer tourist accommodation in the Croatian region of Dalmatia), and it is not affirmed in general language use. Including both the co-occurrence approach and substitution approach, the noun pair *krevet* and *postelja* are only partial synonyms, as the noun *krevet* prevails in usage and shares only 50% of the most frequent adjective collocates with the noun *postelja*. The noun *postelja* is limited to some contexts that reflect stylistic use (*samrtnička/samrtna postelja* 'death-bed') or regional/local language use (*slobodna postelja* 'free bed', *postelja na kat* 'bunk bed'), but the noun *krevet* prevails in neutral or ordinary language use (*slobodan krevet* 'free bed', *krevet na kat* 'bunk bed').

2.2 The synonymous pair *bijes/gnjev* 'rage'

The three Croatian dictionaries reviewed describe the abstract nouns *bijes* and *gnjev* 'rage' as absolute synonyms with the meaning of 'a very strong/great state of displeasure and anger'. Compared to the synonymous pair *krevet/postelja*, the synonyms *bijes* and *gnjev* have almost double the raw frequency, but since they are abstract meanings, they do not have strong collocation potential for MWUs. The corpus search was not limited to regular expressions, rather the first ten collocation candidates on the left (-1) and on the right (1) were chosen from the lemmas *bijes* and *gnjev*. Closed class words have been removed from the list, and adjectives, nouns, and verbs are taken into consideration.

Table 2. Frequency distribution of the lemmas *bijes* and *gnjev* and their first ten collocation candidates in hrWaC. Common MWUs are in bold.

Query <i>bijes</i> 'rage'	total: 22,189	Query <i>gnjev</i> 'rage'	total: 6,004
izazvati bijes 'to cause rage'	620	Božji gnjev 'God's rage'	295
izljev bijesa 'outburst of rage'	567	izazvati gnjev 'to cause rage'	172
navući bijes 'to fill with rage'	368	pravedan gnjev 'just rage'	137
napad bijesa 'attack of rage'	339	pravednički gnjev 'righteous rage'	122
ispad bijesa 'outburst of rage'	298	navući gnjev 'to fill with rage'	102
nalet bijesa 'flurry of rage'	229	gnjev javnosti 'rage of the public'	71
iskaliti bijes 'to wreak rage on sb.'	194	velik gnjev 'big rage'	71
bijes javnosti 'rage of the public'	182	opravdan gnjev 'justified anger'	63
napadaj bijesa 'attack of rage'	180	izljev gnjeva 'outburst of rage'	59
provala bijesa 'outbreak of rage'	124	pun gnjeva 'full of rage'	59

The substitutability of the synonyms *bijes* and *gnjev* in the same collocation context according to the obtained results [Table 2] is 40% (*izazvati* 'to cause', *navući* 'to fill with', *javnost* 'public', *izljev* 'outburst',). The lemma *bijes* often co-occurs with verbs (*cause*, *wreak*, *fill with*) and nouns (*outburst*, *flurry*, *attack*, *outbreak*) which specifies a process or action. The noun *gnjev*, in addition to verbs and nouns with meanings of action, usually co-occurs with adjectives or nouns that denote the holder of rage (*God's*, *righteous*, *the public*) which indicate stylistic (ethical or religious) contexts of use. The noun *bijes* has been expanded in ordinary, non-stylistic use.

3 Conclusion: Absolute synonymy on the collocational level?

The corpus-based approach indicates stylistic and regional limitations in the use of some words that dictionaries define as absolute synonyms. However, the corpus does not provide an answer to the question as to why synonyms are not substitutable in all contexts. Is contextual non-substitutability simply a matter of the communications habits, or is synonymous substitution truly impossible in some contexts? In order to answer this question, in addition to the semantic potential of synonymous pairs, it would be necessary to explore the semantic potential of the most common collocates and to determine why only one synonymous mate co-occurs with some collocates while the other does not have the same potential. The lexicographer should not take corpus results for granted, which implies that care should be taken in defining absolute synonyms. The corpus suggests that (i) absolute synonymy, which is based on the cognitive principle of determining "certain semantic properties in common" [1: 270], should be distinguished at the level of dictionary definitions so that several words can have the same definition in dictionaries, and (ii) partial synonymy depends on both the communication context and the collocates with which synonymous words most often co-occur. Given that stable collocations should confirm the definitions of words in the dictionary, the results of these corpus searches evidently show that absolute synonymy is a very limited lexical phenomenon [2:107], and that it is difficult to associate two dictionary entries that match in both their definitions and in their frequent MWUs.

Acknowledgments

This paper is written within the research project Croatian Web Dictionary – MREŽNIK (IP-2016-06-2141), financed by the Croatian Science Foundation. The research presented in this paper is mostly based on the lexicographic sources of the Institute for Croatian Language and Linguistics.

References

1. Cruse, A.: Lexical semantics. Oxford University Press, Oxford (1986).
2. Edmonds, Ph., Hirst, G.: Near-synonymy and lexical choice. *Computational Linguistics* 28(2), 105–144 (2002)
3. Birtić, M. et al.: Školski rječnik hrvatskoga jezika. Institut za hrvatski jezik i jezikoslovlje – Školska knjiga, Zagreb (2012).
4. Hrvatski jezični portal, <http://hjp.znanje.hr/>.
5. Rječnik hrvatskoga jezika. Šonje, J. (eds.), Leksikografski zavod „Miroslav Krleža“ – Školska knjiga, Zagreb (2000).
6. Lyons, J.: *Linguistic Semantics: An Introduction*. Cambridge University Press, Cambridge (1995).
7. Ljubešić, N., Erjavec, T. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In: Habernal, I., Matousek, V. (eds.) *Text, Speech and Dialogue 2011*, pp. 395–402. Springer, Heidelberg (2011).
8. Ljubešić, N., Klubička, F. {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In: Bildhauer, F., Schäfer, R. (eds.), *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pp. 29–35. Association for Computational Linguistics, Gothenburg: (2014).
9. Gries, S., Otani, N.: Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal* 34, 121–150. (2010).
10. Kjellmer, G.: Synonymy and corpus work: On *almost* and *nearly*. *International Computer Archive of Modern and Medieval English* 23, 19–28 (2003).
11. Apresyan, Y.: *Leksicheskaya semantika. Sinonimicheskie sredstva yazyka*. Nauka, Moscow (1974).
12. Mel’čuk, I.: Collocations and Lexical Functions. In: Cowie, A. P. (eds.) *Phraseology: Theory, Analysis and Applications*, 23–53. Clarendon Press. Oxford (1998).

A Lexical Database for the Analysis of Portuguese MWEs

Sandra Cristina dos Santos Antunes

Center of Linguistics of the University of Lisbon, Lisbon, Portugal
sandra.antunes@gmail.com

Abstract. We present an Access database that will help to analyze and classify Portuguese MWEs according to their linguistic properties. We used a subset of an existing lexicon of Portuguese MWEs that was extracted from a corpus using Mutual Information as a statistical measure, followed by manual validation. The lexicon is organized in a three-level structure: the main lemmas, the group lemmas and the variants of those groups. This subset was imported to the database together with linguistic information about the MWEs and their variants (morphosyntactic structure, syntactic category, frequency, MI, grammatical function, discursive function, etc.). A semantic and syntactic fine-grained typology was established, and, by selecting a particular MWE, we can classify it with respect to its degree of semantic decomposability and syntactic transformation. The database is highly customizable and enables the addition/deletion of semantic or syntactic categories considered important throughout the analysis. In the end, all the information of the database will be exported to a XML format, resulting in a lexicon enriched with linguistic information.

Keywords: Access database, classification of MWE, enriched lexicon.

1 Introduction

This paper presents an on-going work that aims at the analysis, classification and annotation of Portuguese lexical multiword expressions (MWEs). This term is understood as an umbrella term, since it includes any sequence of two or more graphic words that present a high syntactic and/or semantic cohesion, embracing, therefore, different types of word combinations (collocations, compounds, formulae, light-verb constructions, idioms, similes, sayings, clichés, etc.).

It is widely known that these expressions play a crucial role in language ([7], [12], [16], [18]). But, despite the fact that their analysis is being intensely carried out in several linguistic areas, they still pose problems regarding their accurate identification. It is not easy to distinguish different types of expressions, which has been proven by the large quantity of typologies in the literature ([2], [5], [9], [10], [15], [18], [22]).

Considering the Portuguese language, there is a major gap regarding the proper classification of these expressions. In order to fill that gap, we developed a database in Access format that imports a subset of a Portuguese MWEs lexicon (COMBINAPT) and enables the classification of each expression according to its linguistic properties. In the application, each MWE is described according to a comprehensive set of

semantic and syntactic categories. The available set of search functions, using filters, enables the extraction of linguistic regularities.

This paper will briefly present the compilation of the lexicon and the methodology adopted for the MWEs selection (section 2), the database and the criteria used for the MWEs classification (section 3), and the practical applications of the tool (section 4).

2 The lexicon

For this work, we used COMBINA-PT, which is a lexicon of significant lexical word combinations of European Portuguese¹. The MWEs were extracted from a balanced² 50 million word written corpus³ using Mutual Information (MI) [4] as a statistical association measure⁴, followed by manual validation. Table 1 presents the corpus constitution from which the MWEs were extracted.

Table 1. Corpus constitution.

Newspapers	29,344,736
Books	10,917,889
Magazines	7,500,500
Miscellaneous	1,851,828
Leaflets	104,889
Supreme Court Verdicts	313,962
Parliament Sessions	277,586
TOTAL	50,310,890

As described in [17], n-grams of 2, 3, 4 and 5 tokens were extracted from the corpus. Sequences constituted by 3 to 5-grams are contiguous, while 2-grams sequences can be either contiguous or separated by a maximum of 3 tokens.

Considering the large candidate list extracted from the corpus (1.7 million MWEs), it was necessary to hand-check only a subpart of the groups. Following previous studies ([6], [19]), the team firstly selected groups with MI values between 8 and 10, since there is a higher concentration of good candidates around those values. Throughout manual validation, we followed several criteria upon which usually relies the definition of a MWE: lexical and syntactic fixedness, semantic cohesion, frequency of occurrence (which reveals sets of favoured co-occurring forms) and grammatical constituency (only completed constituents were selected).

¹ The lexicon is available at Meta-Share repository: <http://www.meta-net.eu/meta-share>.

² The corpus covers a wide range of textual genres representing, in a proportional way, the language usage.

³ This corpus was extracted from the Reference Corpus of Contemporary Portuguese, a written and spoken monitor corpus, in a total of 311M words: <http://www.clul.ulisboa.pt/en/10-research/713-crpc-reference-corpus-of-contemporary-portuguese>

⁴ The choice of MI relied on the fact that it is reported to differentiate between MWE and non-MWE [28].

The MWEs in the lexicon are organized in order to identify: (i) a main lemma, from which the MWE was selected (e.g., *mistério* ‘mystery’); (ii) a group lemma, i.e., the neutral expression that corresponds to the canonical form (e.g., *revelar o mistério* ‘unravel the mystery’); (iii) all the variants that occurred in the corpus (e.g., *revelar este mistério* ‘unravel this mystery; *o mistério foi revelado* ‘the mystery was unravelled’). Concordances lines for each MWE are also available in KWIC format.

In all, the lexicon comprises 1,180 main lemmas, 14,153 group lemmas and 48,154 variants.

3 The database

In contrast to languages for which there is a wide range of studies regarding MWEs, for Portuguese little work has been done so far. Most of the studies pay particular attention to idiomatic expressions ([3], [14], [28]), compounds ([1], [21], [26]) or light verbs [11], while other types of expressions are classified taking into account their morphosyntactic structure [20].

Given the existence of different types of MWEs (with different degrees of syntactic and semantic cohesion), as well as the difficulty in distinguishing between certain expressions (such as free combinations, collocations and compounds), we created a database that will help to analyze and classify these groups taking into account their lexical, grammatical, syntactic, semantic, pragmatic and discursive properties.

Due to the large amount of COMBINA-PT data, we decided to firstly analyze only a part of the lexicon. At the moment, the database contains 169 main lemmas, 3,230 group lemmas and 1,503 variants. The database contains all the information considered important for the analysis of the MWEs: (i) the morphosyntactic structure; (ii) the syntactic category; (iii) the variants (if any) and their morphosyntactic structure; (v) the frequency of the expression and the variants; (vi) the MI; (vii) the grammatical function; (viii) the discursive function [18]; (ix) the definition (important for idiomatic expressions); (x) a concordance line; (xi) indication of presence/absence of the expressions in a Portuguese reference dictionary [30] (since Portuguese lexicographers do not make full use of corpora (they only use quotations from literary texts), it could be interesting to see to what extent the dictionary entries meet corpus data).

Lemmas that pertain to different grammatical classes are registered in different entries with different numbers: *frio*_{1N}, *frio*_{2ADJ} ‘cold’. The same is true for polysemous MWE, i.e., that can have several senses (such as *sinal verde* ‘green light’, which can be either a traffic sign or an indication of approval): *sinal verde*₁, *sinal verde*₂.

Figure 1, below, shows how the database is structured in a three-level structure (on the left): main lemma, MWEs for each lemma and variants of each MWE. The selection of a main lemma displays the set of related MWE; by clicking on the “+” sign of a MWE, all the variants of that expression will be displayed.

By selecting a particular MWE, all the information regarding that expression will be disclosed (on the right). For the semantic and syntactic analysis, we established a typology, presented in a tree diagram format, with checkboxes that must be selected according to their degree of semantic decomposability and syntactic transformation.

Regarding the semantic analysis, and following [2], [5], [10], [15] and [22], we firstly considered a scale of idiomaticity, composed by three major classes: (i) compositional meaning; (ii) partially idiomatic meaning (at least one of the elements keeps its literal meaning); (iii) total idiomatic meaning. Within each of these three semantic categories, the MWEs are also classified regarding their type: collocations, compounds, formulae, light verbs, idioms, similes, sayings and clichés. A fine-grained distinction can also be made inside some of these types: collocations are subdivided into favoured co-occurring forms and expressions with restricted collocability. This helps to visualize which type of MWE may occur in different semantic categories (collocations, compounds, similes and sayings are spread amongst compositional and idiomatic levels), highlighting the process of lexicalization of some expressions. Although we are trying to draw this dividing line, we are aware that the semantic evaluation of certain expressions may be difficult, not allowing accurate divisions.

From the syntactic standpoint, the observation of the variants of each MWE will help its analysis regarding the lexical and syntactic fixedness [18]. Each expression can be classified as: (i) fixed (no variation); (ii) semi-fixed (nominal/verbal inflection)⁵; (iii) with variation. The variation field is, in turn, subdivided into lexical (permutation, replacement of elements, reduction, etc.), syntactic (passivization, relativization, pronominalization, nominalization, insertion of elements, etc.) and structural (free realizations, exploitations, etc.). It is also known that corpus data may not cover all the possibilities of the language system [24]. For that reason, it is possible to add variants of a MWE that were found in other corpora (duly identified), the internet or through introspection, confirming their lexical-syntactic variation/fixedness.

The database is highly customizable and enables the addition/deletion of semantic or syntactic categories that we may consider important throughout the analysis.

Finally, all the information of the database will be exported to a readable format (XML), resulting in a lexicon enriched with linguistic information ([8], [13], [27]).

4 Applications

This new database will facilitate the encoding of the linguistic properties of MWE, providing a lexicon with several layers of information. As an Access database, it is possible to apply filters and to create reports that enumerate all the criteria that match for any category of MWEs. Filters can be applied to any field and all the combinations are possible. We believe that the resulting lexicon will constitute a value for areas such as: (i) lexicography (the lexicographer can consult the MWEs actually produced by native speakers, and easily observe that different collocates may point towards different meanings of a word [25]); (ii) foreign language acquisition (allowing the production of natural-sounding speech and writing); (iii) natural language processing (helping the development of automatic identification systems); (iv) theoretical linguistics (expediting semantic and syntactic proposals).

⁵ Since Portuguese is a highly inflectional language, practically all the verbs and nouns that occur in MWEs show inflection marks.

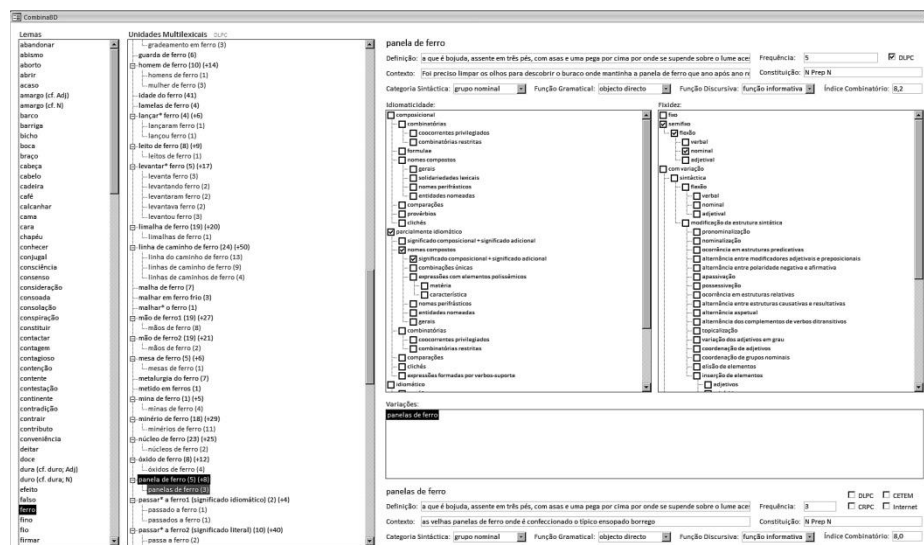


Fig. 1. MWE for the lemma *ferro* ‘iron’ and classification of *panela de ferro* ‘iron cauldron’.

References

- Baptista, J.: Estabelecimento e Formalização de Classes de Nomes Compostos. Master thesis. University of Lisbon, Lisbon (1994).
- Benson, M., Benson, E., Ilson, R.: The BBI Combinatory Dictionary of English: a guide to word combination. John Benjamins Publishing Company, Amsterdam/Philadelphia (1986).
- Chacoto, L.: Estudo e Formalização das Propriedades Léxico-Sintáticas das Expressões Fixas Proverbiais. Master thesis. University of Lisbon, Lisbon (1994).
- Church, K. W., Hanks, P.: Word Association Norms, Mutual Information and Lexicography. In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, pp. 76-83. Vancouver, Canada (1989).
- Cowie, A. P.: Introduction. In: Cowie, A.P. (ed.) Phraseology. Theory, Analysis, and Applications, pp. 1-20. Oxford University Press, Oxford (1998).
- Evert S., Krenn, B.: Methods for the Qualitative Evaluation of Lexical Association Measures. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pp. 188-195. Toulouse, France (2001).
- Fellbaum, C.: An WordNet Electronic Lexical Database. The MIT Press, Cambridge, MA (1998).
- Fellbaum, F., Geyken, A., Herold, A., Koerner, F., Neumann, G.: Corpus-Based Studies of German Idioms and light Verbs. International Journal of Lexicography (19), pp. 349-361 (2006)
- Fernando, C.: Idioms and Idiomaticity. Oxford University Press, Oxford (1996).
- Hausmann, F. J.: Le dictionnaire de collocations. In: Hausmann, F. J., Wiegand, H. E., Zgusta, L. (eds.) Wörterbücher, dictionaries, dictionnaires. Ein international Handbuch zur Lexikographie, pp. 1010-1019. de Gruyter, Berlin (1989).
- Hendricks, I., Mendes, A., Pereira, S., Gonçalves, A., Duarte, I.: Complex Predicates annotation in a corpus of Portuguese. In: Proceedings of the fourth Linguistic Annotation Workshop. Association for Computational Linguistics, pp. 100-108. Uppsala, Sweden (2010).

12. Jackendoff, R.: *The Architecture of the Language Faculty*. The MIT Press, Cambridge, MA (1997).
13. Krenn, B.: CDB – a database of lexical collocations. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens, Greece.
14. Macário Lopes, A. C.: *Texto Proverbial Português: elementos para uma análise semântica e pragmática*. PhD. Dissertation. University of Coimbra, Coimbra (1992).
15. Mel'čuk, I.: Collocations and Lexical Functions. In: Cowie, A. P. (ed.) *Phraseology. Theory, Analysis, and Applications*, pp. 23-53. Oxford University Press, Oxford (1998).
16. Mel'čuk, I.: Phraseology in the language, in the dictionary and in the computer. In Jean-Pierre, C. (ed.) *Yearbook of Phraseology* 3(1), pp. 31-56 (2012).
17. Mendes A., Antunes, S., Bacelar do Nascimento, M.F., Casteleiro, J.M., Pereira, L., Sá, T.: COMBINA-PT: A Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pp. 1900-1905. Genoa, Italy (2006).
18. Moon, R.: *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford Studies in Lexicography and Lexicology. Clarendon Press, Oxford (1998).
19. Pereira, L. A. S., Mendes, A.: An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications. In: *Proceedings of the 10th EURALEX International Congress*, vol. II, pp. 841-849. Copenhagen, Denmark (2002).
20. Ranchhod, I.: O Lugar das Expressões 'Fixas' na Gramática do Português. In: Castro, I., Duarte, I. (eds.) *Razões e Emoção. Miscelânea de Estudos oferecida a Maria Helena Mira Mateus*, pp. 239-254. Imprensa Nacional Casa da Moeda, Lisboa (2003).
21. Rio-Torto, G., Ribeiro, S.: Compounding in contemporary Portuguese. *Probus* 24(1), pp. 119-145 (2012).
22. Sag, I. Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh A. (ed.) *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 1-15. Mexico City, Mexico (2002).
23. Sinclair, J.: *Corpus, Concordance and Collocations*. Oxford University Press, Oxford.
24. Stubbs, M.: *Words and Phrases. Corpus Studies of Lexical Semantics*. Blackwell Publishing, Oxford (2002).
25. Stubbs, M.: A quantitative approach to collocations". In: Allerton, D.J., Nesselhauf, N., Skandera, P (eds.) *Phraseological Units: basic concepts and their applications*. ICSELL 8. Schwabe Verlag Basel (2004).
26. Villalva, A.: *Estruturas Morfológicas. Unidades e Hierarquias nas Palavras do Português*. PhD. Dissertation. University of Lisbon, Lisbon (1994).
27. Villavicencio, A., Copestake, A., Waldron, B., Lambeau, F.: The lexical encoding of MWEs. In: Tanaka, T., Villavicencio, A., Bond, F., Korhonen, A. (eds.) *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*. Barcelona, Spain (2004)
28. Villavicencio, A., Kordoni, V., Zhang, Y., Idiarte, M., Ramisch, C.: Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. In: *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1034-1043. Prague, Czech Republic (2007).
29. Vilela, M.: *Metáforas do Nosso Tempo*. Almedina, Coimbra (2002).
30. *Dicionário da Língua Portuguesa Contemporânea*. Academia das Ciências de Lisboa e Editorial Verbo (2001).

A Contrastive Analysis of Antonymous Prepositional Pairs in Croatian and Russian

Ivana Matas Ivanković^[0000-0002-9796-8346]

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

Abstract. Structures consisting of two prepositional phrases with antonymous prepositions followed by a noun phrase (NP) usually have unique meaning, which provides a reason to treat them as phraseological units. In Croatian and Russian, these structures are *od / om* 'from' + *NP1_{GEN}* + *do / do* 'to' + *NP2_{GEN}*, *s / c* 'from' + *NP1_{GEN}* + *na / ha* 'onto' + *NP2_{ACC}*, *iz / u3* 'out of' + *NP1_{GEN}* + *u / e* 'into' + *NP2_{ACC}*. These are grammatically fixed units, but lexically, they can be placed on a scale from fixed to flexible. On one side of the scale, the form is productive and its members are not lexically restricted. On the other side, they come as lexically filled idioms that cannot be translated word for word. The goal of this paper is to examine the similarities and differences between units with specific syntactic structure in Croatian and Russian, two languages that are genetically related.

Keywords: Antonymous Prepositional Pairs, Croatian, Russian.

1 Introduction

This paper compares syntactic structures in Croatian and Russian that contain two prepositional phrases with antonymous prepositions that refer to one whole. The first prepositional phrase in this kind of pair indicates a starting point, while the other refers to an ending point. These structures are *od / om* 'from' + *NP1_{GEN}* + *do / do* 'to' + *NP2_{GEN}*, *s / c* 'from' + *NP1_{GEN}* + *na / ha* 'onto' + *NP2_{ACC}*, *iz / u3* 'out of' + *NP1_{GEN}* + *u / e* 'into' + *NP2_{ACC}*. This type of unit has been analysed in Croatian [1] and Russian [2, 3], but a detailed comparison of these constructions in the two languages has not yet been performed, especially for idioms. The aim of this research is to establish the similarities and differences between these structures in Croatian and Russian, as such complex structures, if taken for granted, can cause difficulties in second language learning and in translation.

The paper is structured as follows: first, the grammatical properties of the structure are described, then formal idioms are presented, and then substantive idioms are compared. Since there is no parallel Croatian-Russian corpus, the examples in the first

part are taken from the Croatian hrWaC corpus¹ (<http://nlp.ffzg.hr/resources/corpora/hrwac/>) and the Russian National Corpus² (<http://www.ruscorpora.ru/>). The texts were chosen independently but are as similar as possible, meaning that if they were translated from one language to another, the same structure would be used.³

2 Structure

Antonymous prepositional pairs have fixed structure, consisting of two prepositional phrases with antonymous prepositions (*od / om* ‘from’ – *do / do* ‘to’; *iz / u3* ‘out of’ – *u / 6* ‘into’; *s / c* ‘from’ – *na / na* ‘onto’). Formally, prepositional phrases can come separately, but the meaning is incomplete without one of the parts. In the pair *od / om* ‘from’ – *do / do* ‘to’, both nominal phrases following the preposition come in the genitive case. In the pair *iz / u3* ‘out of’ – *u / 6* ‘into’ and *s / c* ‘from’ – *na / na* ‘onto’, the first nominal phrase is in genitive (in Croatian and in Russian, *s / c* can also come with the instrumental), while the second is in the accusative (in Croatian and Russian, the prepositions *u / 6* and *na / na* can also come with the locative case – the accusative indicates a goal). The grammatical structure of these units is determined and fixed, but the nominal phrases that appear in them can be free or fixed.

3 Formal idioms

Fillmore et al. [4: 505] distinguishes between substantive or lexically filled idioms and formal or lexically open idioms. Formal idioms are syntactic patterns dedicated to semantic and pragmatic purposes not knowable from their form alone. Antonymous prepositional pairs with *od / om* ‘from’ – *do / do* ‘to’, *iz / u3* ‘out of’ – *u / 6* ‘into’, *s / c* ‘from’ – *na / na* ‘onto’ are syntactic patterns that can be filled by different noun phrases. As a pair, *od / om* ‘from’ + *NP1_{GEN}* + *do / do* ‘to’ + *NP2_{GEN}* can have (a) spatial meaning of stretching or prolonging: *jedna od najtežih ruta, od Aljaske pa do Ognjene zemlje* ‘one of the most difficult routes, from Alaska to Tierra del Fuego’ / *мышцы ног от бедра до колена* ‘leg muscles from the thighs to the knees’, (b) temporal meaning of lasting: *trajati od kasne zime do kasnog ljeta* ‘last from late winter to late summer’ / *в возрасте от 12 до 17 лет* ‘from the age of 12 to 17’, (c) quantitative meaning: *naklada od 300 do 1000 primjeraka* ‘an edition of 300 to 1000 copies’ / *по цене от 20 до 80 рублей* ‘at the price of 20 to 80 rubles’, (d) the meaning of a range in a broad sense, referring to many constituents belonging to one whole (these noun phrases frame everything in between, whether referring to objects, processes, or something else): *pripremiti sve, od omleta do svježeg soka od naranče* ‘prepare every-

¹The search, used the following regular expressions: [lemma="od"][]{1,3}[lemma="do"], [word="s"&tag="S.*"][tag="....g"]{1,3}[lemma="na"], [lemma="iz"][]{1,3}[lemma="u"].

² In the lexical and grammatical search, the first preposition of the pair was set in the “word” field, while the other one was set in the second “word” field at a distance of 1–3.

³ The first example is in Croatian, while the Russian example follows the slash. The English translation is located in single quotation marks.

thing, from omelettes to fresh orange juice’ / *цветки от розовых до ярко-красных* ‘flowers from pink to bright red’. The preposition *iz / из* ‘out of’ primarily refers to a starting point, but unlike *od / от*, the activity begins in a space that is perceived as closed or bounded. *U / в* ‘into’ has the opposite meaning, referring to a space inside which an action ends. *S / с* ‘from’ with the genitive primarily means that an action has begun on the upper or outer part of an object. *Na / на* ‘onto’ with the accusative means that something or somebody has come to the upper side of something with which it is in contact. Bearing these nuances in mind, as pairs, they have (a) spatial meaning: *vraćanje tekućine iz tkiva u krv* ‘returning fluid from tissue to the blood’ / *экспорт нефти и газа из России в Европу* ‘the export of oil and gas from Russia to Europe’, (b) temporal meaning: *u noći s četvrtka na petak* ‘on the night between Thursday and Friday’ / *перенесение парламентских выборов с декабря на март* ‘the shifting of parliamentary elections from December to March’, (c) a change from one state to another: *prelaziti iz osnovnog u srednje školovanje* ‘transition from primary to secondary schooling’ / *переход из юниоров в мастер-класс* ‘transition from the juniors to the master class’; *prevoditi s hrvatskoga na ukrajinski* ‘translate from Croatian into Ukrainian’ / *переход с приема на передачу* ‘transfer from reception to transmission’. The Russian pairs *от* ‘from’ + NP1_{GEN} + *к* ‘towards’ + NP2_{DAT} (*от мистики к физике* ‘from mysticism to physics’), *с* ‘from’ + NP1_{GEN} + *до* ‘to’ + NP2_{GEN} (*с утра до ночи* ‘from morning till night’), and *с* ‘from’ + NP1_{GEN} + *по* ‘up to’ + NP2_{ACC} (*с марта по декабрь* ‘from March to December’) are also treated as grammatical structures with unique semantics [2, 3], although the prepositions in them are not completely antonymous. These pairs are not typical for Croatian, and would be translated using antonymous pairs (respectively: *od mistike do fizike, od jutra do noći, od ožujka do prosinca*).

Each of these pairs can obtain distributional meaning when the same noun is repeated after the prepositions: *rasprostirati se od srca do srca* ‘spread from heart to heart’, *от берега до берега* ‘from coast to coast’; *skakutati s grane na granu* ‘jump from branch to branch’ / *перепрыгивать с танка на танк* ‘jump from tank to tank’. This manner of pair construction can sometimes be replaced with the preposition *po / по* ‘over’ + noun (*skakutati po granama* ‘jump about the branches’ / *бегать по магазинам* ‘run among the shops’). The repeating noun can have temporal meaning: *iz dana u dan / изо дня в день* (lit. ‘from day into day’) ‘from day to day’. Although nouns can vary, their semantics are restricted to temporal meaning, so they can be treated as substantive idioms.

4 Substantive idioms

Formal idioms can serve as host to substantive idioms. Their lexical makeup is (more or less) fully specified [4: 505]. Based on previous research [1] and searches of the corpus and a dictionary of idioms [5, 6], a list of 34 idioms was made, and Croatian and Russian equivalents were compared. Some idioms can be translated into several equivalents, in which case the most similar equivalent was taken into account. Based on their grammatical and lexical features, the idioms can be fully equivalent, partially equivalent, or non-equivalent. 1) **Fully equivalent idioms** have equivalent

grammatical and lexical structure. Nine idioms are fully equivalent, e.g.: *od glave do pete* (lit. 'from head to heel') / *от головы до пят, пяток* (lit. 'from head to heels') 'from head to foot'; *prelijevati iz šupljeg u prazno* / *переливать из пустого в порожнее* (lit. 'transfuse from hollow into empty') 'to plough the sand'. 2) **Partially equivalent** idioms can be equivalent lexically or grammatically. 2a) Ten idioms are **lexically equivalent**: *ići od Poncija do Pilata* (lit. 'go from Pontius to Pilate') / *ходить от Понтия к Пилату* (lit. 'to walk from Pontius to Pilate') 'to go from pillar to post'; *ići s noge na nogu* (lit. 'go from foot to foot') / *идти нога за ногу* (lit. 'go foot after foot') 'to drag one's feet'. 2b) Two idioms are **grammatically equivalent**: *od jutra do sutra* (lit. 'from morning till tomorrow') / *од темна до темна* (lit. 'from dark till dark') 'around the clock'; *živjeti od prvoga do prvoga* (lit. 'live from the first till the first') / *жить от зарплаты до зарплаты* (lit. 'live from paycheck to paycheck') 'live from paycheck to paycheck'. 3) Thirteen idioms are **non-equivalent**, e.g.: *ni iz džepa ni u džep komu* (lit. 'neither out of one's pocket nor into one's pocket') / *ни пользы ни вреда кому* (lit. 'neither the benefit_{GEN} nor harm_{GEN} to sb') 'to have nothing to gain nor to lose'; *smijati se od uha do uha* (lit. 'laugh from ear to ear') / *смеяться во весь рот* (lit. 'laugh into one's whole mouth') 'smile from ear to ear'.

5 Results of the comparison and conclusion

The analysis has shown that Croatian and Russian are very similar in their use of antonymous prepositional pairs, which is to be expected due to their genetic relatedness. It also shows that prior experience in one language can help in learning the other, but that it can also have a negative influence when using the patterns in the examples that have differing grammatical and/or lexical structure between languages. Fillmore et al. [4] draws a distinction between substantive and formal idioms, however, structures with antonymous prepositional pairs can be placed on a scale from formal idioms to substantive idioms. In lexically open idioms, grammatical structure is equivalent between both languages (*od podruma do krova* – *от подвала до крыши* 'from cellar to roof'). Patterns with the same noun in both pairs are less lexically open (*iz broja u broj* – *из номера в номер* 'from issue to issue'). Distributional patterns with temporal nouns (*iz dana u dan* – *из дня в день* 'day after day', *iz godine u godinu* – *из года в год* 'year after year') are closer to substantive idioms, although the units used with nouns with temporal meaning can vary from smaller to larger ones. In all of these types, syntactic structure can be transferred from one language to another. Unlike in Croatian, the Russian structures *om* 'from' – *к* 'towards', *с* 'from' – *до* 'to', *с* 'from' – *по* 'to' are also idiomatic and can be used in place of antonymous prepositional pairs. Although constitutive prepositions are equivalent in both languages when used separately, they cannot be transferred into Croatian when used as a pair (*переход от системы к системе* > **prijelaz od sustava k sustavu* / *prijelaz sa sustava na sustav* 'transition from system to system'). The substantive idioms in some examples are fully equivalent between the two languages (*s koljena na koljeno* – *с поколения на поколение* 'from generation to generation'), but many are not, and a direct translation from one language to another would result in an error.

Acknowledgements

This paper is written within the research project Croatian Web Dictionary – MREŽNIK (IP-2016-06-2141), financed by the Croatian Science Foundation.

References

1. Kovačević, B., Matas Ivanković, I.: Parni prijedlozi. Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje 33(1), 245–261 (2008).
2. Всеволодова, М. В., Владимирский, Е. Ю.: Способы выражения пространственных отношений в современном русском языке. „Русский язык“, Москва (1982).
3. Всеволодова, М. В. Способы выражения временных отношений в современном русском языке. Издательство Московского университета, Москва (1975).
4. Fillmore, C.J., Kay, P., O'Connor, M.C. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language* 64(3), 501–538 (1988).
5. Menac, A., Fink Arsovski, Ž., Venturin, R.: Hrvatski frazeološki rječnik. Naklada Ljevak, Zagreb (2014).
6. Menac, A., Fink Arsovski, Ž., Mironova Blažina, I., Venturin, R.: Hrvatsko-ruski frazeološki rječnik + Kazalo hrvatskih i ruskih frazema. Knjigra, Zagreb (2011).

An Objective Method of Identifying Teachworthy Multi-word Units for Second Language Learners

James Rogers¹

¹ Meijo University, Nagoya Aichi 463-0012, Japan

jrogers@meijo-u.ac.jp

Abstract. This study aimed to confirm whether a frequency-based approach toward the identification of multi-word units (MWUs) most exemplary of lemma that frequently co-occur would be as reliable as one that relies upon native speaker intuition. This results showed that the frequency-based method produced very similar results to the method that relied upon native-speaker intuition, and thus could be considered more reliable in that potential subjectivity could be avoided.

Keywords: Multi-word unit, concgram, corpora

1 Introduction

Research has shown that MWU knowledge for second language learners improves language processing speed. However learners are often very weak in this aspect of fluency and learning materials do not focus on them. This stems from a lack of resources. In fact, high-frequency collocation and MWU identification is a very complex, time-consuming process in which much research is still needed. This current study will evaluate a method which aims to improve upon the objectivity of one of the steps taken in previous research to identify high-frequency MWUs.

2 Literature Review

Knowledge of MWUs has been found to aid in more efficient language processing (Conklin & Schmitt, 2008). While there is variation in simply defining MWUs, this current study will define them as the MWUs most exemplary of lemmatized concgrams, with a ‘concgram’ being “all the permutations of constituency and positional variation generated by the association of two or more words” (Cheng, Greaves, & Warren, 2006, p. 411).

Despite agreement on the value of MWU knowledge, many studies have shown that learners are not acquiring this fluency (DeCock et al., 1998; Nesselhauf, 2005). This lack of fluency is connected to the fact that learning materials simply do not focus on them (Gitsaki, 1996). This is due to a severe lack of large-scale studies which actually have identified MWUs. Identifying MWUs is a complex and time consuming process, and thus it is difficult to create large-scale comprehensive materi-

als (Rogers, 2017). Thus, much research is still needed simply in developing and evaluating efficient and reliable MWU identification methods.

In Rogers (2017), MWUs were identified by searching for exemplary chunks of high-frequency words and their collocates. Mini corpora were compiled for these, and were then analyzed to extract the most frequent MWU in which the high-frequency pivot word and collocate occurred in. To identify this MWU, native speakers chose the most frequent chunk identified. However, if this core item occurred within any subsequent items and the native speaker judged it to be worthy of teaching, the native speaker could opt to extend the core and have that MWU be the exemplar for the lemmatized concgram instead. For instance, if *come to terms* was identified as the most frequent chunk for the lemma *come/terms*, and the next most common chunk was *come to terms with* and the native speaker felt that this extended version was worth teaching, then it was chosen to represent *come/term*. Extending core MWUs in this way was found to be an essential step in that 53 percent of the sampled items in Rogers' (2017) study were extended upon. However, relying upon native speaker intuition can introduce subjectivity. Thus, a gap in the research exists as to whether or not an objective method can be relied upon for this step.

3 Research Question

Does an objective method of extending exemplary MWUs of lemmatized concgrams beyond their core produce results similar to those using native speaker intuition, and does it affect native speaker intuition-based judgments of teach-worthiness?

4 Procedure

This study evaluated an objective method to extend MWUs of lemmatized concgrams beyond their core in comparison to Rogers' (2017) method which utilized native speaker intuition. It utilized corpus frequency data instead of native speaker intuition to decide whether or not MWUs should be extended or not. These results were compared to the results of Rogers (2017) to determine whether a similar percentage of items were extended or not. The results were examined to determine whether or not extending a MWU beyond its core affects native speaker intuition based judgments of teach-worthiness.

First, the most frequent 500 lemma in Gardner and Davies' (2014) high-frequency academic vocabulary list were utilized as pivot words to search for lemmatized collocates in the academic section of the Corpus of Contemporary American English (COCA) (Davies, 2008). Mini corpora were then compiled for the lemmatized pivot words and collocates from the academic section of the COCA, and these corpora were analyzed to extract the most frequent MWU in which the pivot word and collocate occur in.

Analyzing the data using conventional concordance software such as AncConc (Anthony, 2011) would result in a large amount of noise in each set that would have to be

removed manually. Rogers (2017) solved this issue by utilizing AntWordPairs (Anthony, 2013), a custom piece of software designed specifically for his study. Thus, this same software was utilized in this current study.

This current research is part of a larger study which resulted in approximately 9,000 MWUs. Analyzing such a large amount of data is beyond the scope of what will be explored, and thus a random sample of that data was examined in this study. Data for 100 random pivot words of the initial 500 were utilized.

Collocation implies 'frequent' co-occurrence. This study thus followed Rogers' (2017) parameter of approximately one occurrence per million tokens. However, since this current study is only working with the academic section of the COCA which is one-fifth of the entire corpus, this study's frequency cut-off was set to 100 occurrences or more.

Other researchers have also used relative frequency of co-occurrence compared with the individual words' total independent occurrences to help identify collocations. Church and Hanks (1990) referred to this measure as 'mutual information' (M.I.). Stubbs (1995) and Hunston (2002) both believe that an M.I. statistic of three or higher indicates that two words collocate. However, after analyzing initial data, this cutoff was deemed to be too exclusive, leaving out many clearly useful collocations. M.I. cutoffs of two and one were thus experimented with. An M.I. cutoff of one was deemed too inclusive, while a cutoff of two was found to be most balanced and thus it was used.

After processing with AntWordPairs, results were examined to determine if the core MWU should be extended or not. Any subsequent MWU that still contained the core while having half or more of the frequency of the core was considered the top MWU. Then, the data was further examined to determine if that new MWU identified could be extended any further. Its frequency was used to determine if any other subsequent MWUs contained it and also had half or more of its frequency. This process continued on until it was not possible to continue any longer. MWUs that were extended were then tallied.

Three native speakers then judged the MWUs in regards to their teach-worthiness for learners of academic English. They made a judgment since corpus data can contain noise and does not always produced what the user intends to find. The native speakers were instructed to mark items as not worthy of teaching if they are: odd formulations (noise in the data), proper nouns, too specific to a genre of academic research, have more of a tendency to occur in general English rather than academic English, grammatical formulations that are devoid of meaning, or items that are overly technical.

Then an interrater reliability (IRR) was conducted to assess the degree that coders consistently determined whether the lemmatized concgrams were pedagogically valuable for learners of academic English. Cohen's kappa was utilized to factor out agreement due to chance, and a mean of the kappas for each unique pair of coders was used. Landis and Koch's (1977) scale of interpreting kappa was used to determine the level of agreement. Since ratings may not be evenly distributed across categories, the formula of $IRR = 2(PA) - 1$ (where PA is the probability of agreement due to chance) was used as recommended by Siegal and Castellan (1988). Byrt, Bishop and Carlin (1993) recommend reporting all three kappas when there may be doubt, and thus this was followed.

Finally, the judgments were tallied for items that were extended versus items that were not extended to determine whether or not extending MWUs beyond their core affects native speakers' judgments of teach-worthiness.

5 Results

The 100 pivot words analyzed in this study produced a total of 1,777 concgrams. When the MWUs of these were identified, native speakers deemed 79 percent to be worthy of teaching. 48 percent of these items examined were extended. 41 percent of items judged as being unworthy of teaching were extended, and out of these all cores of that extended item were also judged to be unworthy of teaching.

IRR analysis showed that marginal distributions indicated the presence of both prevalence and, to a lesser extent, bias in the data (Hallgren, 2012). Therefore, following Byrt, Bishop, and Carlin (1993), Cohen's (1960) kappa was supplemented by two other measures of agreement: Siegel and Castellan's (1988) kappa (which guards against bias) and Byrt et al.'s $2(PA) - 1$ (which guards against prevalence). For each of these indices, IRR was computed separately for the three coder pairs then averaged to provide a single IRR index (Light, 1971). Based upon Landis and Koch's (1977) guidelines, IRR was moderate to substantial, $\kappa = .46$; Siegel and Castellan's $\kappa = .45$; $2(PA) - 1 = .62$.

6 Discussion

This study showed potential for improved reliability of utilizing corpus data in comparison to native speaker intuition. In addition, it was revealed that extending an item did not make it more prone to be judged as unworthy of teaching. However, although more objective, the method still needed to be done manually and is time consuming. Future concordance software development certainly should take these issues into consideration.

IRR was shown to be reliable, and the judgements showed that the vast majority of items (78.5 percent) were deemed to be worthy of teaching. However, if only corpus data and the quantitative parameters used in this study were relied upon 21.5 percent of the results would be unworthy of instruction. This large percentage thus points to native speaker intuition still being essential in creating a good quality, ready to use resource.

7 Conclusion

This study found that an objective method utilizing corpus frequency data can be relied upon to produce results similar to what native speakers produce when they use their intuition. Approximately half of the items examined in this study were extended beyond the core top MWU identified using corpus frequency data, which is the nearly the same amount that were extended in a previous study which relied upon native speaker intuition. This study also found that native speaker intuition was shown to be an essential step needed to be taken to remove items deemed not worthy of instruction. While this

method is an improvement since it is more objective, the step still needs to be done manually and is very time-consuming. It is hoped that future concordance software development will take the results of this study into consideration to help improve upon the efficacy and quality of MWU identification.

References

1. Anthony, L.: AntConc (Version 3.2.2) [Computer Software]. Waseda University, Tokyo, Retrieved from <http://www.antlab.sci.waseda.ac.jp/> (2011).
2. Anthony, L.: AntWordPairs (Version 1.0.2) [Computer Software]. Waseda University, Tokyo. Available on request (2013).
3. Byrt, T., Bishop, J., Carlin, J.: Bias, prevalence and Kappa. *Journal of Clinical Epidemiology*, 46(5), 423-429 (1993).
4. Cheng, W., Greaves, C., Warren, M.: From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, 11(4), 411-433 (2006).
5. Church, K., Hanks, P.: Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 76-83 (1990).
6. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46 (1960).
7. Conklin, K. Schmitt, N.: Formulaic sequences: Are they processed more quickly than non-formulaic language by native and nonnative speakers? *Applied Linguistics*, 29, 72-89 (2008).
8. Davies, M.: The corpus of contemporary American English: 425 million words, 1990-present. Available online at <http://corpus.byu.edu/coca/> (2008).
9. DeCock, S., Granger, S., Leech, G. and McEnery, T.: An automated approach to the phrasicon of EFL learners. In: Granger, S. (ed.), *Learner English on computer*, pp.67-79. Longman, London (1998).
10. Durrant, P. Schmitt, N.: To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics*, 47, 157-177 (2009).
11. Gardner, D., Davies, M.: A new academic vocabulary list. *Applied Linguistics*, 35(3), 305-327 (2014).
12. Gitsaki, C.: The development of ESL collocational knowledge (Unpublished doctoral dissertation). University of Queensland, Brisbane (1996).
13. Hallgren, K.: Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23-34 (2012).
14. Hill, J., Lewis, M., Lewis, M.: Classroom strategies, activities, and exercises. In: Lewis, M. (ed), *Teaching Collocation: Further developments in the lexical approach*, pp. 88-116. Language Teaching Publications, Hove (2000).
15. Hunston, S.: *Corpora in applied linguistics*. Cambridge University Press, Cambridge (2002).
16. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174 (1977).
17. Light, R.: Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, 76(5), 365-377 (1971).
18. Nesselhauf, N.: *Collocations in a learner corpus*. John Benjamins, Amsterdam (2005).
19. Rogers, J. What are the collocational exemplars of high-frequency English vocabulary? On identifying multi-word units most representative of high-frequency lemmatized concgrams

- (Unpublished doctoral dissertation). University of Southern Queensland, Queensland (2017).
20. Siegel, S., Castellan, N.: Nonparametric statistics for the behavioral sciences. McGraw Hill, Boston (1988).
 21. Stubbs, M.: Collocations and semantic profiles: On the cause of the trouble with quantitative methods. *Function of Language* 2(1), 1-33 (1995).

English Multi-word Expressions as False Friends between German and Russian: Corpus-driven Analyses of Phraseological Units

Lyubov Nefedova

Moscow Pedagogical State University
1, Malaya Pirogovskaya St., Moscow, 119991 Russia
lfn@mpgu.edu

Abstract. In this paper we outline the issue of borrowing multi-word expressions from English into German and Russian. It is established which forms of borrowed multi-word expressions (original or translated ones) predominate in contemporary German and Russian. The corpus-driven analysis of functioning of English multi-word expressions in the German and the Russian press has revealed direct borrowings and calques. The majority of the multi-word expressions borrowed from English into German over recent decades are direct loans, whereas multi-word expressions borrowed from English into Russian are still calques. Uncommon direct loans in Russian often belong to professional or youth slang. In this case direct loaned multi-words expressions from English are considered in German and Russian as stylistic false friends. Direct loaned multi-words expressions can rarely have semantic differences in both languages and be semantic false friends. The text corpora comparative study helps identify semantic and stylistic false friends and provides examples for a German-Russian and Russian-German dictionary of false friends, which are represented in this paper.

Keywords: multi-word expressions, borrowing, false friends, stylistic false friends, semantic false friends, direct loans, calques

1 Introduction

There are a lot of intersections between German and Russian and both languages share a significant number of words of predominantly Latin and Greek origin that have the same meaning. In some cases, however, their semantics can be radically different or include different emotional or connotative components of meaning. Such cases of formal congruence and semantic non-equivalence are known as “false friends of a translator”.

The only available dictionary of false friends for German and Russian languages by K. Gottlieb was published in 1972 (reprinted in 1985) [4, 5] and is increasingly obsolete in content. It does not represent the recent influence of English on German and Russian and does not contain new words and multi-word expressions from English which are new false friends.

A new dictionary on which we are working is supposed to help Russian learners of German and German learners of Russian to approach the problem of false friends more consciously to act competently in situations of intercultural contact [7, 8, 10, 11]. It constitutes the results of the corpus-driven analyses of actual use of English

word and multi-word expressions in the contemporary German and Russian public discourse. As text corpora we applied corpora of the Institute of German [16] and National Corpus of Russian [17]. The parallel German-Russian corpus could not be used because it does not contain investigated English multi-word expressions.

The main subject of this paper is the phenomenon of the influence of English on the phraseological part of contemporary German and Russian. The goal of research is to present some English multi-word expressions as false friends in German and Russian received through the corpus-driven analyses and to show that there are two kinds: semantic and stylistic.

Linguistic interpretation of collocational data from large corpora is the main method of analysis. This method helps describe multi-word expressions that develop new meanings not found in the original language.

Based on Burger's theory of phraseology we discern depending on the degree of idiomaticity nature (Idiomatizität):

- collocation (Kollokation) *sich die Zähne putzen*;
- part idiom (Teil-Idiom) *einen Streit vom Zaun brechen*;
- full idiom (Voll-Idiom) *jemandem reinen Wein einschenken* [2].

2 What are False Friends between German and Russian today?

The problem of the false friends has been dealt with in the investigations of a number of authors. Almost all observers emphasize that the false friends are pairs of words in two languages that look or sound similar, but differ significantly in meaning.

Since Haensch the concept 'false friends' is not restricted to semantic differences [6]. There are false friends based solely on stylistic differences.

In *Metzler Lexikon Sprache* by Glück the definition of false friends is extended. Not only words but also multi-word expressions are designated as false friends ("komplexe Ausdrücke"). Besides, the feature of similarity is specified: the false friends are in two (or more) languages phonologically, graphically or morphologically similar words (or phrases). And it is accentuated that words and phrases can have different meaning gradations, reference ranges or connotations:

[...] in zwei (oder mehr) Sprachen phonologisch, graphisch oder morphologisch ähnliche Wörter (oder komplexe Ausdrücke), die jedoch unterschiedliche Bedeutung(sschattierung)en, Referenzbereiche bzw. Konnotationen haben: it. *lametta* Rasierklinge vs. dt. *Lametta*; engl. *sympathetic* verständnisvoll vs. dt. *sympathisch* usw. [3: 195-196].

Because the false friends are words or multi-word expressions in two (or many) languages the linguists speak about interlingual false friends.

The problem of phraseological, or idiomatic, false friends has not been researched exhaustively. They are defined as set phrases in two languages that have the same literal meaning but differ as regards their idiomatic meaning or their sociolinguistic and stylistic features [1].

Today the list of false friends between German and Russian is populated by words which were borrowed from English, they can lure the learners into some uncomfortable traps. Many borrowed Anglicisms in German and Russian don't mean exactly

what one might expect. There are total false friends (they have completely different meanings), for example the words *germ. Rating* – *rus. Рейтинг* (*reiting*) have absolutely different meanings. The words *germ. User* and *rus. юзер* (*juzer*) are partial false friends. Both words have the same meaning 'somebody who uses the computer'. But the German word means also as a slang word 'somebody who regularly takes drugs'.

3 Phraseological borrowing from English into German and Russian

The influence of English on other European languages is so great that their vocabularies are nowadays enriching themselves not only through the borrowing of separate words from the former, but also more and more phrases or multi-word expressions entering them. Such multi-word expressions are interphraseologisms: they are used in many languages [12].

Both German and Russian phraseological systems are now considerably influenced by English: new borrowings in phraseology fill lacunae in the languages. Direct borrowings from English, not their calques are typically used by modern Germans. The Russian language also borrows lots of English phrases but not directly, most of them are calques or latent Anglicisms [13].

When addressing the issue of phraseological borrowing from English into contemporary German we can register a great number of direct borrowings. All directly borrowed multi-word expressions are word-combinations of the type *adjective* + *noun*, both components being autosemantic words. A lot of loaned multi-word expressions in German are collocations *Electronic Banking*, *Global City*, *Late Show*, *Lean Management*, *Personal Trainer* and part idioms: *Blind Date*, *Daily Talk*, *Global Player*, *Soft Skills* (the figurative component is underlined). Many loan phrases in German and in Russian are terms belong to technical languages, e.g. *Global Player*, *Golden Handshake*, *Global Village* (economics), *Soft Skills* (sociology), *Golden Goal* (sport), *World Wide Web* (EDP) [14].

In the dictionary of neologisms, these multi-word expressions are not always represented as phrases per se [18]. Most of them have different normative spellings: as two separate words, as one word or with a hyphen, so they can even be considered as composites, e.g. *Functional food*, *functional food*, *Functionalfood*, *Functional-Food*. Russian speakers normally try to adapt phrases borrowed from English, so calques are more popular than direct borrowings, whereas German speakers make use of either a direct loan phrase or a translated one, e.g. *Functional food* and the calque *funktionelle Lebensmittel*, the former being more preferable, though. In Russian, only the calque *функциональные продукты* (*funktionalnyye produkty*) is in use.

Russian speakers also use lots of English loans but they are mostly loan multi-word expressions which are nativized in Russian. The direct borrowing of multi-word expressions from English is not common for Russian. Some English direct multi-word expressions are used as proper names in Russian, e.g. *Functional Food* as the name of a company, *Global Village* as the name of an English language school or a fair. Such multi-word expressions are usually transliterated and are also partially adapted.

4 English multi-word expressions as new false friends between German and Russian: corpus-driven analyses

Multi-word expressions from English are new false friends between German and Russian. We can reveal both types of false friends: semantic and stylistic false friends. The majority of them are stylistic false friends. I give below some examples.

The direct loaned multi-word expression *Highpotential (Hi-Po)* is a full idiom in German and means ‘junger Akademiker, der beste fachliche und soziale Kompetenzen und damit geeignete Voraussetzungen für eine Führungsposition besonders in einem international tätigen Unternehmen hat’ (a young employee who has been identified as having the best professional and social competences and suitable prerequisites for a leadership position, especially in an international company) (<http://www.owid.de/artikel/315700?module=neo&pos=16>). The set expression has a figurative metonymic meaning in German and it is a neutral phrase.

I come to this conclusion based on many examples of their use in context, in contemporary German public discourse, for example: *US-Forscher haben nun untersucht, was echte High Potentials ausmacht und wie man von ihnen lernen kann* (Zeit Online, 12.01.2012, Die Besten unter den Besten). *Nach MasterCard wechselte ich zur General Electric – zur GE Money Bank. Dort war ich High Potential und konnte mit den Besten des Unternehmens Projekte machen* (NEW12/JUN.00229 NEWS, 21.06.2012, S. 68,69,70; Der Durchstarter).

The German multi-word expression *hohes Potenzial* is a collocation, it has a direct meaning, for example: *Das Land hat hohes Potential im Blick auf die Zukunft* (<http://dict.leo.org/forum/viewUnsolvedquery.php?idThread=115540&lp=ende&lang=de>). The translation tips of this collocation from German to English are: “the country has a promising future”, “the country’s future potential is high”, “there is a high future potential for the country”, “the country has great potential for its future”, “the country is full of potential for its future” (ibid.).

The equivalent of *High potential (Hi-Po)* in Russian is a calque, it is a periphrasis *молодой сотрудник/ молодая сотрудница с потенциалом* (*molodoy sotrudnik/molodaya sotrudnitsa s potentsialom*) (a young employee with potential). The main sphere of use of multi-word expressions directly borrowed from English into Russian is professional discourse or youth subculture. The multi-word expression *хай потенил (хай-по)* (transliteration of the English set expression) is used in Russian as a slang phrase: *Хай потенил решают задачи с помощью микрорешений программ* (<https://vk.com/id4943278>) *High potentials (Hi-Po’s) solve tasks with the help of mikro learning programmes*. The multi-word expression is written like an English phrase too: *мы планируем сосредоточиться на областях развития лидерства топ-менеджеров, менеджеров среднего звена и high potentials* (we are planning to focus on areas of development of leadership of head-managers, mid-level managers and high potentials) (<http://www.trainings.ru/library/articles/?id=9052>).

A striking example of a loan multi-word expression borrowed from English into Russian which has undergone a semantic revaluation is *elektronnyy kesh* (*electronic cash*). In German the expression *Electronic Cash* has a meaning ‘bargeldlose Zahlungsmittels einer EC-Karte’

(<http://www.owid.de/artikel/298337?module=neo&pos=13>), for example: *In zahlreichen westeuropäischen Ländern wird auch Electronic Cash akzeptiert* (A00/JUN.41196 St. Galler Tagblatt, 15.06.2000, Ressort: TB-LBN (Abk.); Ohne Not auch ohne «Nötli»).

The Anglicism *electronic cash* corresponds to two multi-word expressions in Russian, i.e. *электронные деньги* (*elektronnyye dengi*) (money) or *электронные платежи* (*elektronnyye platezhi*) (payments) and *электронный кэш* (*elektronnyy kesh*). The direct loaned multi-word expression phrase *elektronnyy kesh* acquires a new meaning in Russian –not ‘electronic money’ like in English and German, but ‘money for gangsters, drug traffickers, terrorists, killers and corrupt officials’ [Nefedova, Polyakov 2014]. An example from Russian: “... задачей Центрального банка России является твердое заявление, что электронный кэш в Россию не будет допущен” ... *the task of the Central Bank of Russia is to firmly state that electronic cash will not be allowed in Russia* (http://www.e-reading.club/chapter.php/84362/65/Yurovickiii_Denezhnoe_obrashchenie_v_epohu_peremen.html).

The analysis of various contexts of use of the multi-word expressions *electronic cash* and *elektronnyy kesh* in German and Russian text corpora allows identifying them as semantic false friends.

Further corpus-driven analyses of English multi-word expressions in German and Russian can give other examples of semantic and stylistic differences and the corpora “should be the sole source of our hypotheses about language” [15].

Such multi-word expressions are idiomatic semantic false friends, but they are not represented in the only available dictionary of false friends for German and Russian by K. Gottlieb, published 1972, which is becoming increasingly obsolete in content. A new dictionary is supposed to help Russian learners of German and German learners of Russian approach the problem of false friends more consciously and act more competently in situations of intercultural contact.

5 English multi-word expressions in the new dictionary of false friends for German and Russian

The new dictionary of false friends for German and Russian will include English multi-word expressions that are semantic and stylistic false friends. The content of the dictionary will be so arranged that it will be used by Russian learners of German and German learners of Russian alike.

Let us look at two examples of how multi-word expressions will look like in the new dictionary of false friends between German and Russian.

Each article in the dictionary designates formal consentaneous pairs of words and has a cross construction. In the German part of the dictionary, a corresponding German lexical unit is accompanied by a basic grammatical description. It has a canceled translation which is automatically assumed but could be a trap for the learner, and finally a correct translation. The word use is illustrated with an example from everyday speech which is translated into Russian.

In the Russian part, a reverse translation of the correct Russian equivalent is given, backed with an example of its use.

Presentation of multi-word expressions as semantic false friends

E

	German	Russian
	Electronic Cash	электронный кэш
electronic cash	электронный кэш электронные деньги, электронные платежи	
Δ Was macht Electronic Cash so sicher? Почему электронные платежи надежны?		
электронный кэш	Geld für Gangster, Drogendealer, Terroristen, Killer und korruptierte Beamte	
Δ Электронный кэш не будет допущен в Россию. Geld für Gangster und Drogendealer wird nach Russland nicht zugelassen.		

Presentation of multi-word expressions as stylistic false friends

H

	German	Russian
	High Potential	хай потеншл (хай-по) / high potential
High Potential	хай потеншл / high potential (professional / youth slang) молодой сотрудник / молодая сотрудница с потенциалом	
Δ Nach dem Studium kam er als High Potential zu einer Bank und war innerhalb weniger Jahre zum Projektleiter aufgestiegen. После учебы как молодой сотрудник с потенциалом он стал работать в банке и через несколько лет стал руководителем проекта.		
хай потеншл / high potential (professional / youth slang)		High Potential
Δ Хай-по или хай потеншл – новый тренд в управлении кадрами. High Potential ist ein neuer Trend in der Personalverwaltung.		

Stylistic differences between English multi-word expressions in German and in Russian are labeled *professional / youth slang*.

The dictionary will give examples of multi-word expressions in use from text corpora of the contemporary German and Russian public discourse.

6 Conclusion

In the light of the growing internationalization of German and Russian the problem of false friends gets more relevant. Considering that the vocabulary of both languages develops in the ethnocentric cultures, the number of false friends is increasing rapidly. New false friends in German and Russian are borrowed Anglicisms, but they are not reflected in the dictionary of false friends for German and Russian. In addition, there are more new multi-word expressions which are used differently. That is why one of the important goals of modern linguistics is to update the dictionary of false friends to the current standard of lexicography and include English multi-word expressions.

This paper illustrates semantic and stylistic differences of meaning of English multi-word expressions in German and Russian and also some new forms of presenting of phraseological units in the German-Russian dictionary of false friends based on corpus-driven approach. Corpus linguistics is a research approach that supports empirical investigations of language use, primarily of new lexical units, and serves as a reliable source of material for lexicography.

References

1. Al-Wahy, A. S.: Idiomatic false friends in English and Modern Standard Arabic. *Babel* 55 (2), 101–123 (2009).
2. Burger, H.: *Phraseologie: Einführung am Beispiel des Deutschen (Grundlagen der Germanistik (GrG), Band 36)*. 3. Auflage. Erich Schmidt Verlag GmbH & Co, Berlin (2007).
3. Glück, H. (ed.): *Metzler Lexikon Sprache*. 4. aktualisierte und überarbeitete Auflage. J.B. Metzler, Stuttgart/Weimar (2010).
4. Gottlieb, K.H.: *Deutsch-Russisches und Russisch-Deutsches Wörterbuch der „Falschen Freunde des Übersetzers“*. Sovyetskaya Entsiklopediya, Moskva (1972).
5. Gottlieb, K.H.: *Sprachfallen im Russischen: Wörterbuch der „falschen Freunde“ Deutsch und Russisch: Ein Lern- und Nachschlagewerk*. Hueber, Ismaning (1985).
6. Haensch, G.: „Faux amis“. In: *Lebende Sprachen: Zeitschrift für fremde Sprachen in Wissenschaft und Praxis*. Fachblatt des Bundesverbandes der Dolmetscher und Übersetzer 1, 16 (1956).
7. Nefedova, L.: *Lexikografie und Interkulturalität. Zum Projekt für ein deutsch-russisches und russisch-deutsches Wörterbuch der falschen Freunde des Übersetzers*. In: Scheller-Boltz, D., Weinberger, H. (eds.) *Lexikografische Innovation – Innovative Lexikografie. Bi- und multilinguale Wörterbücher in Gegenwart und Zukunft: Projekte, Konzepte, Visionen*, S. 231–247. Olms Verlag, Hildesheim/Zürich/New York (2017).
8. Nefedova, L.: *Sprachfallen: Wie kann man sie umgehen? Über den Umgang mit „falschen Freunden“ des Deutschen und des Russischen*. In: Földes, Cs. (ed.) *Interkulturelle Linguistik als Forschungsorientierung in der mitteleuropäischen Germanistik*. (Beiträge zur interkulturellen Germanistik; 8), S. 149–165. Narr Francke Attempto, Tübingen (2017).
9. Nefedova, L., Polyakov, O.: *Set Expressions Borrowed from English into German and Russian: Direct Loans or Calques?* In: Gural, S. (ed.) *Procedia – Social and Behavioral*

- Sciences. The XXVI Annual International Academic Conference, Language and Culture, 27-30 October 2015, vol. 200, pp. 83–86 (2015).
10. Nefedova, L.: Russian-German and German-Russian Dictionary of False Friends. In: Karpova O.M., Shilova, E.A. (eds.) *Heritage Lexicography as Supporting Tool for International Council for Monuments and Sites (ICOMOS): Proceeding of the International Workshop*. Florence, July 21-23, 2014, pp. 71–74. Ivanovo State University, Florence/Ivanovo (2014).
 11. Nefedova, L.: O novom tipe nemetsko-russkogo i russko-nemetskogo slovarya “lozhnyih družey perevodchika: k voprosu ob innovatsiyah v leksikografii. In: Kitadzë M., Minasyan, S. (eds.) *Sotsiokulturnye i filologicheskie aspekty v obrazovatel'nom i nauchnom kontekste*, pp. 410–415. University Kyoto Sangë, Kyoto (2014).
 12. Nefedova, L.: On the Use of Interphraseologisms in the Journalistic Discourse of German and Russian Linguocultures. In: Gural, S. (ed.) *Procedia – Social and Behavioral Sciences*. The XXV Annual International Academic Conference, Language and Culture, 20-22 October 2014, vol. 154, pp. 130–137 (2014).
 13. Nefedova, L., Polyakov, O.: On Some Aspects of the Borrowing of Phrases from English into German and Russian. In: Arsenteva E. (ed.) *Phraseology in Multilingual Society*, pp. 141–155. Cambridge Scholars Publishing, Newcastle upon Tyne (2015).
 14. Nefedova, L.: Zu einigen Aspekten der phraseologischen Entlehnungen aus dem Englischen im heutigen Deutsch. In: Arsenteva E. (ed.) *Phraseology in Multilingual Society*. *Sbornik statey mezhdunarodnoy frazeologicheskoy konfrentsii „Europhras“ 19-22 avgusta 2013*, vol. 1, pp. 200–207. Tatarskoe respublikanskoe izdatel'stvo „KHETER“, Kazan (2013).
 15. Tognini-Bonelli, E.: *Corpus Linguistics at Work*. (= *Studies in Corpus Linguistics* 6). Amsterdam/Philadelphia (2001).
 16. Deutsches Referenzkorpus, <https://cosmas2.ids-mannheim.de/cosmas2-web/>, last accessed 2017/06/15.
 17. Russian National Corpus, <http://www.ruscorpora.ru/>, last accessed 2017/06/15.
 18. Neologismenwörterbuch, <http://www.owid.de/wb/neo/start.html>, last accessed 2017/06/15.

Towards the Generation of Bilingual Chinese-English Multi-word Expressions from Large Scale Parallel Corpora: An Experimental Approach

Benjamin K. Tsou^{1,3} Derek F. Wong² Ka Po Chow³

¹ City University of Hong Kong

² University of Macau

³ ChiLin (HK) Limited

btsou99@gmail.com

Abstract. There is increasing practical need for bilingual lexicon in wide ranging applications such as CAT systems and in cross-lingual information retrieval. The need for more sophisticated and more extensive resources involving bilingual multi-word expressions has increased even more for the bilingual processing of complex texts such as legal documents, technical names, contracts and patents. We report here a preliminary attempt at successful efforts to mine bilingual MWEs from 300+k Chinese-English patents.

Keywords: Parallel Corpus, Bilingual Alignment, Multiword Expressions.

1 Introduction

There is increasing practical need for bilingual glossaries in wide ranging applications such as CAT systems and in cross-lingual information retrieval. The need for more sophisticated and more extensive resources involving bilingual multi-word expressions has increased even more for the bilingual processing of complex texts such as legal documents, technical names, contracts and patents. Besides the availability of considerable computational power, the cultivation of the prerequisite database has special requirements, such as, 1) Identification of thousands of relevant patent families across two languages; 2) Identification of parallel bilingual linguistic segments; 3) Evaluation of bilingual equivalence; 4) Pruning of poor results from the database, etc.

This task of data cultivation and curation is by no means a simple one, especially when comparability of Chinese and English patents was not easily determined. In addition to considerable noisy data, there are also considerable challenges in the bilingual sentence alignment during the cultivation of the training corpus on which the system is modelled. In a recent study based on a corpus of more than 300,000 “comparable” Chinese and English patents (Tsou 2017), it was found that the longest Chinese sentence had 1,417 words and the longest single document had 249,322 words in the wide-ranging collection of Chinese patents. The database had a total of more than 500 billion words and characters and 10.5 trillion characters, and on the average, there were 30,000 words and/or characters for each document.

When the source language is Chinese, the patents have multiple additional challenges, such as: 1) Very long sentences consisting of continuous strings of characters; 2) Grammatical complexity: from frequent nominalization to telegraphic style; 3) Frequent subordination, implicit constructions, and dangling modifiers; 4) Problematical anaphora reference; 5) Active vs Passive forms; 6) incorrect characters; 7) Extensive neologisms; 8) Synonymy and polysemy; 9) wrong punctuation.

Words are conventionally segmented in the Western languages. However, an unusual challenge not faced by patents written in English, for example, but found in Chinese patent text is the opaque tokenization of the words represented by generally continuous strings of Chinese characters in texts. They are occasionally punctuated by a smaller set of punctuation marks than found in English, for example, but without the use of capitalization to delineate proper nouns or the use of italization to indicate special terms or emphasis. There is therefore a critical challenge with the proper tokenization and identification of name-entities and novel words or multi-word expressions are relative unique to Chinese text processing.

But the Chinese language has other uncommon problems with structural ambiguities beyond the classical “old man and woman” parallel, partly because of the combinatorial possibilities of its mono-morpho-syllabic writing system.

There are still other challenges in Chinese patent processing, such as:

- (a) Content formatting: Broken sentences; Wrong encoding; OCR errors
- (b) Different data formatting: USPTO, EPO, WP, SIPO have own formats;
- (c) Source of Data: Conversion from pdf into XML
- (d) Text extraction from graphs and diagrams

Our efforts to obtain Chinese-English multiword expressions have several stages.

2 Cultivation and curation of parallel corpus

The successful development of corpora of bilingually aligned sentences has been made following extensive attempts involving the cultivation and curation of a sizable prerequisite database of comparable Chinese and English patents. The English patent documents are harvested from the website of WIPO (World Intellectual Property Organization), while the Chinese patent documents are originated from SIPO, State Intellectual Property Office, the official patent office in China. The patent data is then divided into different sections, among which only title, abstract, claims and descriptions are our concern.

The correspondence between the English and Chinese patents are established through their cross-references. Since the comparable patents are not strictly parallel, some prior art of individual alignment methods might not be effective. For example, length-based alignment method might not be accurate since it does not consider content similarity. Moreover, new technical term poses a challenge to the bilingual dictionary-based alignment algorithm.

Following persistent efforts we have been able to harvest a database of 300+k comparable Chinese-English patents. (Lu et al. 2009, 2011, 2015).

3 Filtering of bilingual sentence pairs

We use a sequence of algorithms to filter good quality parallel sentence pairs. First we use a publicly available dictionary to perform preliminary alignment based on a dictionary-based similarity score of a sentence pair (Utiyama and Isahara, 2003).

In the resultant set of sentence pair candidates obtained from the above preliminary alignment, some pairs will be of very incomparable lengths. They are filtered based on a sentence length ratio which is found empirically. For example, if a sentence pair has around 50 words in the English sentence while there are more than 333 characters in the Chinese one, it is removed.

The parallel sentence candidates are further filtered by learning an IBM translation model Model-1 and the translation similarity score is computed by combining the translation probability value of both directions based on the trained model.

We make use of a number of existing alignment tools including Champollion (Ma 2006), Hunalign and Microsoft Sentence Aligner to arrive at a preliminary parallel by design if not on purpose corpus. Champollion applied dynamic programming based on a probabilistic score which allows us to obtain some correspondence other than 1-to-1, like 1-to-2, 2-to-2, etc.

It should be noted that the resultant database of 10+ million parallel sentence pairs are useful for training and fine-tuning MT system (Goto, 2013). However, they may not all be linguistically well-formed as they are the result of optimization of string matching (Koehn, 2010). For example, (1) “abdominal sac” --(膈 14下腹部区域内的)腹囊; (2) “ability of the peptide” --多肽(单体解离的)能力; (3) “access time”--【访问】时间.

4 Extraction of bilingual multiword expressions

As a further attempt to obtain linguistically well-formed entries, an alternative automatic approach to acquire the bilingual phrases from the above parallel corpus was made on the basis of Tian et al. (2011, 2014). We first derive the phrases from monolingual data, i.e. the source sentences of parallel data, by considering the possible n -gram of words. For those of high frequency words or phrases, we evaluate how likely a phrase can be constructed by two (words or) phrases p_i and p_j , by considering the following significant score:

$$\text{Score}(P_1, P_2) = \frac{f(P_1 \oplus P_2) - \mu_0(P_1 - P_2)}{\sqrt{f(P_1 \oplus P_2)}}$$

where $f(\cdot)$ and $\mu_0(\cdot)$ are the frequency and the mean under null hypothesis of independence of two phrases (El-Kishky et al. 2014). \oplus is the concatenation operator. The equation computes the number of standard deviations away from the expected number of occurrences, and this score can be considered a generalization of the t -statistic for identifying dependent bigrams. To extend the identified phrases to its bilingual, we first derive the word alignment information for the bilingual data using the model proposed by Dyer et al. (2013). The word alignments act as vital infor-

mation and are used to project the phrase boundaries from the source sentences to the target side of the bilingual data (Zeng et al., 2014). For those of unaligned words or phrases, we simply ignore it from the induction process. The described bilingual phrases extraction model is a kind of unsupervised approach, where all the statistics are automatically derived from a given parallel corpus aligned at sentence level.

The bilingual multi-word expressions obtained are fed into an online translation engine for further automatic evaluation. Each monolingual part of a pair of bilingual terms is input to a translation engine to obtain a translation, which is compared with the other part of the bilingual pair in terms of Levenshtein distance (LD). The following results are obtained: LD=0:11.9%, LD=1:25.5%; LD=2:29.3%; LD=3:11.9%; LD=4: 11.0%; LD \geq 5:10.3%.

We noted that from the figures, only 11.9% are identical with online translation result. The remaining near 90% are not identical but are still potentially valid entries, and are more valuable because they reflect the actual language in this particular domain which are not readily manifested through any publicly available resources. Further empirical studies show that lower edit-distance entries require less manual modifications. When the edit distance is equal to 1 or 2, about 65% are valid entries without modifications. 5% can be useful after manual modifications. For distance 3 or 4, about 55% are valid entries, while about 10% can become useful after modifications.

Our efforts have produced over 6 million MWE candidate entries. Further mostly straightforward manual efforts have yielded 1 million good bilingual MWEs thus far with another 1.5 million expected on completion of the project. The human efforts mostly entail the pruning of redundant constituents and in some cases the recovery of missing constituents as given in the examples in section 3. The final harvest rate should be about 40% or approximately 2.5 million bilingual MWEs.

5 Concluding Remarks

We have shown that highly selective bilingual multiword expressions in the field of Chinese-English patents can be rigorously mined from a much larger and carefully cultivated database of comparable patent consisting of more than 300+k Chinese-English patents involving more than 10.5 trillion Chinese characters. They can be highly useful for efforts in computer-assisted translation and cross-lingual information retrieval, for example. Several areas continue to pose challenges in the expeditious cultivation of these bilingual terms: (1) inherent noise in data because of OCR errors; (2) constituent redundancy, including cases of numbers, letters, modification and punctuation; (3) constituent omission, including cases of missing headword, pre-headed/post-headed modification, and incomplete sentences.

Resolving these challenges will enhance the semi-automatic cultivation of linguistically sound multiword expressions. Computational methods seem to plateau at a certain level regarding many of these linguistic issues and human efforts are still needed at present stage, and this remains an open research area.

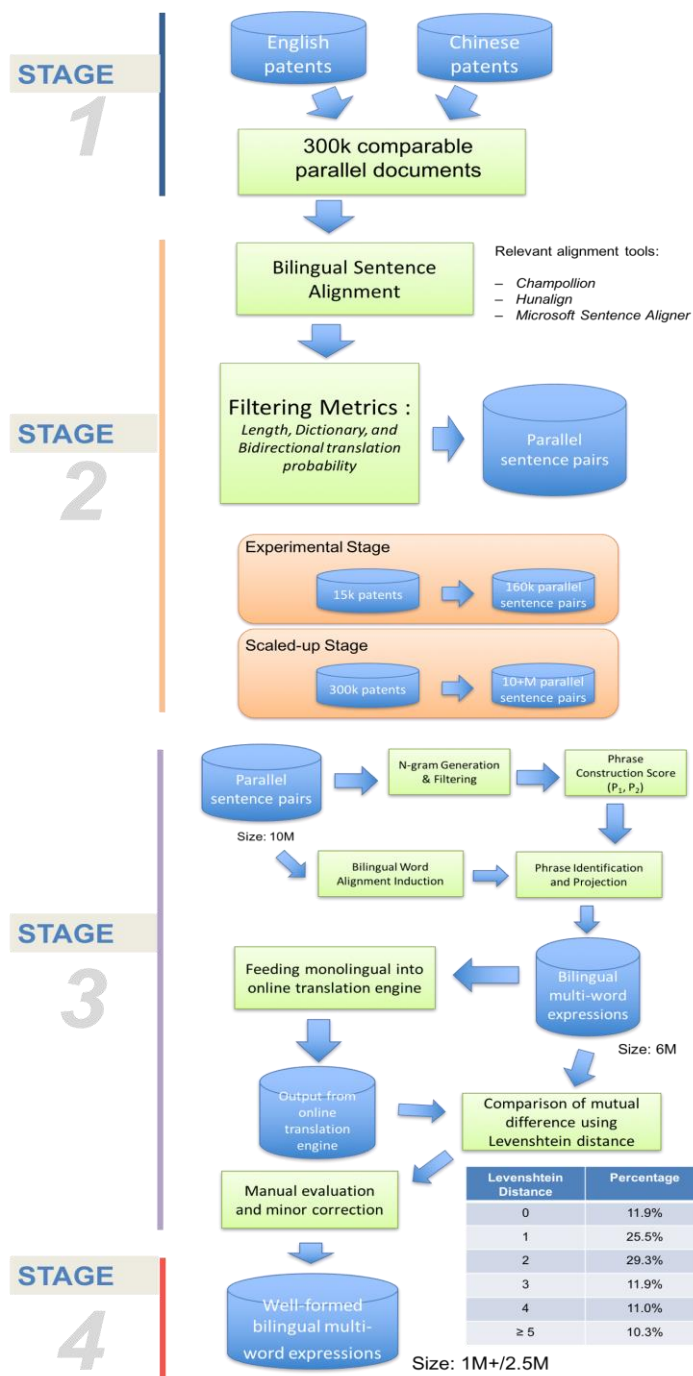
References

1. Dyer, Chris, Victor Chahuneau, and Noah A. Smith.: A simple, fast, and effective reparameterization of IBM Model 2. *Proceedings of NAACL-HLT*, pp. 644–648. (2013).
2. El-Kishky, Ahmed, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han.: Scalable Topical Phrase Mining from Text Corpora. *PVLDB* 8(3), 305–316(2014).
3. Goto, Isao, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou.: Overview of the patent machine translation task at the NTCIR-10 workshop. *Proceedings of NTCIR-10 Workshop Meeting*. (2013).
4. Koehn, Philipp.: *Statistical machine translation*. Cambridge University Press, the United Kingdom (2010).
5. Liu, Jialu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han.: Mining Quality Phrases from Massive Text Corpora. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. pp. 1729–1744. Melbourne, Victoria, Australia (2015).
6. Lu, Bin and Benjamin K. Tsou.: Towards Bilingual Term Extraction in Comparable Patents. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC'23)*. (2009).
7. Lu, Bin, Benjamin K. Tsou, Jingbo Zhu, Tao Jiang, and Olivia Y. Kwong.: The Construction of an English-Chinese Patent Parallel Corpus. *MT Summit XII 3rd Workshop on Patent Translation*. (2009).
8. Lu, Bin, Benjamin K. Tsou, Tao Jiang, Oi Yee Kwong and Jingbo Zhu.: Mining Large-scale Parallel Corpora from Multilingual Patents: An English-Chinese example and its application to SMT. In *Proceedings of the 1st CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2010)*. Beijing, China (2010a).
9. Lu, Bin, Tao Jiang, Kapo Chow and Benjamin K. Tsou.: Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT. In *Proceedings of the Workshop on Building and Using Comparable Corpora*. (2010b).
10. Lu, Bin, Benjamin K. Tsou, Tao Jiang, Jingbo Zhu and Olivia Kwong.: Mining Parallel Knowledge from Comparable Patents. In: *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*. IGI Global. (2011).
11. Lu Bin, Chow K.P., Tsou B.K.: Comparable Multilingual Patents as Large-Scale Parallel Corpora. In: Sharoff S., Rapp R., Zweigenbaum P., Fung P. (eds) *Building and Using Comparable Corpora*. Springer, Heidelberg (2013).
12. Lu Bin, Benjamin K. Tsou and Ka Po Chow.: Cultivating Large-scale Parallel Corpora from Comparable Patents: From Bilingual to Trilingual, and Beyond. *Linguistic Corpus and Corpus Linguistics in the Chinese Context (Journal of Chinese Linguistics Monograph Series No.25)*(Tsou, Benjamin, and Kwong, Olivia. (Eds).), 334-355 (2015).
13. Ma, Xiaoyi.: Champollion: A Robust Parallel Text Sentence Aligner. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. Genova, Italy (2006).
14. Tian, Liang, Fai Wong, and Sam Chao.: Phrase Oriented Word Alignment Method. In Wang, Hai Feng (Ed.), *Proceedings of the 7th China Workshop on Machine Translation* pp. 237–250. Xiamen, China (2011).
15. Tian, Liang, Derek F. Wong, Lidia S. Chao, and Francisco Oliveira.: A Relationship: Word Alignment, Phrase Table, and Translation Quality. *The Scientific World Journal* 1–13 (2014).
16. Tsou, Benjamin K.: Challenges And Advances IN Natural Language Processing of Chinese Patents-From Machine Translation To Cognitive Filtering. Invited paper, East meets West Forum, 2017 European Patent Office, Vienna (2017).

17. Utiyama, Masao and Hitoshi Isahara.: Reliable measures for aligning japanese-english news articles and sentences. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 72–79. Sapporo, Japan (2003).
18. Zeng, Xiaodong, Lidia S. Chao, Derek F. Wong, Isabel Trancoso, and Liang Tian.: Toward Better Chinese Word Segmentation for SMT via Bilingual Constraints. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1360–1369. Baltimore, Maryland (2014).

Appendix:

Filtering of Bilingual Chinese-English Multi-word Expressions from Large Scale Parallel Corpora of Comparable Patents



Phraseological Meaning and Image

Roza Ayupova

Kazan Federal University, Kazan Tatarstan 420008, Russia, 18, Kremlevskaya str.
rozaayupova@gmail.com

Abstract. The article is devoted to the study of the phraseological unit formation process, which is viewed from linguistic and semasiological angles. Taking a free word combination, prototype of the widely known phraseological unit, we analyze the process of its passing all stages of phraseologization – becoming a phraseological unit. Semasiologically it is a conversational implicature becoming a conventional one. Eventually, it is called phraseological meaning. As component parts of phraseological units one can often observe utilization of words which are semantically not collocable, e.g.: *pigs fly*, *speak daggers*. This paradoxical collocation provides brightness of the image of the new linguistic sign – the phraseological unit. In this paper, we will try to show the mechanism of conversational implicature becoming a conventional one and the role of paradoxes in the process of formation of new linguistic signs, which are secondary nominations of a denotatum. The empirical material of the present research consists of only a special group of phraseological units.

Keywords: Phraseological unit, phraseological meaning, prototype, implicature, linguistic sign.

Introduction

The fact that phraseological units (PU) are in the focus of many linguists' attention is related to two factors: firstly, phraseology makes up a very significant part of the lexical fund of any language; secondly, PUs are specific units with a complicated meaning. The aim of the current paper is to study the reason for semantically not collocating words becoming component parts of PUs.

Our research is based on the contemporary linguistic and semasiological methodology presupposing application of the method of etymological analysis, the method of componential analysis, the method of discourse analysis, the method of comparative analysis etc.

Complex nature of PUs is traced back to the prototype of each unit and the process of its development to acquire the current meaning and form.

The term phraseology in its broad meaning encompasses also paremiology, therefore as examples in our research we use proverbs or sayings alongside with PUs.

Process of Phraseologization

As it was already noted, phraseological nomination is a secondary nomination which emerged to name objects, properties and processes in a more expressive way, therefore, most PUs are endowed with bright connotation. Each of them as a free word combination once was utilized by a speaker (addresser of the text) figuratively with the purpose of impressing the listener (addressee of the text).

So, when using the word combination ***reach the wool sack*** for the first time instead of saying *s/he became a lord chancellor* an addresser wanted to sound more expressive. He knew that the addressees were well aware of the fact that a lord-chancellor sits on the sack full of wool, that is why he used the word combination mentioned above, implying the information about the person getting the position of lord-chancellor. Using the term suggested by P. Grice, we call it a conversational implicature when speaking about its use as a free word combination [1]. The technique of transference of meaning utilized in this example is sure to be metonymy, as there used to be a close correlation between the processes expressed by the free word combination and the figurative meaning it was used in or “using one entity to refer to another that is related to it” [2] – the position and seat usually taken by a person occupying this position.

Afterwards, the word combination was used with the above mentioned implicature by other speakers. As a result, the phrase ***reach the wool sack*** is associated with the meaning “to become a lord- chancellor”. Gradually, the word combination abstracted from the semantic meanings of its component parts, since then it acquired a new phraseological meaning, and the conversational implicature turned into a conventional one. The process of phraseologization is completed with fixing the phrase in the phraseological dictionary. Juxtaposing the first nomination of this process with the PU one discovers that the latter is endowed with the image, which is related to the direct meaning of the prototypical word combination ***reach the wool sack***. The presence of this image provides the expressiveness of the PU under consideration, which is one of the most important components of phraseological meaning. Expressiveness is defined by the well-known Russian phraseologist A.V. Kunin as “a bright figurative property of a linguistic unit conditioned by imagery, intensity and emotiveness” [3, 179].

Traditionally, imagery is understood as the ability of linguistic units to create a vivid and sensuous image about objects and phenomena of reality. Accordingly, imagery is closely connected with the direct meaning of the free word combination which is homonymic with the definite idiom.

A.V. Kunin explains that a receptor perceives the content of the notion realized by the idiomatic meaning and the semantic information enclosed in the prototypical word combination. These two ‘pictures’ give birth to the third one, which makes up phraseological imagery. The further from each other the two compared objects are, the brighter the image is [2]; the brighter the image, the more expressive the PU is.

In the above mentioned example metonymy is used as a technique of deriving figurative meaning. As the analysis shows illogicality, e.g.: *when pigs fly, pie in the sky, have a flea in one’s ear, eat one’s words*; hyperbolic metaphor, e.g.: *eat/ drink (sth.) till / until it comes out of one’s ears*; litotes, e.g.: *under smb’s nose, as rare as hen’s teeth* tend to bring together the most paradoxical phenomena.

Coming back to the process of phraseologization, one should note one more thing, important from the semiotic viewpoint. If letters and words as linguistic signs have no logical connection between their form and content, according to Charles Pierce’s classification of signs, they fall into the group of symbols, PUs have this logical connection based on the technique of deriving figurative meaning. That is why; they appear in the group of indexes.

“In semiotic terms, a sign system is a kind of field of related things, and their meaning comes from how they relate to each other” [4, 3]. It remains true in linguistics also, where the meaning of an utterance or a phrase is very much dependent on how different words – their component parts – collocate with each other.

In this place, it is worth mentioning the example used by the American linguist Noam Chomsky, one of the founders of structural grammar: “Colorless green ideas sleep furiously” [5]. The sentence is completely grammatical, yet completely nonsensical. The same can be said about the word combination *pigs fly*, while as N+V structure is one of the most typical structures in the English language.

But the speaker who first used it intentionally, with the definite implicature, to let the addressee know that the thing which in the other part of the utterance is mentioned will never take place, as pigs never fly, e.g.: *You will get your money back when pigs fly*. As Pavol Štekauer writes:

“By implication, any naming act is necessarily preceded (or dominated) by a network of ‘objectively’ existing relationships. By implication, the naming act is a cognitive phenomenon relying on the intellectual capacities of a coiner” [6, 13].

In accordance with some sources, the figure of swine in the air was first mentioned by Winthrop, an English Puritan explorer, who settled in Massachusetts, in

his story “The History of New England” [7]. One more existing viewpoint about the etymology of this PU claims that it was first found in a list of proverbs in the 1616 edition of John Withals's English-Latin dictionary - A Shorte Dictionarie for Yonge Begynners: ***Pigs fly in the ayre with their tayles forward.***

This form of the proverb was in use for nearly two hundred years. Other animals also could be mentioned instead of pigs. It is presupposed that pigs are the least likely to fly, because of their bulkiness and their habit of “rooting in earth”, that is why the component *pig* appeared to be the most timeless one [7]. And the reason for it is its sounding the most paradoxical in this linguistic environment. Now this PU is transformed into ***when pigs fly*** and used commenting on sarcastically any prediction that sounds too optimistic.

Conclusion

As our analysis witnesses, for the purposes of creating brighter images, phraseology resorts to bring together the following paradoxical concepts: an agent and an action that cannot be fulfilled by it, e.g.: ***one's face falls*** – one shows one's disappointment, dismay etc. by one's expression; an object and an action this object cannot undergo, e.g.: ***eat one's words*** – (be forced) to take back what one previously has told was true, certain etc. (because of changed circumstances); an object placed to somewhere, where it cannot be, e.g.: ***pie in the sky*** – a future reward after death, considered as a replacement for a reward not received on earth, something good that is unlikely to happen; an object and a property it cannot be endowed with, e.g.: ***a cultural desert*** – a place or community where there is little or no artistic or academic activity or any interest take in such pursuit. Bringing together such paradoxical phenomena results in a bright phraseological image.

References

- Grice, P. Logic and conversation. In Syntax and Semantics 3: Speech acts., eds. Cole, P. and Morgan, J. p. 41-58. Academic Press, New York: Academic Press (1975).
- Lakoff, G. and Johnson, M., Metaphors we live by. Chicago: University of Chicago press (2008).
- Kunin, A. Kurs frazeologii angliiskogo yazyka. Vysshaya shkola, Moscow: Fenix (1996).
- Maasik, S, Solomon, J. Signs of Life in the USA. Boston: Bedford/St. Martin's (1994)
- Chomsky, N. Syntactic Structures. The Structures. Mouton, The Hague/Paris (1957).
- Štekauer, P., Lieber, R. Handbook of Word-formation. Springer, Rotterdam (2005).
- The Phrase Finder. <http://www.phrases.org.uk/meanings/pigs-might-fly.html>. Last accessed 2017/4/9.

Student Research Workshop

Language and Power in Czech Corpora

Irene Elmerot¹[0000-0002-9809-8207]

¹Stockholm University, SE-106 91 Stockholm,
Sweden irel5167@student.su.se

Abstract. The author focuses on quantitatively examining the linguistic othering in printed media discourse in the Czech Republic, using the Czech National Corpus. The method used so far has been a corpus-based discourse analysis based on the adjectives preceding the keywords for each part of the project, now moving on to include reporting verbs. The theoretical starting point is that power relations in a society are reflected in that society's mainstream media, and that the language usage in these media contributes to the worldview of its recipients, in some cases even helps to construct it. Frequent but widely dispersed stereotypical and negative phrases and collocations are examples of a power language that may not be visible at once, but slowly enters the general discourse in a society. This project aims to survey these linguistic othering phrases in the Czech media discourse, as comprehensively as possible, and shed some light on their appearance over time.

Keywords: othering, discourse analysis, corpus linguistics, Czech

1 Introduction

This paper presents an ongoing project on how language relates to power and how othering is depicted in the language of a small but well-known country in the heart of the European continent.

2 Past project: Linguistic othering in Czech printed media

2.1 Roma vs. Gypsy – a short discourse and corpus analysis

The first article in this project was published in 2016 [6]. The subcorpus SYN of the Czech National Corpus (at that time about 5.170.696 lemmata¹ and 2.685.127.310

¹ For the sake of clarity, the following definitions are used: **denomination**: name of a group of people.

lemma (pl. lemmata): form of a word representing all forms of that word.

othering: the action of labelling someone who belongs to a different, often subordinate, social category with the purpose of exclusion from the sender's social category.

tokens [4]) was then used to analyse the Czech lemmata for Roma and Gypsy (*Rom* and *Cikán*) with their adjacent (position L-1) adjectives, to see what differences there were in the discourse of the most popular Czech printed media from 1989 to 2009, depending on which denomination had been used. A statistical analysis of the frequencies of adjectives adjacent to the two denominations was then performed. The main theory was a parallel to Masako Fidler's idea [8] about finding a "more automatic mental representation" of these "others". Most surprising of the results was that for both denominations, about a third of the adjectives in position L-1 were geographically related (as in "Romanian", "Czech", or "local"). The negative adjectives were about the double when adjacent to the lemma Gypsy compared to the lemma Roma. The neutral words were, on the other hand, almost the double for the lemma Roma compared to the lemma Gypsy. One adjective, "unadaptable" (*nepřizpůsobivý*), has become so popular in recent years as a collocation to the Roma people in the Czech Republic that it has created at least one article [13] and one book [12]. This adjective was found, but couldn't really be classified as a collocation in this material. These are not very surprising results, but hereby confirmed for a "small" language by a large source material.

2.2 Linguistic othering of minorities in the Czech Republic (Follow-up)

In the follow-up [7] a similar method of analysing the frequency of adjectives adjacent to denominations of people was used, but this time the nouns for the minority groups Roma, Ukrainians and Vietnamese were analysed. The theory now focused more on the hypothesis that power relations in a society are reflected in that society's mainstream media, and that these media's language usage contributes strongly to the receivers' worldview, in some cases even helps to construct it (cf. inter alia [8], [1] and [10]). The material was still the SYN series (now version 4) of the Czech National Corpus, and the time frame 1990–2014. The amount of lemmata had increased to 7.427.573, and tokens to 4.349.023.692 with this version [5]. The performed search returned 29.657 hits for the lemma *Rom*, 4.470 for Vietnamese (*Vietnavec*) and 5.335 for Ukrainian (*Ukrajinec*). Based on the previous research for these minority groups, the Roma were most likely to be linguistically othered in a similar way to the Vietnamese (who have been a large minority group in the Czech Republic since the 1960's). It was not clear what would be found about the Ukrainians, who are a Slavic people like the Czechs, and have been immigrating in larger groups after 1989. The geographical words proved indeed to be much more frequent for Roma than for the other groups, e.g. of the total before the lemma Roma (29,657 absolute frequencies), "Czech" made up ten per cent. Before the lemma Vietnamese, there were a large frequency of the word "lonely" or "self" (*samotný*), and the word for "small" (*malý*) was also rather frequent – there was a double-check in context to make sure it didn't mean "child", but this did not seem to be the case. Both Roma and Vietnamese were on occasion considered "unadaptable". The Ukrainians were found to be more positively

Also, "Roma" denominates all people of Roma descent, whether they call themselves Roma, Gypsy or something else. "The lemma Gypsy" or "the lemma Roma" then means the lemmata in the corpus searches or analyses.

depicted, but also often depicted as drunk (*opily*). In total, the negative/positive ratio was 5.21 for the lemma *Roma*, 3.49 for *Vietnamec* and only 0.91 for *Ukrajinec*. A Pearson chi2 test has been performed on this, and the probability (P) value turned out to be 0.00, which means that we can reject the hypothesis that the variables are independent, i.e. that the observed differences in the sample data are systematic. Even with the neutral adjectives removed, the chi2 test shows the same result.

3 Future project: Linguistic power structures in Czech media

3.1 Gender structures shown in Czech media language (Work in progress)

During the autumn of 2017, the plan is to broaden the research scope of this project. Null hypothesis: There is no gender differentiation in the reporting verbs about women and men in the source material. Hypothesis 1: There is gender differentiation in the reporting verbs about women and men in the source material. To test this, the material will again be the most updated SYN series of the Czech National Corpus, and the keywords professional terms, like *poslankyně* (female member of parliament), *úřednice* (female office worker) and perhaps *učitelka* (female teacher), to compare theoretical professions with a more practical one where women are in majority. These keywords searches would then be filtered with reporting verbs like the Czech verbs for “claim”, “assert” or “establish”. The same searches and filtering would then be made for the male counterparts (MP, office worker and teacher) to see if the verbs differ in frequency, and what that may tell us about the linguistic othering structures created or added to by Czech media when it comes to working women of the middle and higher classes. One of the hypotheses will then be rejected. Previous research to form a basis will then be Baker [1], especially chapter 4 on how men and women are represented in a language, that is how mental images of the female professionals are created using “signifying practices and symbolic systems” [1, 89] in the language, and what the cumulative language usage tells us when such a large source material as the SYN version 5 corpus is analysed. Such corpora are likely to tell us what kind of language usage large numbers of people encounter regularly. Also research surveys such as Oates-Indruchová’s article [11] on the continued gender critique in the Czech Republic from 1948 onwards, and – for the background – such work as Rebecca Nash’s article [9] on gender scholars in the Czech Republic during the 1990s will be used.

3.2 Language structures: class, gender, minorities & the city vs. countryside cleft (planned project)

Since the aim is to turn this into a Ph.D. project in 2018, there is also the possibility of a larger project, where the corpus research could go wider in time and hopefully stretch back at least one century, since the Czech National Corpus is hoping to expand their amount of older corpora. The theory is not yet set in stone, but the previous research would include inter alia R. Čech’s paper on Language and ideology [2] and T.

Váňa's Language power potential [14], as well as what M. Fidler and V. Cvrček are doing in their Needle in a Haystack project [10]. When it comes to the method to be used, the new collocation candidate function of the Czech National Corpus may come in handy here, to compare possible collocations to these words in the two corpora before and after 1989, and thereby extract expressions pointing towards the power structures and mental representations visible in the Czech media, as both Colson [3] and Fidler [8] mention. However, as Colson [3] points out, much caution must be taken when using such automated collocation extraction tools, as there may be collocations consisting of more than two words. Therefore, perhaps his CPR method may come in handy, or a similar method. The future will tell.

References

1. Baker, P. Using corpora to analyze gender. London: Bloomsbury Academic (2014).
2. Čech, R. Language and ideology: quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949–2011). *Quality & Quantity* 48:2, 899–910 (2014).
3. Colson, J.-P. Set Phrases around GLOBALIZATION: An experiment in corpus-based computational phraseology. In F. Alonso, Almeida, I. Ortega Barrera, E. Quintana Toledo & M.E. Sánchez Cuervo (eds.), *Input a Word, Analyze the World. Selected Approaches to Corpus Linguistics*, 141–152. Newcastle: Cambridge Scholars Publishing. (2016).
4. Czech National Corpus wiki, <http://wiki.korpus.cz/doku.php/en:cnk:syn:verze3>, last accessed 2017/06/13.
5. Czech National Corpus wiki, <http://wiki.korpus.cz/doku.php/en:cnk:syn:verze4>, last accessed 2017/06/13.
6. Elmerot, I.: Är en zigenare mer oanpassningsbar än en rom? En pilotstudie om kollokationer för orden *Cikán* och *Rom* i modern, tjeckisk tidningstext. *Slovo. Journal of Slavic Languages, Literatures and Cultures* 57, 99–110 (2016).
7. Elmerot, I.: *Hodný, zlý a ošklivý (The Good, the Bad and the Ugly) : The representation of three minority groups in printed media discourse from the Czech Republic*. Bachelor's thesis, Stockholm University (2017).
8. Fidler, M.: The others in the Czech Republic: Their image and their languages. *Multilingualism and Minorities in the Czech Sociolinguistic Space*, a special issue of the *International Journal of the Sociology of Language* 238, 37–58 (2016).
9. Nash, R. Exhaustion from explanation – Reading Czech gender studies in the 1990s, *European Journal Of Womens Studies*, 9, 3, 291–309 (2002).
10. Needle in a Haystack research outputs, <https://www.brown.edu/research/projects/needle-in-haystack/404>, last accessed 2017/06/14.
11. Oates-Indruchová, L. Unraveling a Tradition, or Spinning a Myth? Gender Critique in Czech Society and Culture, *Slavic Review*, 75, 4, 919–943 (2016).
12. Pallas, H.: *Oanpassbara medborgare: historien om förföljelsen av de tjeckiska romerna*. Atlas, Stockholm (2016).
13. Slavičková, T.: Investigating nepřizpůsobivý as a key word in critical analysis of Czech press reports on Roma. *Korpus-Gramatika-Axiologie* 11, 69–82 (2015).
14. Váňa, T. Language power potential. *The Annual of Language & Politics and Politics of Identity*, vol. VI (2012).

Teasing Apart Russian Idioms And Homonymic Compositional Expressions

A Word Embedding Approach

Marina Pchelina¹ and Jae-Woong Choe²

^{1,2} Korea University, Seoul 02841, Republic of Korea
{marina_p, jchoe}@korea.ac.kr

Abstract. We test and evaluate a context-based method for MWEs compositionality detection that utilizes word embeddings. Embeddings for individual words are used to get representations of target expressions and their context. In making a judgement on compositionality/idiomaticity of an expression, our algorithm relies on the expectation that when a MWE is used literally constituents retain their original meanings and are semantically related to surrounding context words, which is not normally true of idiomatic usage. Context is the only factor that decision about compositionality of an expression is based on, which adds to simplicity and universality of the method. We test the recently introduced idea of applying Principal Component Analysis to represent semantic composition and argue that its performance is at least as good as the state of the art.

Keywords: Compositionality Detection, Word Embeddings, PCA.

1 Introduction

Non-literal uses of language pose a significant problem for many NLP tasks, especially when they look exactly the same as literal, consider expressions in bold in sentences 1) and 2).

- 1) *Jinny was so startled that she nearly **kicked the bucket** over.*
- 2) *Chatterton and Fagg and a few more like them who've since **kicked the bucket**.*

For efficient natural language understanding, there should be a way to automatically tell apart those two cases. Great attention has been paid to this issue and considerable results have been achieved, but the majority of proposed methods rely on hand-crafted

lexicons and databases. Such approaches tend to be restricted in terms of language and range of expressions they can detect.

We believe that it is local context that should be the main cue in deciding on compositionality of a given expression. We focus on solutions that are simple and general enough to cover broader types of non-compositional uses of expressions and, theoretically, to work on any language.

2 Method

The basic intuition behind our method is that when an expression is meant literally it is compositional in the sense that its constituents a) retain their original meaning and b) are usually semantically related to context words [1]. Therefore, if representation of a phrase used literally is compared to representation of its local context, they are expected to be similar, and in case of non-compositional usage, otherwise.

While representing individual words with embeddings has proved to be highly effective, computationally representing larger pieces of text is an ongoing issue, simple vector averaging being currently the state of the art. The newest idea is to draw on geometrical aspects of word vector spaces and apply Principle Component Analysis to represent context as a linear subspace [2].

The method, which we adopt here, consists in applying PCA to the linear space constituted by word embeddings and finding a linear subspace that the first few principle components create, so that the original data is represented in a compact way with minimal loss. The target expression's vector is then projected into the linear subspace, and that projection is compared with the original vector [5]. In all cases, we use cosine similarity to compare vectors and calculate threshold as the mean of all similarity scores in our data set.

3 Experiment Setup

It was important to look specifically at MWEs that could be used literally and idiomatically without notable bias toward any of the two and regardless of a particular grammatical form, for it seems those pose the most difficulty for NLP tasks. That is why they had to be picked out manually¹. We disregarded syntactic structure and length of the MWEs to check universality of the method, so the list included expressions of the form ADJ+N, PR+N, V+PR+N and others².

We had 50 target expressions, with one instance of compositional and one of non-compositional use for each, making it 100 cases overall. For context, 10-15 content words were considered from both sides of a MWE in question³. We obtained original 300-dimensional word embeddings from rusvectors.org, specifically, from the model

¹[We used wikitionary.org, russkiyyazik.ru.

²[Because only content words got embeddings, the number of embeddings per expression was 1-3.

³[Russian National Corpus was queried for expressions to obtain contexts [6].

trained with word2vec's CBOW algorithm on 900 mln words web corpus, because it performed best on simple word similarity tests [3].

4 Results and Discussion

In the first set of experiments, target expressions were represented as compound vectors calculated over all constituents with simple averaging, multiplication and PCA; context was represented⁴ with either averaging or PCA. Results are reported in Table 1.

Table 1. Accuracy with MWEs represented as compound vectors.

Phrase/context	PCA	Average
PCA	0.87	0.55
Average	0.49	0.89
Multiplication	0.79	0.52

It is immediately clear that better results are obtained when both expressions and context are represented in the same way, which apparently has to do with the fact that different composition methods provide different resulting vectors – averaging introduced negative similarity values, while PCA did not.

In the second set of experiments, we aimed to see if MWEs could be represented as an embedding of their single constituent with a minimum, maximum or furthest from the mean (extreme) score. Since individual words can have different degrees of idiomaticity within an expression, it would allow for fuller coverage of potentially idiomatic language, e.g. partly compositional expressions. Results are reported in Table 2.

Table 2. Accuracy with MWEs represented as single constituent embeddings.

Phrase/context	PCA	Average
Min	0.76	0.79
Max	0.88	0.86
Extreme	0.89	0.87

This time, scores seem to be more stable, with PCA being slightly more accurate than averaging. Overall, 17 phrases out of 50 (34%) were labelled correctly in all tests, a few examples are: *закинуть удочку* “throw the fishing rod”, *сидеть на чемодане* “sit on a suitcase”, *прижаться к стене* “press against the wall”⁵. As for the rest, where there were false predictions, overwhelmingly, they formed patterns.

⁴□ Multiplication was found not suitable for representation of larger chunks of language.

⁵□ Idiomatic meanings are, respectively, “make a cautious inquiry”, “be ready to leave at any moment”, “expose sb.”.

Firstly, relying on minimal similarity scores did not provide accurate enough results, which might be due to fact that it is fairly easy for two randomly taken words to have similarity score close to zero, which just means irrelevance. Judging by maximum scores gave better results. Secondly, scores obtained with PCA for context representation are higher than with averaging: mean accuracy scores after all experiments are 78 and 74.6, respectively. Finally, error analysis showed that wrong predictions were given in cases where target expressions either were used in a very peripheral role and did not have enough similar words around, or contained very frequent or general words that could not get distinctive embedding representations.

As for the limitations, however, precision scores in our experiments did not vary drastically, which might indicate that it is the embeddings themselves that should be improved, possibly with exemplar-based models, which allow multiple embeddings per word and have a potential to improve accuracy for compositionality detection [4].

5 Conclusion

We found the context-based test for telling apart compositional and non-compositional uses of same MWEs to be fairly effective while being potentially independent of a particular language. Additionally, our experiments hint that PCA could be considered as a suitable way of representing context, because it can take into account tens and hundreds of factors (here, words) and, at least on our dataset, it outperformed the state-of-the-art averaging. Another finding is that when potentially idiomatic expressions are not cherry-picked to be of the same kind, it might be reasonable to judge them by single constituent embeddings.

References

1. Doumas, L. A. A., Hummel, J. E.: Modeling human mental representations: What works and what doesn't and why. In: Holyoak, K. J., Morrison, R. G. (eds.) *The Cambridge handbook of thinking and reasoning*, pp. 73–91. Cambridge University Press, Cambridge, UK (2005).
2. Gong, H., Bhat, S., Viswanath, P.: Geometry of Compositionality. In: *AAAI Conference on Artificial Intelligence*, <https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14699>, last accessed 2017/06/21.
3. Kutuzov, A., Kuzmenko, E.: WebVectors: a toolkit for building web interfaces for vector semantic models. In: Ignatov, D. et al. (eds.) *AIST 2016, CCIS*, vol 661, pp. 155-161. Springer, Cham (2017).
4. Reddy, S., McCarthy, D., Manandhar, S., Gella, S.: Exemplar-based word-space model for compositionality detection: Shared task system description. In: *ACL 2011 Workshop on Distributional Semantics and Compositionality*, pp. 54-60. ACL, Portland, USA (2011).
5. Widdows, D.: Semantic vector products: Some initial investigations. In: *Proceedings of QI 2008, 2nd International Symposium on Quantum Interaction*. College Publications, Oxford, UK (2008).
6. Russian National Corpus Homepage, <http://www.ruscorpora.ru>, last accessed 2017/06/15.

Observations on phonetic and metrical patterns in Spanish-language proverbs

Jordi Martínez Martínez, Gemma Bel Enguix and Liliana Torres Floress

Universidad Nacional Autónoma de México, Instituto de Ingeniería,
Grupo de Ingeniería Lingüística, Mexico City, Mexico
jor.mtzmtz@gmail.com
gbele@iingen.unam.mx
lilianatorres0412@gmail.com

Abstract. This paper aims at starting a new research line on Spanish proverbs with the help of computational tools. The first step to reach this goal was the compilation of a corpus of proverbs from existing digital sources. Later, a phonetic study of the items of the corpus was carried out, approaching the syllabic structure and metrical features, as well as regularities and patterns in the configuration of paremiological units. For the future, we expect to complete this work by extending the research to syntactic and semantic features and build a system capable of searching for analogue structures in large collections of texts.

Keywords: paremiology, poetic meter, rhyme, Spanish proverbs

1 Introduction

1.1 Phonetic and metrical patterns in proverbs

Exploring the relationship between metrical or phonetic features and proverbs is, by no means, a new approach in paremiological studies: reflections on the nature of such kind of patterns are common in literature, more often than not, as casual remarks which may still await for a more rigorous approach [1].

Nevertheless, phonetic regularities found in proverbs have been hypothesized to serve as an instrument of internal organization [16], as a sign of an utterance's *proverbiality* [1,2] or as a mnemotechnical device [6].

Empirically oriented approaches to the study of phonetic regularities in proverbs seem to be scarce, and, as such, their possible applications in paremiography and paremiology have not been explored, i.e. as criteria for search queries in databases or as variables in automatic or manual identification tasks.

In Spanish-language proverbs, such patterns have been observed to occur as rhyme (specially as phrase-final assonance), isosyllabism, binary (prosodic, syntactic) structure, and a certain intonational pattern. Several works approach the topic, among others [1,2,3,4,7,9,12,13,15].

1.2 Objectives

In this paper we aim to go deeper into a corpus-based computational approach to Spanish paremiology, that has not been enough developed so far.

Moreover, we have the goal to corroborate some remarks on the phonetic regularities found in Spanish proverbs. The following hypotheses were considered:

1. Binary proverbs were expected to predominate over non-binary proverbs.
2. Rhymed proverbs were expected to predominate over rhymeless proverbs.
3. In binary and rhymed proverbs, assonance was expected to be the dominant type of rhyme, over consonance.
4. Proverbs formed by two or more units were expected to exhibit isosyllabicity.
5. Octosyllabic proverb hemistichs were expected to predominate over non-octosyllabic hemistichs.

To do so, we describe a basic methodology which consisted on the creation of a corpus of proverbs with metrical and phonetic labeling.

2 Methodology

2.1 The corpus

Three on-line databases were used as basis for our corpus: *Refranero mexicano* [11], and the Spanish sub-corpora of both *Refranero multilingüe* [14], and *ParemioRom* [8]¹.

The selection of these repertoires was based on the following criteria: (a) documentation of their materials, (b) their digital nature, which facilitated the processing of the proverbs extracted from each of them, and (c) the diatopic variety associated with the sources from which the proverbs were extracted.

The initial corpus consisted of a total of 4734 proverbs. To filter repeated or similar ones, the FuzzyWuzzy library [5] was used, called from the Python programming language (Version 3.6.1).

FuzzyWuzzy uses Levenshtein's distance to compare strings of characters [5]. The similarity parameter for matched strings was set at 85%. This allowed us to identify groups of proverbs that exhibited similarity equal or greater to the set parameter.

Out of any group of matched proverbs, we decided to keep those belonging to *Refranero mexicano*; while in those instances where the matched strings were made up of proverbs from the Peninsular Spanish sources, we decided to keep the proverbs that belonged to *ParemioRom*.

In the instance of a group of matched strings that belonged to a single source, it was decided to retain the first selected member of such group.

After applying this filter to the initial corpus, the repertoire was reduced to 4620 proverbs. The composition of the corpus is set out in Table 1.

¹ Some proverbs were not taken into account: dialogue proverbs, incomplete or censored proverbs and some blank-slot proverbs which contained its variants in the same string (i.e., *El que siembra, o siembre, o sembró, su maíz, que se coma su pinole*).

Table 1. Composition of the corpus.

Source	Region	Proverbs	%
<i>Refranero mexicano</i>	Mexico	1975	42.75
<i>Refranero multilingüe</i>	Spain	1622	35.11
<i>ParemioRom</i>	Spain	1023	22.14
Totals	-	4620	100.00

2.2 Corpus processing

The separation of proverbs into smaller units was carried out in two phases. Firstly, the punctuation marks already present in the proverbs were taken advantage of: commas (,) and semicolons (;) were combined with adjacent vertical bars (|) automatically. The second part consisted in a manual revision of the output, with the aim of dismissing irrelevant divisions (i.e. some vocatives) and dividing the proverbs that did not signal internal divisions through punctuation, but with other types of boundaries, such as rhyme, conjunctions or, to a lesser extent, other signs not initially considered, such as colons (:).

We agreed to denominate such internal prosodic constituents as *hemistichs*, following Pérez Martínez [10], in an analogy of its use in metrical studies.

The divided proverbs were processed with a series of simple rule-based scripts called from the Python programming language. These take as input a divided proverb, on whose hemistichs apply a set of syllabification rules, a stressed syllable identifier², and a resyllabification cycle of whose output is inferred a binary string (interpreted as series of stressed (1) and unstressed (0) syllables), equivalent to the metrical structure of the hemistichs, as illustrated by the examples in Table 2.

Table 2. Examples of inputs and outputs of the rule-based scripts. Vertical bars (|) denote hemistich boundaries; stressed and unstressed syllables are represented by 1 and 0, respectively.

Divided proverb	Metrical structure
A cada pajarillo le gusta su nidillo	0100010 0100010
Pobre y ticolotero, no te irás con dinero	1000010 0010010

From this process, the following variables were obtained for each hemistich: number of *graphical syllables*, number of *beats*, position of the last stressed syllable and metrical structure.

² It should be noted that the scripts employed in this investigation automatically consider any monosyllabic word as unstressed.

The identification of metrical syllables entailed the distinction of at least four different variables: *phonological syllables*, *graphical syllables*, *beats*, and *metrical syllables*.

Phonological syllables were assumed to be equivalent to the underlying syllabical representation of the utterance (the proverb), which was not taken into account in this study; *graphical syllables* were obtained through the output of the syllabification cycle of the scripts used and correspond to the syllabic divisions of isolated word tokens; *beats* were obtained through the output of the rules of resyllabification cycle and correspond to the units of a string of strong (stressed) and weak (unstressed) syllables. Finally, *metrical syllables* correspond to the output of the metrical analysis of the previous output, according to Spanish poetic meter conventions.

The number of metrical syllables (MS) per hemistich was determined on the basis of the number of beats per hemistich plus the application of a subtraction or addition operation given the following situations: $MS = \text{beats} - 1$, if the last stressed beat is the third-to-last pulse of the string; $MS = \text{beats}$, if the last stressed beat is the penultimate beat of the string; $MS = \text{beats} + 1$, if the last stressed beat is the last beat of the string.

Rhyme labeling was done manually. Two types of rhyme were distinguished: assonance and consonance. Only hemistich-final rhymes were taken into account. Consonance was defined as the perfect match of the phonic material contained between the nucleus of the last stressed syllable and the right edge of two or more hemistichs. Assonance, on the other hand, was taken as the partial coincidence of phonic material contained between the nucleus of the last stressed syllable and the right edge of two or more hemistichs, restricted only to syllabic nuclei.

Consonant rhymes were required to match perfectly the syllabic nuclei involved, including vowels /i, u/ associated to the left edge of syllabic nuclei in ascending diphthongs.

Only one dialectal feature was taken into account in the labeling of rhymes, that of the distinction between phonemes /s/ and /θ/ in most European varieties of Spanish. Therefore, some rhyme schemes were labeled as assonant or consonant depending on the geographical region associated with the source material (Mexico, in the case of *Refranero mexicano*, or Spain, in the case of *Refranero multilingüe* and *ParemioRom*).

The finest labeling was concentrated only on proverbs of two hemistichs, the most numerous, where the two rhyme types are distinguished. Proverbs of three or more hemistichs that presented a type of rhyme on the edge of their hemistichs were only labeled as *rhymed*. Proverbs formed by a single hemistich were considered to be rhymeless.

3 Results

3.1 Hemistichs

Number of hemistichs. As shown in Table 3, binary proverbs, formed by two hemistichs, made up almost three quarter parts (74.94%) of the corpus, followed

by single hemistich proverbs, proverbs formed by three hemistichs and proverbs formed by four hemistichs. Proverbs formed by five or more hemistichs made up only 0.86% of the corpus. A total of 9561 hemistichs were counted.

Table 3. Proverbs by number of hemistichs.

Hemistichs	Proverbs	%
1	572	12.38
2	3463	74.94
3	345	7.47
4	200	4.33
5	22	0.48
6	14	0.30
8	3	0.06
10	1	0.02
Total	4620	100.00

Hemistich length. As shown in Table 4, octosyllabic hemistichs were the most numerous, making up 22.62% of the corpus, but, unlike the distribution of proverbs by number of hemistichs, greater variety was found, especially among the ones with a length of eight metrical syllables or less (82.84% of the corpus).

Table 4. Hemistichs by number of metrical syllables.

MS	Hemistichs	%
<4	539	5.64
4	996	10.42
5	1555	16.26
6	1429	14.95
7	1238	12.95
8	2163	22.62
9	627	6.56
>9	1014	10.61
Total	9561	100.00

3.2 Rhyme

Presence of hemistich-final rhyme. Rhymed proverbs made up a slight majority (55.37%) of the corpus over rhymeless proverbs (44.63%), as seen in Table 5.

Rhymed proverbs by number of hemistichs. Distribution-wise, there seems to be a great correlation between binary structure and presence of rhyme, as up to 79.44% of rhymed proverbs were formed by two hemistichs (Table 6).

Table 5. Proverbs by presence of rhyme.

Rhyme	Proverbs %	
Rhymed	2558	55.37
Rhymeless	2062	44.63
Total	4620	100.00

Table 6. Rhymed proverbs by number of hemistichs.

Hemistichs	Rhymed proverbs %	
2	2032	79.44
3	298	11.65
4	190	7.43
5	20	0.78
6	14	0.55
8	3	0.12
10	1	0.04
Total	2558	100.00

Type of rhyme. The most common type of rhyme found in rhymed proverbs formed by two hemistichs was consonance (53.40%), with a slight difference over assonance (46.6%), as shown in Table 7.

Table 7. Types of rhyme in binary proverbs.

Rhyme	Proverbs %	
Assonance	947	46.60
Consonance	1085	53.40
Total	2032	100.00

3.3 Isosyllabicity

Isosyllabicity was exhibited by only 1052 proverbs (22.77%) out of a total 4620. Then again, a strong affinity between this feature and binary structure was found, as 1012 (96.2%) out of 1052 isosyllabic proverbs were formed by two hemistichs.

Metrical length seems to play an important role as well, as 47.05% of isosyllabic proverbs contained octosyllabic hemistichs (Table 8).

3.4 Summary

Regarding our hypotheses, binary proverbs did indeed predominate over non binary proverbs, forming up to 74.94% of the corpus. Octosyllabic hemistichs

Table 8. Isosyllabic proverbs found in the corpus.

MS	Hemistichs						Totals	%
	Two	Three	Four	Five	Six			
2	3	0	0	0	0		3	0.29
3	12	0	2	0	0		14	1.33
4	95	3	1	0	0		99	0.93
5	154	3	7	0	0		164	9.41
6	129	3	4	0	1		137	15.59
7	77	3	0	0	0		80	7.6
8	482	8	5	0	0		495	47.05
9	32	0	0	0	0		32	3.04
10	13	0	0	0	0		13	1.24
11	9	0	0	0	0		9	0.86
12	4	0	0	0	0		4	0.38
13	0	0	0	0	0		0	0
14	0	0	0	0	0		0	0
15	1	0	0	0	0		1	0.1
16	1	0	0	0	0		1	0.1
Totals	1012	20	19	0	1		1052	100.00

(22.62%) were the most common metrical type in our corpus, although there seems to exist a variety of metrical types, especially in hemistichs of eight or less metrical syllables. Rhymed proverbs were found to be more common than rhymeless proverbs, but only by a slight margin. Consonance, in turn, prevailed over asonance, but in similar conditions. Isosyllabicity exhibited a rather restricted presence in our corpus: only 22.77% of the proverbs were isosyllabic, and such feature seems to be strongly correlated to binary structure (96.2% of isosyllabic proverbs) and octosyllabic hemistichs (47.05% of isosyllabic hemistichs).

4 Conclusions and future work

This work has been focused on a preliminary quantitative analysis of internal structure, isosyllabicity and rhyme in a corpus of Spanish proverbs. The results seem to confirm this genre has strong metrical and prosodic features. However, some results, like the preference for consonant rhyme and the low presence of isosyllabicity, are surprising considering the mainstream opinions so far.

The future work has to include the identification of contexts where the proverbs appear and the identification in corpora of prosodic structures with the same patterns than the ones analyzed in here. The main objective for this work is the automatic retrieval of proverbs in large collection of texts.

Acknowledgements We thank Octavio Augusto Sánchez, author of the Python scripts used in this investigation, for the permission granted for using his original work, and the mexican Red Temática en Tecnologías del Lenguaje for the support to our project.

References

1. Anscombre, J.C.: Estructura métrica y función semántica de los refranes. *Paremia* 8, 25–36 (1999)
2. Arora, S.: The perception of proverbiality. In: Mieder, W. (ed.) *Wise Words: Essays on the Proverb*, pp. 3–29. Routledge, Abingdon-on-Thames (1984/2015)
3. Baehr, R.: *Manual de versificación española*. Gredos, Madrid (1973)
4. Casares, J.: La frase proverbial y el refrán. *Revista Universidad Pontificia Bolivariana* 27(95), 36–49 (1964)
5. Cohen, A.: *FuzzyWuzzy. Fuzzy String Matching in Python* (2017), <https://github.com/seatgeek/fuzzywuzzy>
6. Corpas Pastor, G.: *Manual de fraseología española*. Gredos, Madrid (1996)
7. Crida Álvarez, C.A., Sevilla Muñoz, J.: La problemática terminológica en los estudios paremiológicos. *Anuari de filologia. Estudis de lingüística* 5, 67–77 (2015)
8. Gargallo Gil, J.E., Álvarez Pérez, X.A.: El Proyecto Paremiom. *Refranes meteorológicos y geoparemiología romance. Estudis Romànics* 36, 313–324 (2014)
9. Navarro Tomás, T.: *Manual de entonación española*. Guadarrama, Madrid, 4 edn. (1974)
10. Pérez Martínez, H.: *Refranero mexicano*. Lengua y Estudios Literarios, Academia Mexicana de la Lengua, Fondo de Cultura Económica, México (2004)
11. Pérez Martínez, H.: *Refranero mexicano* (2008), <http://www.academia.org.mx/universo:lema/obra:Refranero-mexicano>
12. Quilis, A.: *Métrica española*. Ariel Letras, Ariel, Madrid, 15 edn. (2013)
13. Sevilla Muñoz, J., Crida Álvarez, C.: Las paremias y su clasificación. *Paremia* 22, 105–114 (2013)
14. Sevilla Muñoz, J., Zurdo Ruiz-Ayúcar, M.I.T.: *Refranero multilingüe* (2009), <http://cvc.cervantes.es/lengua/refranero/>
15. Taylor, A.: *The Proverb*. Harvard University Press, Cambridge, MA (1931)
16. Toporov, V.N.: Folk Poetry: General Problems. In: Sebeok, T. (ed.) *Current Trends in Linguistics*, vol. 12. Linguistics and Adjacent Arts and Sciences, pp. 684–739. Mouton, The Hague (1974)

Towards a Corpus-lexicographical Discourse Analysis

Emma Franklin

Lancaster University, Bailrigg, Lancaster, LA1 4YW

Abstract. This working paper presents the progress made thus far in the development of a corpus-lexicographical approach to discourse analysis, more specifically the application of Hanks' [5, 6] Corpus Pattern Analysis (CPA) procedure to a (critical) discourse analysis task. The theoretical basis of CPA is explained, followed by some practical applications of CPA, namely lexicography and the proposed method of discourse analysis. Examples are taken from an ongoing investigation into the use of 'killing' verbs in contemporary British English, which draws upon two corpora: the British National Corpus (BNC) and the animal-themed 'People', 'Products', 'Pests' and 'Pets' (PPPP) corpus [8]. Preliminary findings suggest that a CPA-assisted, or corpus-lexicographical, discourse analysis is one with a strong theoretical basis, whose transparency and systematicity empowers the analyst to make precise and persuasive arguments.

Keywords: Discourse Analysis, Corpus Pattern Analysis, Lexicography.

1 Introduction

Current methods of discourse analysis are numerous and wide-ranging, to the point that such terms as "discourse analysis" and even "critical discourse analysis" are almost meaninglessly vague. The word "discourse" is, itself, polysemous, and the aim of this paper is not to attempt to untangle its nuances. Rather, this work seeks to carve out a new, potential route to understanding meaning in discourse using corpus-lexicographical methods, and some of the progress made thus far in this approach is presented here.

In over-simplified terms, for the purposes of this brief discussion, "discourse" is understood to refer in its non-countable form to "language in use", and in its countable form – or "big D" form [3] – to a "conventional practice". Critical Discourse Analysis (CDA) is defined most simplistically as "discourse analysis 'with attitude'" [11, p. 96]. More specifically, it seeks "to uncover and de-mystify certain social processes in this and other societies, to make mechanisms of manipulation, discrimination, demagoguery and propaganda explicit and transparent" [12, p. xiv].

Though discourse is arguably language above the level of sentence or clause, it is constituted by much smaller units of language which need to be analysed as such. Corpus-assisted discourse analyses already take this route, traditionally via the use of statistically generated word lists, collocates, keywords, and so on [2]. The approach proposed here does not make use of most of these methods, but does rely on corpus data and uses corpus analysis software to generate concordance lines for manual inspection. It employs Hanks' [5, 6] Corpus Pattern Analysis procedure, the output of which is

considered in light of existing literature as well as historical and political context. The result is an empirical, semantically motivated, critical analysis of argument structure across text types. It is lexicographical in that it entails the creation of a corpus-based lexicographical entry for a given word, as per the *Pattern Dictionary of English Verbs* (PDEV), introduced in Section 2. It does not currently make use of automated natural language processing, e.g. semantic parsing, but instead relies on manual analysis for accurate classification of arguments and delimitation of word senses.

The rest of the paper is organised as follows. Section 2 outlines the theoretical background and main features of Corpus Pattern Analysis (CPA), followed by some examples of practical applications of CPA, namely lexicography and corpus-lexicographical discourse analysis, including a short case study. Section 4 concludes the paper with a very brief summary of the potential rewards and challenges of taking this approach.

2 Corpus Pattern Analysis and the PDEV

Corpus Pattern Analysis (CPA), developed by lexicographer Patrick Hanks, seeks “[to elucidate] the relationship between syntagmatic patterns and activated meanings” [5, p. 92]. Following in the Neo-Firthian tradition, CPA examines the behaviour of words in their contexts, and in doing so establishes the linguistic patterns with which word senses are associated. Words, Hanks argues, do not have meaning but “meaning potential”; their meanings are only activated by the lexical patterns in which they exist [5] and, like Sinclair, Hanks finds meaning to be inextricably linked to form (cf. [10]). So far, CPA has mostly been employed in lexicography, namely the *Pattern Dictionary of English Verbs*¹, under the *Disambiguation of Verbs by Collocation* (DVC) project².

CPA is underpinned by Hanks’ [6] Theory of Norms and Exploitations (TNE), which centres on the phenomenon of prototypical language use (norms) and exploitations of these norms. CPA takes a similar approach to that of the COBUILD [9] and Hector [1] projects, and bears some similarities to Construction Grammar [4]. However, CPA is more concerned with lexical semantics, and it relies wholly on corpus evidence of usage. A *pattern*, in the CPA sense, “consists of a valency structure ... together with sets of preferred collocations” [6, p. 92]. Patterns can be *norms* (patterns of normal, conventional, everyday usage) or *exploitations* (creative patterns of language use), though the distinction between the two is not an absolute one [6, p. 4].

Following Pustejovsky [7], CPA employs *semantic types*, which are logical constructs for groups of lexical items, arranged in a hierarchical semantic ontology. For example, the verb *sip* selects as its direct object lexical items such as *beer*, *water*, *whiskey*, and *tea*, which form a lexical set represented in the *CPA Ontology*³ by the semantic type of [[Beverage]]. A [[Beverage]] is a [[Liquid]] is a [[Fluid]] is [[Stuff]] is an [[Inanimate]] is a [[Physical Object]], and so on. The CPA Ontology is unique, in that it was not devised *a priori*, but instead was progressively built and altered during the

¹ <http://pdev.org.uk>

² <http://gtr.rcuk.ac.uk/projects?ref=AH/J005940/1>

³ <http://pdev.org.uk/#onto>

course of the project, and can be considered to be data-driven and specific to the corpus upon which it is based (the British National Corpus (BNC), predominantly).

The semantic types from the CPA Ontology occupy argument slots, for example, the subject, object and prepositional object slots. CPA patterns are anchored to *implicatures*, which form an integral part of a word’s “syntagmatic profile” [6] and which describe the entailment of a particular pattern. For example, the most common pattern associated with the verb *drink* is listed in PDEV as (1), with the implicature, (2).

[[Human]] drink [[Beverage]] ({up | down}) (1)

[[Human]] takes [[Beverage]] into the mouth and swallows it (2)

Words in double square brackets are semantic types. The round brackets in (1) denote optionality, i.e. in this instance, an adverbial is not always present. Curly brackets denote specific lexical items; in this case, *up* and *down* cannot be substituted.

Finally, it should be noted that CPA is concerned with conventionality; it does not classify what is *possible* in language, but what is *typical*. CPA patterns, like semantic types, represent central, canonical forms of language as opposed to all potential ones.

3 Putting CPA into Practice

3.1 Doing Lexicography with CPA

The standard CPA procedure is described in detail elsewhere [5, 6]. To summarise:

- The analyst generates a concordance for a node word and takes a random sample of concordance lines, starting with around 250. In the interests of producing generalisable results, a large, general-language corpus is used as a source of data.
- Lines are manually grouped together based on their shared syntagmatic properties – their valency, arguments, presence or absence of adverbials, etc. This involves identifying norms (prototypical phraseology) and from there deciding which instances are likely to be exploitations. Establishing such patterns “calls for a great deal of lexicographic art” [5, p. 88].
- The analyst sorts these grouped lines into patterns by tagging each line with a pattern number, and then writing up the patterns and their implicatures into a kind of dictionary entry (see Fig. 1).

Using CPA for lexicography results in an empirically well-founded dictionary entry which gives the proportions of different word senses in the data. In other words, meaning becomes somewhat measurable. Lexicography that more accurately represents natural language use is valuable not only for language learners and teachers, but also for computational linguists interested in semantic probabilities for the purposes of word-sense disambiguation. Measuring the presence of word senses in “general” language also makes it possible to compare meanings across texts.

drink Add pattern Stretch Shrink more Concordance (OEC , enTenTen12 , BNC) Ontology Renumber				
Sample size	250	(out of 1844)	Semantic class	Drinking
Status	complete	▼	Difficulty	▼
#	%	Pattern & primary implicature		
1.	40.40%	[[Human]] drink [[Beverage]] ((up down)) [[Human]] takes [[Beverage]] into the mouth and swallows it		
2.	3.60%	[[Animal]] drink ([[Water]]) [[Animal]] takes ([[Water]]) into the mouth and swallows it		
3.	32.80%	[[Human]] drink [NO OBJ] ((heavily excessively more than ...)) [[Human]] drinks alcoholic beverages, typically in excessive amounts In many cases, [[Human]] has health and social problems as a result of this		
4.	0.80%	[[Human]] drink [[Eventuality = Experience]] {in} pv [[Human]] eagerly cognitively and emotionally assimilates [[Eventuality = Experience]]		

Fig. 1. Non-public-facing PDEV entry for the verb *drink*.

3.2 Doing Discourse Analysis with CPA

CPA for lexicographical purposes involves the use of large, general, reference corpora, such as the BNC. As discourse analysts tend to be interested in one particular type of discourse, or how discourses differ from one another, a corpus-lexicographical discourse analysis will also involve carrying out CPA on a specialised corpus or subcorpus. For highly specialised or technical language, a new ontology of semantic types may have to be created from scratch. In most cases of contemporary British English investigations, however, the PDEV's CPA Ontology will act as a useful starting point.

By way of example, my ongoing doctoral research is a corpus-assisted investigation into 'killing' phraseology in contemporary British English, with a particular focus on human-animal relations and how 'killing' events are represented across discourses. Killing is a process involving multiple participants, e.g. agents and patients, or 'killers' and 'killees', making verbs an ideal place to start; predicates act as the pivot of a clause, and so to analyse a verb is to uncover the arguments it governs. As CPA is systematic, empirical, and particularly well-suited to verbs, it forms the basis of the analysis. The specialist corpus used in this project is the 'People', 'Products', 'Pests' and 'Pets' (PPPP) corpus [8]. It comprises almost 9 million words of animal-related discourse in contemporary British English from a range of text types and genres; see Table 1 for the composition details. The BNC is used as a reference corpus, and PDEV entries are referred to where available. Corpus software AntConc⁴ is used for generating concordances, and Microsoft Excel is used for sampling, tagging, sorting and analysing.

The procedure for CPA-assisted discourse analysis, in this project, is as follows:

- Consult the PDEV to see whether the verb in question already has an entry. If not, take a 250-line random sample of the verb's concordances from a POS-tagged version of the BNC and carry out CPA using the CPA Ontology.
- Take a 250-line random sample of the verb's concordances from the POS-tagged version of the PPPP corpus and carry out CPA, using the PDEV/BNC patterns as

⁴ <http://www.laurenceanthony.net/software.html>

a loose guide. The CPA Ontology is a basis from which to start creating a new ontology tailored to the specialised corpus over time; this is an iterative process.

- Compare occurrence and distribution of patterns across the two samples, and observe differences in semantic types. Discuss these in context of the literature.

Table 1. PPPP Corpus Composition, from [8].

Subcorpus	No. of files	No. of types	No. of tokens
Broadcasts	83	19835	614378
Campaign literature	470	16488	306680
Legislation	843	10201	627127
Food websites	258	7503	87118
Journals	1609	93567	5698531
News	1023	28777	466340
MO Project contributions	103	9931	174938
Focus groups	19	8277	229059
Interviews with text producers	17	8068	157664
Interviews with dog keepers	19	8698	309719
Total	4444	211345	8671554

Case study: *destroy*

Given that it refers in some contexts to killing, *destroy* was selected as a candidate for analysis. The above steps were carried out and the same patterns were found in both the PPPP and the BNC samples, though in different proportions. Pattern 1, which refers to the attacking or damaging of a physical object, is equally prominent in both samples. Patterns 2 and 3, which refer to abstract senses of destruction (e.g. of confidence, and a human opponent, respectively), are less prominent in the PPPP corpus sample. The proportion of Pattern 4, which refers to the killing of animals (and fetuses) by humans, is (as expected) far higher in the PPPP sample. See Fig. 2 for the pattern details.

#	BNC	PPPP	Pattern & primary implicature
1	60.40%	59.49%	[[Human Animal Institution Event Artifact]] destroy [[Physical_Object]] [[Human Animal Institution Event Artifact]] damages or attacks [[Physical_Object]] until it is completely ruined
2	28.00%	5.06%	[[Anything]] destroy [[Property Abstract_Entity State_of_Affairs]] [[Anything]] causes [[Property Abstract_Entity State_of_Affairs]] to no longer exist
3	5.60%	1.27%	[[Human 1]] destroy [[Human 2 Human_Group]] [[Human 1]] utterly defeats [[Human 2 Human_Group]]
4	3.20%	34.18%	[[Human]] destroy [[Animal Animal_Group Fetus]] [[Human]] kills unwanted [[Animal Animal_Group Fetus]]

Fig. 2. Patterns and implicatures for the verb *destroy*, as found in the BNC and PPPP samples

CPA makes it possible to distinguish not only inter-pattern differences (the pattern boundaries) but also intra-pattern variation, including anomalies. While pattern distribution across corpora is useful, the real value to discourse analysis is found within the boundaries of the patterns themselves. For example, *destroy* in Pattern 1 takes as its object a Physical Object. These are typically physical objects of the inanimate variety, particularly artifacts and buildings, which is intuitive given the etymology of *destroy* (from the Latin *destruere*, lit. “unbuild”). In the BNC these tend to be houses and vehicles, and in the PPPP corpus these are more often nests, setts and other animal homes.

Less predictable and less straightforward to deal with are those examples which feature unusual arguments. Take, for example, lines (3-6), found in the PPPP sample.

Did a meteorite really *destroy* the dinosaurs? (3)

[...] otters have *destroyed* entire populations of large fish in some fisheries [...] (4)

[...] everything will be *destroyed*, the animals, the plants, the water, the land. (5)

Scottish Ministers may (a) cause to be *destroyed* any semen, egg or embryo [...] (6)

Having carried out CPA, we can say that although these examples make sense and are *possible*, they are not particularly normal or *typical*. Given that all four involve killing, and not merely damage, we might consider them instances of Pattern 4. However, Pattern 4 refers to the killing of (unwanted) animals by humans, usually in an official, procedural context. Hence, (3) and (4) must instead belong to Pattern 1; they certainly do not fit with Patterns 2 and 3. The same goes for (5) and (6); although they involve killing, they are not describing the sort of event typically construed by Pattern 4. Their objects include animals and fetuses, but they also involve inanimate objects. Verb senses do not change mid-argument, unless in a creative exploitation such as wordplay [6, p. 72]; therefore, the same sense of the verb must apply to all entities in this co-hyponymy. If ‘water’ and ‘eggs’ cannot be killed, then we know that these examples are referring instead to destructive *damage*, i.e. belong to Pattern 1. Using CPA, it is therefore possible to say that while animals are sometimes *destroyed* in the same sense as inanimate objects such as houses and cars, humans are not. This is an assertion now provable with evidence.

In terms of my research, this finding contributes to answering the research question, “In what ways are animals conceptualised as persons, and in what ways are they conceptualised as things?”, which can only be answered in full once evidence has been gathered from a wide range of ‘killing’ verbs and their patterns. Nevertheless, this example is one small step towards demonstrating the inherent anthropocentrism of (English) language, and the persistent, widespread and insidious habit of likening nonhuman beings to inanimate things. In light of literature found in human-animal studies and critical animal studies, this might feasibly be interpreted as an attempt by the speaker to justify the routine exploitation of animals by humans, or at the very least as a betrayal of their view – subconscious or otherwise – that some animals are more similar to insentient objects than they are to humans, and as such deserve their subordinate status.

4 Conclusions

The very brief example given in Section 3.2 is not a full analysis, owing to space limitations, but it points towards a route along which CPA might be used as an empirical basis for (critical) discourse analysis. The systematicity of the CPA procedure ensures that conclusions drawn from the data are reached methodically and with measurable evidence. I might assert, for instance, that humans objectify other animals with their language, and I might – as has been done by many in human-animal studies – be able to provide several examples of this type of oppressive language. However, such assertions are made less convincing by not having accounted for the whole picture (the ‘whole picture’ being, admittedly, a sample of the whole picture). By starting with a word and using CPA to map the meanings of that word across large samples of text, we can demonstrate with more accuracy where, when and in what ways its norms are being used and exploited, often subtly and even unknowingly, to further a particular ideology.

There are some challenges to using this approach. It currently involves a lot of manual tagging, and at its most effective CPA requires a corpus-specific ontology of types, which takes time and dedication. However, once the groundwork has been laid, the discourse analyst has an empirically well-founded and robust basis from which to explore meaning. The classification of concordance lines in terms of their arguments rather than surface-level representations has advantages in terms of generalisability, and the use of CPA does not preclude – rather bolsters – other forms of textual analysis.

References

1. Atkins, S.: Tools for Computer-aided Corpus Lexicography: the Hector Project. *Acta Linguistica Hungarica*, 41, 5-72 (1993).
2. Baker, P.: *Using Corpora in Discourse Analysis*. Continuum, London (2006).
3. Gee, J.: *Social Linguistics and Literacies: Ideology in Discourses*. 5th edn. Routledge, London (2015).
4. Goldberg, A. E.: *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago (1995).
5. Hanks, P.: *Corpus Pattern Analysis*. In: *Proceedings of the 11th EURALEX International Congress*, pp. 87-97, Lorient, France (2004).
6. Hanks, P.: *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, MA (2013).
7. Pustejovsky, J.: The Generative Lexicon. *Computational Linguistics* 17(4), 409-441 (1991).
8. Sealey, A., Pak, C.: First catch your corpus: methodological challenges in constructing a thematic corpus. *Corpora* (forthcoming).
9. Sinclair, J.: *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Collins, London (1987).
10. Sinclair, J.: *Corpus, Concordance, Collocation*. Oxford University Press, Oxford (1991).
11. van Dijk, T.: Multidisciplinary CDA: a plea for diversity. In: Wodak, R., Meyer, M. (eds.) *Methods of Critical Discourse Analysis*. Sage, London (2001).
12. Wodak, R.: *Language, Power and Ideology: Studies in Political Discourse*. Critical Theory, Vol. 7. John Benjamins, Amsterdam (1989).

Digital Storytelling and the 21st Century Classroom: a powerful tool in phraseological units learning

Annalisa Raffone¹

¹ “L’Orientale” University, Naples NA 80121, Italy
araffone@unior.it

Abstract. 21st Century society is characterised by an abundance of information due to the growing availability of emerging technologies.

This paper explores the impact that Digital Storytelling, regarded as a new multimedia learning tool, has in enhancing critical thinking and learning motivation in teaching a second language and culture.

The first part of the paper presents a theoretical framework for understanding the power of this instructional tool and how a technology-integrated learning environment could have a positive effect on student learning.

The second part of the paper shows the results of a study whose aim was to focus on the analysis of MWEs from a DST application for secondary school students, in order to demonstrate how DST could be an effective multimedia tool in teaching phraseological units in second language acquisition.

Keywords: Digital Storytelling, Technology-Integrated learning environment, Phraseological units.

1. Introduction

Current researches on effective educational practices[1] have shown that Digital Storytelling (DST) is a powerful instructional tool for both students and educators.

21st Century students are the first generation to use computers, smartphones, digital music players and other tools of the digital age which, along with e-mails and instant messaging, are integral parts of their life. Prensky defines them as *digital natives* because students today are all “native speakers’ of the digital language of computers, video games and the Internet”[15].

This research project aims to investigate the developments and the effects that the use of DST provides in the classroom in teaching and learning foreign languages and cultures.

At its core, DST combines the art of telling stories with a variety of digital multimedia, including computer-based graphics, recorded audio, text, video clips, and music, so that it can be played on a computer or uploaded on a website. In this way, DST allows computer users to become creative storytellers, developing interesting stories that are typically just a few minutes long (three-to-five min.), have a variety of uses

(from personal tales to historical events or teaching materials), revolve around a chosen theme and often contain a particular point of view.

According to Robin[16], integrating visual images with written text both enhances and accelerates student comprehension, and DST constitutes a meaningful approach for energizing instructors and motivating students. Thanks to the advanced, available, low-cost and user-friendly today's multimedia, students are able to no longer passively listen to lectures (as with PowerPoint slides or paper-based textbooks) but they are actively engaged in the learning process, experiencing feelings of ownership and taking responsibility for their learning.

As Ohler[14] suggests, the greatest potential of DST lies in the fact that it provides digital natives the opportunity to speak in their own language, because media is the environment in which they feel comfortable.

However, along with the development of content understanding, students also develop planning skills in a useful and meaningful manner: when creating their own stories, students are asked to research a topic, look for pictures, record their voice, choose a particular point of view, which force them to create storyboards, story maps, scripts and other planning materials. In this way, they find themselves involved in what Ohler calls "the media-production process"[14], a process that consists in creating, editing, and sharing original work.

Moreover, using their own viewpoints gives students a sense of ownership because the stories they tell are full of their personal feelings and ideas, that are also expressed in a personal manner, so it is clear that DST can help them to capture and expand their imagination and develop their communication skills.

In addition, when digital stories are shared on the Web, students have the opportunity to view the work of others, so that they learn cultural differences, expand their own knowledge and give value to their experience.

Growing developments in neuroscience and neuropsychology[5] are proving that stories have a biological basis[4]: sensory experience is mostly forgotten, moments after we hear, see, touch and taste it. However, researchers have demonstrated that to remember, we move sensory information from a slower part of our brain to the fast-working hippocampus[13], which participates a complex process of consolidation of our memory. This process seems to require reviewing the experience again and again in our minds. The Romans would say *repetita iuvant*.

Second language learning requires the management of four main skills: speaking, writing, listening and reading, which lead to effective communication.

Lexical acquisition is one of the most important parts of language learning and also one of the most discussed topics in second language pedagogy.

However, the way through which learners acquire lexicon and which is the best way for it to be taught is still a discussed matter.

In order to increase their language skills and be able to speak and write correctly in L2, students need to acquire lexicon through repeated exposure.

Laufer defines lexical acquisition as a 'cumulative process': "Each additional exposure to the same word may enrich and strengthen the learner's knowledge of it. The question to ask here is: how many exposures to a word are needed before the learner can recall or recognize the meaning of a word?"[10].

As the traditional formal teaching provides a monotonous and boring environment, using DST in second language teaching seems to be an interesting educational way suitable for enhancing the efficiency of learning and teaching.

In this study, the data collected refers to the English language acquisition by secondary school students in order to enrich their own lexicon and make them aware of the differences in the usage of the most frequent words.

Teaching lexicon by using multimedia tools seems to be the best way to encourage students learning a second language.

It's for this reason that this research project also aims to demonstrate that DST provides teachers a unique way to help students retain new information and comprehend difficult topics without taking a long time.

The idea of this project is that DST can effectively represent a bridge between existing knowledge and new material.

1.1 The Study

The purpose of this study was to focus on the analysis of MWEs from a DST application for secondary school students, in order to demonstrate how DST could be an effective multimedia tool in teaching phraseological units in second language acquisition.

The data in this study is based on a digital storytelling language course called *Muzzy* developed by BBC.

The choice to analyse a single digital story lies in the fact that this project represents the first research conducted on the use of phraseology in the field of DST.

Setting. *Muzzy* is a language-learning program by which educators can teach secondary school students a second language through animated and funny stories.

Muzzy was chosen among different digital stories because it is one of the most used language teaching course in Italy for students between the ages of 11 and 14 years old.

In *Muzzy* each episode is designed to build on the previous episode through words and concepts students have just learned. The use of repetitions makes learning easy and funny.

The story has been analysed in the English language. It is divided into six parts and consists of 67 scenes both printed in a book and included in a DVD.

The corpus consists of 4622 tokens and it was analysed entirely. Each episode explains different particular linguistic phenomena related to level A2, according to CEFR reference levels.

Data collection and data analysis. The data was collected through the following procedures:

- The 67 scenes of the story have been organized into 6 parts corresponding to a corpus consisting of 6 different episodes.

- Each sentence of the entire story has been analysed to identify the use of MWEs in order to put forward hypothesis on how phraseology is employed in DST. Moreover, a list of the total number of verbs in the digital story has been compiled to specifically identify verbal MWEs by using PARSEME annotation guidelines.

1.2 Discussion and Conclusions

The results of this study show a significant usage of MWEs that includes both VMWEs and other types of MWEs.

The MWEs in this digital story were identified by following PARSEME criteria.

In order to identify VMWEs, the first step was to compile a list of the total number of verbs used in this digital story; the second step was to analyse each of them in order to categorize the VMWEs according to their peculiarities.

The total number of verbs used in this digital story is 73.

First of all, a great number of *Light-Verb constructions* was recorded, in particular in scenes 29 to 39 and 40 to 45: *have a shower, take a rest, have breakfast/lunch/dinner* are only some of the most used ones.

Also, numerous *idioms* can be pointed out, such as *I beg your pardon, be careful, well done, come on* – which are endlessly repeated throughout the entire story.

As for *phrasal verbs*, although only 18 of the total number of verbs were used as phrasal verbs, they were repeated 38 times. Among them, the most recurring ones are: *point at, ask for, come from, come back, go up and go in*.

The repetition of one phrasal verb in the same phrasal construction is about 2,1%.

The trickiest aspect of PV is that they can have multiple meanings and that these meanings can change depending on the particle.

As *Muzzy* is a digital story designed for secondary school students, it's mostly for this reason that the total number of phrasal verbs is low. Nevertheless, two episodes of this digital story (in particular, scenes 46 to 58 and 59 to 67) are mainly dedicated to PV, that are constantly repeated in the same phrasal constructions to make their assimilation and comprehension easier and clear. This allows VMWEs to be learned by implicit learning by unconsciously meeting multi-word sequences repeatedly in funny contexts.

Although different *collocations* can be identified (among the most repeated ones: *dirty job, roller-skate, lunch-time, dinner-time, sitting-room*), an overall analysis allows to affirm that the number of VMWEs is greater than MWEs containing only nouns and adjectives. In fact, they're in a ratio of 1:12.

As Lam[8] maintains, trying to remember a list of individual uses of a part of speech is hardly helpful and positive for learners as they mechanically repeat like they are singing a lullaby, without really understanding the meaning of a term and why it is used in a precise context.

Instead, the incessant recall of verbs, nouns, adjectives, prepositions mixed with a clear and pleasant animated story, music and sound effects make students learn lexicon and its functions easily.

Repetition aids familiarity: students get the hang of terms and expressions and start using them naturally.

Indeed, both the meaning and the amusing situations of the digital story also make lexicon easy for students to remember and this can stimulate their cognitive development encouraging them to practice L2 and persuading themselves that they are able to work out the meaning of the words they do not know.

As Kennedy affirms, “research in cognitive science has shown that frequency of occurrence and frequency of experience establishes words and collocations as units of learning, and becomes a determinant in their use”[7].

This paper is based on the idea that DST talks to students, gives voice to their experiences and is able to simplify even difficult topics, constructing a learning environment in which they can learn in an effective, creative and meaningful way. So, this study only represents a first analysis conducted on the use of phraseology in DST and needs to be enriched by a larger corpus whose creation will be the starting point to test the effective incidence of MWEs in the field of DST.

References

1. Alismail, H.A.: Integrate Digital Storytelling in Education. In: Journal of Education and Practice, vol.6, No.9 (2015).
2. Atton, C.: Alternative Media. SAGE Publications Ltd, London (2001).
3. Bruner, J.: Actual Minds, Possible Worlds. Harvard University Press, Cambridge, Massachusetts and London (1986).
4. Calabrese, S.: Neuronarratologia. Il futuro dell’analisi del racconto. Archetipolibri, Bologna (2009).
5. Calabrese, S., Ballerio, S.: Linguaggio, letteratura e scienze neuro-cognitive. Ledizioni, Milano (2014).
6. Hulstijn, J. H.: Mnemonic methods in foreign language vocabulary learning: theoretical considerations and pedagogical implications. In: Coady, J., Huckin, T. (eds): Second language vocabulary acquisition: a rationale for pedagogy, pp. 203-224. Cambridge University Press, Cambridge (1997).
7. Kennedy, G.: Phraseology and language pedagogy. In: Meunier, F., Granger, S.: Phraseology in Foreign Language Learning and Teaching. John Benjamins Publishing Company, Amsterdam/Philadelphia (2008).
8. Lam, Y.: Applying cognitive linguistics to teaching the Spanish prepositions *por* and *para*. Language awareness, 18 (1), 2-18 (2009).
9. Lambert, J.: Digital Storytelling: Capturing Lives, Creating Communities. Routledge, New York and London (2013).
10. Laufer, B.: Focus on Form in Second Language Vocabulary Learning. John Benjamins Publishing Company, University of Haifa, p.227 (2005).
11. Lorincz, K., Gordon, R.: Difficulties in Learning Prepositions and Possible Solutions. In: Linguistic Portfolios, vol. 1, Art. 14 (2012).
12. Meunier, F., Granger, S.: Phraseology in Foreign Language Learning and Teaching. John Benjamins Publishing Company, Amsterdam/Philadelphia (2008).
13. Milner, B., Squire, L.R., Kandel, E.R.: Cognitive Neuroscience and the Study of Memory. Neuron, Vol.20, 445-468, 1998.
14. Ohler, J.: Digital Storytelling in the Classroom. New Media Pathways to Literacy, Learning and, Creativity. Corwin Press, California, p.11 (2013).

15. Prensky, M.: Digital Natives, Digital Immigrants. In: On the Horizon, vol. 9, No.05, p.1 (2011).
16. Robin, B.R.: Digital Storytelling: A Powerful Technology Tool for the 21st Century Classroom. Routledge, The Ohio State University (2008).
17. Yuksel-Arslan, P., Yildirim, S., Robin, B.R.: A phenomenological study: teachers' experiences of using digital storytelling in early childhood education. Educational Studies 42(5), 427-445 (2016).

Author Index

- Antunes, Sandra, 137
Ayupova, Roza, 169
- Baena Lupiáñez, María del Carmen, 19
Ballier, Nicolas, 113
Bel Enguix, Gemma, 182
Blagus Bartolec, Goranka, 132
- Carvalho Ribeiro, Sarah Virginia, 122
Chang, Jason, 95
Chang, Jim, 95
Chen, Jhih-Jie, 95
Chen, Mei Hua, 95
Choe, Jae-Woong, 178
Chow, Ka Po, 162
Condamines, Anne, 104
Costa Lima, Paula Lenz, 122
- Elmerot, Irene, 174
- Foufi, Vasiliki, 36
Franklin, Emma, 190
- Hashio, Shimpei, 70
- Kircili, Kathrin, 127
- Larsen-Walker, Melissa, 78
Llongo, Victoria, 87
- Martínez, Jordi, 182
Matas Ivanković, Ivana, 143
- Nefedova, lyubov, 154
Nerima, Luka, 36
Nigam, Amber, 28
- Pamies-Bertrán, Antonio, 60
Pchelina, Marina, 178
Poirier, Éric, 1
- Raffone, Annalisa, 197
Ramos Ruiz, Ismael, 60
Rogers, James, 148
Ruiz Yepes, Guadalupe, 11
- Steyer, Kathrin, 45
- Torres Flores, Liliana, 182
Tsou, Benjamin K., 162
- Warnier, Maxime, 104
Wehrli, Eric, 36
Wong, Derek F., 162
- Yamauchi, Nobuyuki, 70
Yang, Ching-Yu, 95
Yusupova, Seda, 53
- Zimina, Maria, 113