Towards a Cross-linguistic Study of Phraseology across Specialized Genres

Ana Roldán-Riejos [1][0000-0002-9635-2814] and Łukasz Grabowski[2][0000-0002-3968-9218]

[1] Universidad Politecnica de Madrid, Spain
ana.roldan.riejos@upm.es
[2] University of Opole, Poland; University of Ostrava, Czechia
lukasz@uni.opole.pl; Lukasz.Grabowski@osu.cz

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

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

1 Introduction

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

EUROPHRAS 2019, pages 140–144,
Malaga, Spain, September 25-27, 2019.
https://doi.org/10.26615/978-2-9701095-6-3_019

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

1.1 Main aims

The aims of the project are threefold:

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

2 Methodology

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

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

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

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

3 Preliminary results

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

Table 1**.** Criteria adopted for the analysis across languages

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

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

Table 2. Features considered in the cross-genre analysis

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

Characteristic discourse features	Downtoners, Intensifiers Approximators Personal pronouns. Passive voice	Downtoners, Intensifiers Approximators Personal pronouns	Downtoners, Intensifiers Approximators Passive voice Modal Verbs
Visuals	Diagrams, tables, charts	Photographs, drawings	Tables, diagrams, photographs, drawings

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

4 Conclusions and future work

(a) This work has considerable potential to contribute to cross-linguistic research on recurrent phraseologies in specialized text types and genres, and its results may, among others, help improve LSP vocabulary learning and fluency (cf. Boers & Lindstromberg 2008; Roldán-Riejos & Úbeda 2018a; 2018b), and provide potentially useful data for translators, terminologists, lexicographers and researchers from various discourse communities.
(b) The future study will offer an opportunity to test and fine-tune methodologies of identification and analysis of collocations and longer multi-word items when applied to data written in typologically different languages. It is essential since many approaches (e.g. lexical bundles (Biber et al. 1999), Pattern Grammar (Hunston & Francis 2000), phrase frames (Fletcher 2002)) have been developed and applied largely to English-language material. For example, it might be interesting to see how to apply the distributional criteria (frequency, distribution range, coverage, various collocational strength metrics) to identification of recurrent word combinations in texts written in English, Spanish, Polish and Russian. This will also provide an opportunity for reflection on how to measure the amount of formulaic language in texts written in typologically different languages.

References

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

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