Phraseological complexity measures in learner Italian. Integrating eye tracking, computational and learner corpus methods to develop second language pedagogical resources.

PHRAME
Phraseological complexity
measures in learner Italian

CLOSING CONFERENCE
27-28 November 2023

Book of Abstracts

Table of Contents

Scientific and organising committee

Giulio Biondi (Università degli Studi di Perugia)
Veronica D'Alesio (Sapienza Università di Roma)
Irene Fioravanti (Università per Stranieri di Perugia)
Luciana Forti (Università per Stranieri di Perugia)
Valentina Franzoni (Università degli Studi di Perugia)
Sabine Koesters Gensini (Sapienza Università di Roma)
Francesca La Russa (Sapienza Università di Roma)
Francesca Malagnini (Università per Stranieri di Perugia)
Alfredo Milani (Università degli Studi di Perugia)
Maria Roccaforte (Sapienza Università di Roma)
Valentino Santucci (Università per Stranieri di Perugia)
Stefania Spina (Università per Stranieri di Perugia)
Fabio Zanda (Università per Stranieri di Perugia)

Keynote presentations

Learner Corpora and Phraseology, Philip Durrant

University of Exeter

In this keynote talk I will discuss corpus research as a methodology for studying phraseology in learner language. I will focus in particular on two methodological issues that face researchers in this area: the issue of meaningful quantification and the issue of valid interpretation.

The issue of meaningful quantification arises when researchers use numerical data to summarise the language found in a corpus. Quantification is a process of *constructive simplification*, whereby the detailed information that appears in individual texts is summarised in numerical form to help the researcher notice broader patterns of language use. While this is a powerful strategy, it is also important to pay attention to what information gets lost and how this affects the claims we can make.

In the first part of the talk, I will look at a method of quantification that has become particularly popular in collocation research: that of assigning texts a *phraseological sophistication* value, based on association measures. Again, taking the example of child writing in England, I will look in detail at the types of simplification this approach involves, at the information it leaves out, and at the implications for what conclusions we can draw.

The issue of valid interpretation arises because learner language is the complex product of many interacting factors. These include, amongst other things, learners' language abilities, their cognitive skills, communicative goals and task motivation; the nature of the topic under discussion; register conventions (and learners' perceptions of those conventions); and the context in which a text is produced (e.g., the availability of reference materials and tools such as a word processor or grammar checker, the time available). A key strength of corpus research is that it can capture authentic textual products at the confluence of all of these influences. However, this leaves researchers with the problem of drawing valid inferences about specific constructs of interest – of deciding, for example, what patterns indicate about learners' language knowledge, their motivation, their register awareness, etc.

In the second part of this talk, I will discuss this issue in relation to phraseology. Taking the example of a corpus elicited from child writers in schools in England, I will focus on the influence of individual writer preferences and the pedagogical context on the use of collocation use and reflect on the implications of this for learner corpus studies of phraseology more broadly.

Bringing together the discussion of the issues of meaningful quantification and valid interpretation, I will argue that they highlight the importance of taking an exploratory approach to learner corpus phraseology research that is open to the full range of influences on learner texts, that pays close attention to specific contexts and usages, and that constantly interrogates its categories and measures.

Defining, measuring and interpreting complexity, Gabriele Pallotti

Università di Modena e Reggio Emilia

The term “complexity” has gained considerable currency over the past decades, and has taken on a wide range of meanings. In this presentation, I will argue for a more restricted interpretation, focusing exclusively on formal, structural properties of linguistic units. Based on such a theoretical definition, I will critically review measures operationalising this construct, discussing their strengths, weaknesses and potential applicability to second language research, in order to establish a small set of indicators to be used routinely in the interest of replicability and knowledge accumulation in the field. In addition, I will discuss the relationship between complexity and difficulty and the associated notions of proficiency and development. More particulary, I will be concerned with how complexity should be interpreted in the domain of language acquisition, questioning a simplistic view like ‘the more, the better’.

Phrasal processing: Past, present, and future, Anna Siyanova-Chanturia

Te Herenga Waka – Victoria University of Wellington

Recent years have seen a growing interest in the processing mechanisms that underlie the on-line processing of multi-word expressions (MWEs). MWEs encompass a large set of sequences above the word level, such as collocations, binomials, lexical bundles, idioms, and so on. In this talk, I review recent psycholinguistic evidence attesting to faster processing and easier semantic integration of different types of phrasal configurations compared to novel sequences in various populations – L1 and L2 adults, L1 children, as well as individuals with developmental dyslexia. Current trends and directions for future research are discussed.

Paper presentations

Exploring phrasal verbs of action and motion in learner and native speaker narrative writing, Katherine Ackerley and Erik Castello

University of Padua

Good knowledge and effective use of phrasal verbs (PVs) are considered a marker of English proficiency (Garnier & Schmitt, 2016), and can add phraseological richness to the language (Gardner & Davies, 2007). Yet the array of PVs represents a major challenge to learners (Chen, 2013). This study investigates use of PVs by Italian learners of English in narrative writing, identifying the variety of lexical verbs and adverbial particles used at different proficiency levels, and then comparing their use with that of native speakers (NS). Texts produced by 201 Italian students and 73 L1 speakers of English for the COREFL corpus (Lozano et al., 2021) were analysed. The prompt was a video clip from Charlie Chaplin's *The Kid*. The learner corpus was split into four subcorpora, according to the results of a proficiency test: A2 (13 texts), B1 (56), B2 (79) and C1 (53 texts).

The learner and NS corpora were POS-tagged using CLAWS, then AntConc was used to retrieve all instances of adverbial particles preceded continuously or discontinuously by a lexical verb. These combinations were then checked manually and looked up in a phrasal verbs dictionary if necessary. An initial analysis of the data revealed a consistent increase in both the frequency and variety of PVs at each CEFR level (frequency = 13.78 pkw at B1; 18.48 at B2; 25.75 at C1; 32.73 in the NS corpus; PV variety = 47 at B1 level; 66 at B2; 86 at C1; 123 by NS).

Given the extensive variety of PVs identified in the two corpora, we decided to focus mainly on those used to express actions and motion in space (Ibarretxe-Antuñano, 2017).

Our research questions are:

- Which lexical verbs and particles combine to convey action and motion in the two corpora?
- In terms of use of these phrasal verbs, how does learner and NS phraseology differ?
- What nuances of meanings are expressed through use of the adverbial particles?

Preliminary findings suggest that PVs appear to be skilfully exploited by NS and advanced learners to concisely express elaborate and nuanced meanings (e.g. run off, run away, pass off, pick back up), while the phraseology of lower level learners is less successful.

We will conclude by discussing how the results can inform approaches to teaching PVs of action and motion.

References

Ibarretxe-Antuñano, I. (2017) (Ed.), *Motion and Space across Languages: Theory and applications*. Amsterdam/Philadelphia: John Benjamins Publishing.

Chen, M. (2013). Overuse or underuse. *International Journal of Corpus Linguistics*, *18*(3), 418–442.

Gardner, D., & Davies, M. (2007). Pointing out Frequent Phrasal Verbs: A Corpus-Based Analysis. In *Quarterly 41*(2), 339–359.

Garnier, M., & Schmitt, N. (2016). Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System*, *59*, 29–44.

Lozano, C., Díaz-Negrillo, A., & Callies, M. (2021). Designing and compiling a learner corpus of written and spoken narratives: COREFL. In C. Bongartz & J. Torregrossa (Eds.), *What's in a Narrative? Variation in Story-Telling at the Interface between Language and Literacy* (21–46). Bern: Peter Lang.

Reading versus listening to authentic materials in L2 Italian: What leads to greater vocabulary learning gains? Mahnaz Aliyar and Anna Siyanova-Chanturia

Te Herenga Waka – Victoria University of Wellington

While the value of reading for L2 incidental vocabulary acquisition is widely acknowledged, the role of listening remains relatively less explored (Feng & Webb, 2020; Nation, 2013). This holds particularly true regarding audiobooks. Despite their growing popularity and increasing availability, authentic audiobooks have not been subject of any incidental vocabulary acquisition studies in the field of second language acquisition. Additionally, while several studies have investigated incidental learning of multi-word expressions (MWEs) from various modes of input (Pellicer-Sánchez, 2017), further research is needed to determine whether different modes of input contribute to the acquisition of MWEs as compared to single words. Given aural input plays a critical role in L2 vocabulary development and use (Webb & Nation, 2017), there is a driving need for research on how listening, compared to reading, contributes to incidental learning of single words and MWEs (Feng & Webb, 2020). Using an authentic Italian novel and the audiobook of the same novel, this study aims to fill these gaps by examining the effects of reading versus listening on incidental learning of 22 single words and 19 MWEs.

Ninety-five Iranian university students of L2 Italian (advanced proficiency level) were randomly assigned to one of the following groups: 1. reading half of the Italian novel "*L'amica Geniale: Infanzia, Adolescenza*" (Ferrante, 2011) for pleasure; 2. listening to the audiobook of the same novel for pleasure; 3. control group who engaged in a different L2 learning activity unrelated to the experiment. The experimental procedure consisted of a four-week reading or listening treatment, preceded by a pretest, and followed by an immediate and a three-week-later posttest. The results were analysed using mixed-effects modelling in R. The results showed that L2 incidental vocabulary learning occurred through both reading and listening, and the gains were retained in both modes of input three weeks after engagement with the input. Moreover, listening yielded greater amounts of vocabulary gain. The findings of the study have important pedagogical implications for the effectiveness of authentic audiobooks in incidental vocabulary learning and retention.

References

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition*, *42*(3), 499–523.

Ferrante, E. (2011). *L'amica geniale: infanzia, adolescenza* (Vol.1). Roma: Edizioni e/o.

Nation, I.S.P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Pellicer-Sánchez, A. (2017). Learning L2 collocations incidentally from reading. *Language Teaching Research, 21*, 381–402.

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Exploring Phraseological Patterns in English Non-Finite Clauses: A Corpus-Based Study of Business English, Olfa Ben Amor

University of Monastir

The availability of large-scale corpora and the sophisticated data-processing tools have further shed light on the analysis of phraseology. This study investigates the phraseology pertinent to the English non-finite clauses, namely to-infinitive, -ing and past participle clauses headed by four headword targets: adjectives, adverbs, nouns and pronouns, and analyses the semantic and discourse functions encoded in these structures. To this effect, a specialised corpus is compiled from business English texts exhibiting two distinct genres (academic vs. news). The academic genre includes research articles gleaned from four academic journals (*The Journal of Financial Economics*, *The Journal of Monetary Economics*, *The Quarterly Journal of Economics* and *The Journal of International Management)*, and Tunisian students' graduate MA dissertations (15 MAs) and PhD theses (7PhDs) collected from five different institutions in Tunisia in the discipline of business. The news genre includes business news texts culled from *The Economist* (12 texts) and *Financial Times* (8 texts). The data collection took place between 2018 and 2020 and the total number of words in the corpus amounts to 1,763,930 words. The corpus is analysed using *NooJ* software package (http://nooj4nlp.org/) (Silberztein, 2020). Two levels of analysis are adopted: *(i)* the level of investigating the different lexico-grammatical items that co-pattern with one another to formulate the different phraseologies of the non-finite clauses, and *(ii)* the level of analysing these patterns into semantic sets and discourse functions. The analysis has revealed interesting differences and similarities between the three sub-corpora in the frequency and degree of fixedness of non-finite phraseology (ranging from semi-restricted, restricted and idioms). The results indicate that Tunisian novice academic writers may need further instructional support in how to use these phraseologies more effectively in their academic writing, namely semi-restricted phraseology which tend to be more prone to errors. Other generic differences stood out in the use of semantic patterns and discourse functions. These differences highlight the need to adopt a phraseologically-oriented approach to language teaching in EAP and ESP context.

References

Silberztein, M. (2020). NooJ V7.0 [software]. *Formalising Natural Languages with NooJ2020*.

Phraseological Patterns in Learner Academic English: Insights from Corpus-Driven Approaches, Sibel Aybek and Cem Can

Cukurova University

"The language we use every day is composed of prefabricated expressions, rather than being strictly compositional" (Gray & Biber, 2015, p. 125). These expressions constitute the phraseology of a language, which is a characteristic feature of language owing to "the tendency of words to occur, not randomly, or even in accordance with grammatical rules only, but in preferred sequences" (Hunston, 2002, p. 137; Groom, 2005). Phraseological competence is a critical component of second language acquisition (Sinclair, 1991) and phraseological units specific to the academic language are significant indicator of language development by reducing the processing effort (Nesselhauf, 2005) for academic English learners. Adopting a corpus-driven approach, this study employs both quantitative and qualitative methods to analyse phraseological patterns. The research design involves a threefold comparison: between native speakers of English and advanced EFL learners, and between novice (L1 and L2) and expert academic writers. Frequency lists, keywords and lexical bundles are used to explore the phraseological patterns across three corpora, namely Turkish International Corpus of Learner English (TICLE), Louvain Corpus of Native English Essays (LOCNESS) and British Academic Written English Corpus (BAWE). Initial quantitative analysis involves generating frequency lists, identifying keywords, and extracting lexical bundles. This is followed by a qualitative contrastive analysis to delve deeper into the nuances of phraseological patterns, drawing insights from Granger's (1998) principles of overuse and underuse in learner corpora. The results indicate that using both quantitative and qualitative contrastive analysis to explore learner academic English can give valuable insights to the broader understanding of phraseological competence in second language learners. We discuss the results in light of the under- and overuse of multi-word expressions in English L2 novice academic texts, informing the contrastive study of L2 phraseology, and practical implications for English language teaching.

References

Gray, B., & Biber, D. (2015). Phraseology. In D. Biber, D.; & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics,* Cambridge University Press, 125–145.

Groom, N. (2005) Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes 4*(3), 257–277.

Hunston, S. (2002). *Corpora in Applied Linguistics.* Cambridge University Press.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus.* John Benjamins.

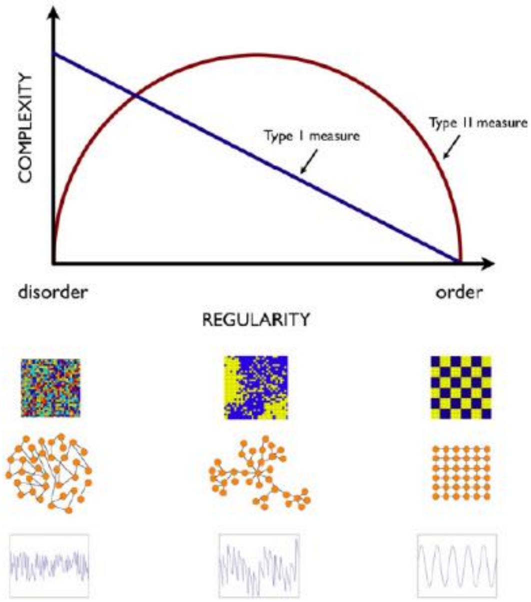
Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Between order and disorder: an ecological view on lexical diversity measures, Arianna Bienati and Paolo Brasolin

Eurac Research, Bolzano

Starting from the influential study by Paquot (2019), many scholars have conceptualised phraseological complexity as composed by two intertwined components: phraseological diversity and sophistication (e.g., Vandeweerd et al., 2022). To compute phraseological diversity, lexical diversity measures such as root type-token ratio have been used. Although there have been significant contributions to the debate on the validity of such indices (Kyle et al., 2021; McCarthy & Jarvis, 2010), a thorough comparison with theoretical observations from other fields of research would greatly enhance our understanding of the specific type of complexity that these measures truly capture.

Typological literature has distinguished two types of complexity (Dahl, 2009): Kolmogorov complexity refers to the length of the shortest description needed to represent a string of symbols. Gell-Mann effective complexity is understood as the length of description required to specify the set of regularities present in a string. According to Kolmogorov complexity, a string that lacks any regular patterns or structure would require a very long thus complex description, as each element of the string would need to be individually accounted for. From the perspective of Gell-Mann effective complexity, instead, a string without any regularities would have no complexity, since it contains no structured patterns.



In ecology (Parrott, 2010), a clear-cut distinction is made between measures that operationalise the Kolmogorov complexity (“Type 1 measures”) vs. the Gell-Mann effective complexity (“Type 2 measures”). Type 1 measures favor random sequences and can be visualised as linear functions on the continuum between order and disorder; type 2 measures, instead, can be visualised as a convex function that reaches its peak when a system strikes a balance between rules and exceptions (Fig. 1). Intuitively, the latter might be the kind of complexity we want to be able to measure in a text.

This contribution will therefore try to answer the following research question: which type of complexity (Kolmogorov vs. Gell-Mann) is measured by the most commonly used lexical diversity indices?

In addressing this question, we will conduct a simulation to evaluate lexical diversity in Italian texts employing TTR-based metrics along with entropy-based ones (as found in Garner, 2020), including Type 2 measures, such as *fluctuation complexity*, as described in Parrot (2010). Each measure will be calculated varying the text length by truncation to study the length dependency of its behavior. Additionally, each measure will be calculated on texts which have different expected behaviors in terms of Kolmogorov and Gell-Mann complexity: real texts –

expert-authored (e.g., Repubblica corpus) and non-expert-authored (e.g., LEONIDE, ITACA)[1] – are expected to display a certain degree of regularity, thus being representative of a medium to high Gell-Mann complexity; synthetic texts sampling the real texts using either the original word distributions or a uniform distribution (i.e., every word has an equal probability of appearing), instead, do not show this regularity, thus representing a medium to high Kolmogorov complexity, but a low Gell-Mann complexity. We will then be able to distinguish between type 1 and type 2 measures by identifying the set of texts (real or synthetic) for which the various lexical diversity measures yield the highest values. Results will contribute to the assessment of metrics' validity for different conceptualisations of complexity.

References

Dahl, Ö. (2009). Testing the assumption of complexity invariance: The case of Elfdalian and Swedish. In G. Sampson, D. Gil, & P. Trudgill (Eds.), *Language Complexity As an Evolving Variable* (pp. 50–63). Oxford University Press.

Garner, J. (2020). The cross-sectional development of verb–noun collocations as constructions in L2 writing. *International Review of Applied Linguistics in Language Teaching*, *60*, 909–935.

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. *Language Assessment Quarterly*, *18*(2), 154–170.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*, 121–145.

Parrott, L. (2010). Measuring ecological complexity. *Ecological Indicators*, *10*(6), 1069–1076.

Vandeweerd, N., Housen, A., & Paquot, M. (2022). Comparing the longitudinal development of phraseological complexity across oral and written tasks. *Studies in Second Language Acquisition*, 1–25.

[1] Size and scale of cited corpora:

Corpus	Size (in n. of tokens)	N. of documents
Repubblica corpus (from: http://sslmit.unibo.it/repubblica)	380,823,725	572,515
LEONIDE_IT (from: http://hdl.handle.net/20.500.12124/25)	93,000	844
ITACA (not yet published)	382,964	635

The 'growth' of academic phrases: a contrastive corpus-based study on phraseological complexity in Romanian learner English, Madalina Chitez, Andreea Dinca, Ana-Maria Bucur, and Kristian Miok

West University of Timisoara and University of Bucharest

Previous studies have established a relationship between phraseological complexity and language proficiency in a particular L2 learner group (Vandeweerd et al., 2021). But how do phrases 'grow' from a certain linguistic level to the next in the case of Romanian learners of English? Our study sets out to gain an understanding of the phraseological complexity (Vandeweerd et al., 2022) of the academic writing produced by university-level Romanian L2 learners so that we can assess the gap between this profile and the expert writer phraseological profile. In our view, novice writers, in contrast to expert writers, are still in the process of mastering linguistic and rhetorical skills necessary for writing various pieces of academic genres. Expert-level L2 writing is produced by Romanian scholars who write in English and have published their work in high quality academic journals from their fields. For the analysis, we compare data from the ROGER corpus (Chitez et al., 2021), a bilingual Romanian-English novice academic writing corpus with data from the EXPRES corpus (Chitez et al., 2022), a bilingual Romanian-English expert academic writing corpus. Our methodology involves applying the standard phraseology complexity measures (e.g. Vandeweerd et al., 2022): diversity and sophistication of adjectival modifiers and direct objects. For the selection of phrases, we use Paquot's method of extracting and analysing syntactic co-occurrences in a corpus (Paquot et al., 2021). One objective of the study is to verify whether the Romanian writers' phrases are more complex at the expert level of writing compared with the novice level, especially considering the demonstrated reverse correlation between linguistic complexity for academic writing in certain disciplines and publication standards (Bucur et al., 2022). Building up on a previous study by Dinca and Chitez (2021), we also aim at identifying category types, such adjectival modifiers (Vandeweerd et al., 2022), which improve significantly once the writers are more proficient in writing. Since our corpora contain L1 writing samples as well, we expand the analysis towards L1 phraseology (texts in Romanian) so that we can identify patterns of phraseology interference between L1 (see also Muresan et al., 2022) and L2 at the novice versus the expert level. We argue that the present study will shed new light on the phraseology used by the Romanian academic writers, to be included in further corpus-driven contrastive studies, while also proving useful in expanding our understanding of mother tongue influence on English L2 production.

References

Bucur, A. M., Chitez, M., Muresan, V., Dincă, A., & Rogobete, R. (2022, June). EXPRES Corpus for A Field-specific Automated Exploratory Study of L2 English Expert Scientific Writing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference – LREC 2022,* 4739–4746.

Chitez, M., Bercuci, L., Dinca, A., Rogobete, R. and Csurös, K. (2021). *Corpus of Romanian Academic Genres (ROGER)*. West University of Timisoara. Available at https://roger-corpus.org/.

Chitez, M., Rogobete, R., Muresan, V. and Dinca, A. (2022). *Corpus of Expert Writing in Romanian and English (EXPRES)*. West University of Timisoara. Available at https://expres-corpus.org/.

Dinca, A., & Chitez, M. (2021). Assessing learners' academic phraseology in the digital age: a corpus-informed approach to ESP texts. *Journal of Teaching English for Specific and Academic Purposes 9*(1), 71–84.

Mureșan, V., Rogobete, R., Bucur, A.-M., Chitez, M. and Dincă, A. (2022). Phraseology in Romanian Academic Writing: Corpus Based Explorations into Field-Specific Multiword Units. In D. Anca, M. Chitez, L. Dinu and M. Dobre (Eds.), *Recent Advances in Digital Humanities. Romance Language Applications* (pp. 29–48). Bern: Peter Lang.

Paquot, M., Naets, H., & Gries, S. T. (2021). Using syntactic co-occurrences to trace phraseological complexity development in learner writing: Verb+ object structures in LONGDALE. In: B.L. Bruyn & M. Paquot (eds.), *Learner corpus research meets second language acquisition*, (pp. 122–147). Cambridge: Cambridge University Press.

Vandeweerd, N., Housen, A., & Paquot, M. (2021). Applying phraseological complexity measures to L2 French: A Partial Replication Study. *International Journal of Learner Corpus Research* 7(2), 197–229.

Vandeweerd, N., Housen, A., & Paquot, M. (2022). Comparing the longitudinal development of phraseological complexity across oral and written tasks. *Studies in Second Language Acquisition*. 1–25.

Discovering the potential of automated phraseological interference error detection: Transformer-based approach, Darya Kharlamova

Higher School of Economics, Moscow

Formulaic language is a crucial part of L2 acquisition. Present in both L1 and L2, formulaic expressions may help language learners in L2 comprehension and production (Conklin & Carrol, 2018). However, interference with L1 can cause errors in production as well (Weinreich, 1979). The present paper explores the possibilities of detecting L1 Russian interference errors in English learner texts with a fine-tuned Transformer-based neural network. The research focuses on the mistakes connected with formulaic expressions and phraseologisms (defined in accordance with the criteria in Gries (2008).

We accumulated a dataset of over 3600 erroneous sentences from the essays in the REALEC corpus (Vinogradova & Lyashevskaya, 2022), classified the mistakes into Synonyms, Copying expression, and Tense semantics following Weinreich (1979). For the Transformer training, we chose the SpaCy architecture (Montani et al., 2023) and RoBERTa-base (Liu et al., 2019).

We prepared two variants of the resulting neural network. In each case, the data was split into a training set and a test set with a 70/30 ratio. The first one (a) is a pipeline consisting of three separately trained Transformers, one for each of the tags. It detects the majority of the mistakes, but it also mixes categories and gives false-positive results. The second variant (b) is a single Transformer capable of detecting all types of mistakes. It extrapolates errors that were not present in the dataset. While it overlooks many of the errors, its predictions are mostly correct.

An example of the automated markup can be found in the table below (highlighted are the automatically detected mistakes).

Tag	(a)	(b)
Copying expression	*<...> can increase pices on harmful production<...>.*	*<...> shares of Samsung increasing from 2011 year <...>.*

The metrics for the resulting Transformers can be found in the table below. These metrics have been calculated using standard F-score, Precision and Recall formulae on 30% test set.

	(a)			(b)
	Copying expression	Synonyms	Tense semantics	
F-score	79.16	67.47	90.65	84.11
Precision	94.00	71.19	86.9	89.7
Recall	68.36	64.12	94.74	79.18

The main conclusion is that, given a sufficient dataset, Transformers can be rather effective in a task as sophisticated as detecting L1-motivated phraseological mistakes. Currently, cross-validation experiments have not been performed. However, we are planning to use 5-fold cross-validation on later stages of the project to analyse how the models generalise across different subsets of the data.

References

Conklin, K., & Carrol, G. (2018). Cognitive and psycholinguistic perspectives on formulaic language: Chapter Three. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (1st ed.). Routledge.

Gries, S. Th. (2008). 1. Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology* (pp. 3–25). John Benjamins Publishing Company.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python.*

Vinogradova, O., & Lyashevskaya, O. (2022). Review of Practices of Collecting and Annotating Texts in the Learner Corpus REALEC. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings* (Vol. 13502). Springer International Publishing.

Weinreich, U. (1979). *Languages in Contact: Findings and Problems*. De Gruyter Mouton.

The effect of the reference corpus on lexical and phraseological sophistication measures: Reflecting on their reliability and validity in L2 English, Elen Le Foll and Raffaella Bottini

University of Cologne and Lancaster University

Measures of lexical and phraseological sophistication can be used to evaluate learners' vocabulary comparing it to a reference corpus. Lexical sophistication metrics use a reference corpus to extract a frequency score for every word in a target text and then compute the average frequency value of the whole text. Phraseological sophistication measures follow a similar procedure: first using a reference corpus to extract an association measure (AM, e.g., Mutual Information score) for each instantiation of a selected lexico-grammatical pattern (e.g., adjective + noun) in the target text, then averaging these AM scores across the entire text (Paquot, 2019).

Brysbaert and New (2009) argue that the choice of a reference corpus that matches the register of the target texts to be analysed is crucial, and Tidball and Treffers-Daller (2008) recommend the use of spoken reference corpora in research on L2 speech. Similarly, Egbert (2017) highlights the importance of a reference corpus' situational variables. However, to date, most studies on L2 production have relied on a mixture of written and spoken corpora, without matching the mode and/or register of the reference corpora to the target texts. The fact that the representativeness of the reference corpus and its comparability to the target texts are rarely evaluated raises potential replicability and validity issues.

In this study, we measure the effect of the reference corpus on sophistication scores in L2 English, comparing three reference corpora that represent different language modes and registers. We base our analysis on the ICNALE corpus (Ishikawa, 2023), a unique dataset of L2 spoken and written English (ranging from A2 to B2 CEFR proficiency levels), whose design allows us to control for the potential effects of topic and production time. We argue that the use of a reference corpus that does not match the learner production mode, or combines different registers and/or language modes, negatively impacts the reliability and validity of lexical and phraseological sophistication measures.

References

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977–990.

Egbert, J. (2017). Corpus linguistics and language testing: Navigating uncharted waters. *Language Testing, 34*(4), 555–564.

Ishikawa, S. 2023. *The ICNALE guide: an introduction to a learner corpus study on Asian learners' L2 English*. Routledge.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*(1), 121–145.

Read, J. (2000). *Assessing vocabulary.* Cambridge University Press.

Tidball, F., & Treffers-Daller, J. (2008). Analysing lexical richness in French learner language: What frequency lists and teacher judgement can tell us about basic and advanced words. *French Language Studies, 18*(3), 299–313.

Phraseological complexity measures and the assessment of L2 writing, Agnieszka Leńko-Szymańska, Piotr Pęzik, and Michał Adamczyk

Universisty of Warsaw and University of Łodź

The recent years there has been a surge of interest in the role of phraseology in second language (L2) acquisition and assessment (cf. Wray, 2002; Meunier & Granger, eds., 2008). Learner corpora and corpus tools are particularly valuable in exploring the development of L2 phraseological competence both quantitively and qualitatively (e.g., Bestgen & Granger, 2014; Paquot, 2019). This presentation stays within this research trend. Its aim is to investigate a relationship between measures of phraseological complexity and raters' scores attributed to L2 English essays at the B2 level.

The data used in this study were 497 argumentative essays which were evaluated holistically and analytically by 5 tandems of raters. The analytical rubric included four marking categories: content, organization, accuracy and vocabulary. Four types of relational collocations were extracted from the learner texts: verb + noun, adjective + noun, verb + adverb and adverb + adjective. Six different measures of frequency and association were computed for each extracted collocation based on the reference corpus (*British National Corpus*) and the learner corpus. They were: frequencies in reference corpus and in learner corpus (per 1 million tokens), Pointwise Mutual Information (MI), LogDice, $\Delta P_{forward}$ and $\Delta P_{backward.}$ They are commonly used metrics in collocational studies.

Several statistics were computed for each L2 essay: the total number of items (types) of each collocation category and overall, as well as collocations' mean and median frequencies and association scores. Finally, two linear regression models were run, taking the phraseology-related statistics as predictors, and the raters' holistic and vocabulary marks and as the outcome variable. The models were computed for all types of relational collocations jointly and then again only for the verb+noun collocation type.

The results demonstrated that the predictive power of the models built for all the relational collocations as well as for the verb+noun collocations was very low for both the holistic and vocabulary marks. The qualitative analysis of selected high-raking and low-ranking essays demonstrated that learner texts with high scores contain instances of creative word associations which are attested in the reference corpus, but their association scores are low.

On the whole, the study points to a lack of robust relationship between measures of phraseological complexity and essay scores. This result may indicate that such a relationship does not exist. Yet, an alternative explanation will also be discussed.

References

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing, 26,* 28–41.

Meunier, F., & Granger, S. (Eds.). (2008). *Phraseology in Foreign Language Learning and Teaching*. John Benjamins Publishing Company.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*(1), 121–145.

Wray, A. (2002). *Formulaic Language and the Lexicon.* Cambridge University Press.

The figurative and phraseological component of Italian academic register, Davide Mastrantonio

Università Ca’ Foscari Venezia

“Academic Italian” can be understood as a language variety distinct both from ordinary communication and from jargons (Ferreri, 2005; Spina, 2010; D'Aguanno, 2019). As far as Italian is concerned, this variety is a rather recent object of study, still needing careful investigation; among other things, the results of this investigation are expected to affect the processes of academic literacy also from an L2/LS perspective. A promising research path seems to be a systematic analysis aimed at shedding light on the following aspects (Mastrantonio, 2022): i) the communicative functions involved in academic communication; ii) their lexical and formal counterparts; iii) the acquisitional progression of the forms and the words used. As an example, the cause-effect relationship may be expressed by juxtaposition of utterances (1) or by means of connectives (2), or it may be encoded at the verbal phrase (3); note that this variation affects both register and acquisitional aspects:

(1) è arrivato in ritardo: c’è lo sciopero dei mezzi [he was late: there’s a transportation strike]
(2) è arrivato in ritardo **perché** c’è lo sciopero dei mezzi [he was late because of a transportation strike]
(3) il suo ritardo **è (stato) determinato da**llo sciopero dei mezzi [his delay was due to a transportation strike]

In this perspective, an important role seems to be played by the figurative and phraseological component; what I mean is that academic expressions can be expressed by metaphors and can involve combinations of single lexical units:

(4) Tra le malattie croniche diffuse nell’Occidente medievale **un posto di rilievo** va sicuramente **attribuito** alla lebbra (A. Luongo, La Peste Nera, Roma, Carocci, 2022, p. 39) [among chronic diseases spread during Middle Ages, an importat role is certainly played by leprosy]

The expression *attribuire un posto di rilievo* (literally ‘to attribute/assign a prominent place’) must be analysed from several points of view. a) First of all, we are faced with a collocation (Masini, 2019), albeit one endowed with a certain degree of variability: as a matter of fact, we can also say *assegnare un posto di rilievo* or *spettare un posto di rilievo* (the last one is the inaccusative equivalent of the previous two). b) From the point of view of register and linguistic competence, *attribuire* and *assegnare* can be seen as middle-high register synonyms of dare. c) As far as the functional aspects are concerned, the expression *attribuire un posto di rilievo* is used to declare the importance of the object of discourse, hence justifying the choice of speaking of such an object. d) From a semantic point of view, we note that the expression under consideration is based on a spatial metaphor pivoting on the word *posto*. The figurative and phraseological nature of this locution becomes clearer when compared to the following rephrasing (which has the same communicative function but is less marked from a register viewpoint):

(5) Tra le malattie croniche diffuse nell’Occidente medievale **è molto importante** la lebbra.

All things considered, what I propose is to start a qualitative classification of academic expressions focusing on the figurative and phraseological component, in order to contribute to the description of the academic register and its teaching (Mastrantonio, 2021). The classification will be based on a corpus of university textbooks and research articles, predominantly taken from the humanities.

References

D'Aguanno, D. (2019a). Il lessico accademico per l'insegnamento della scrittura nelle scuole superiori. In M. Palermo, E. Salvatore (Eds.), *Scrivere nella scuola oggi. Obiettivi, metodi, esperienze*, *Atti del II Convegno ASLI Scuola (Siena, Università per Stranieri, 12-14 ottobre 2017)*, Cesati: Firenze, 93–106.

Ferreri, S. (2005). *L'alfabetizzazione lessicale: studi di linguistica educativa*. Aracne, Roma.

Masini, F. (2019). *Multi-Word Expressions and Morphology*, in *Oxford Research Encyclopedia*, Linguistics, Oxford University Press.

Mastrantonio, D. (2021). *L'italiano scritto accademico: problemi descrittivi e proposte didattiche*. *Italiano LinguaDue*, *13*(1), 348–68.

Mastrantonio, D. (2022). *Capire i testi accademici: il continuum tra comunicazione ordinaria e lingua per lo studio*. *Italiano a stranieri*, *31*, 25–30.

Spina, S. (2010). AIWL: una lista di frequenza dell'italiano accademico. In S. Bolasco et al. (Eds.), *Statistical Analysis of Textual Data*, *Proceedings of the 10th International Conference "Journées d'Analyse statistique des Données Textuelles" (9-11 June 2010) - Sapienza University of Rome*, LED: Milano, 1317–25.

Torn between L1 and L2 patterns? Collocationality levels in L2 English production by speakers of L1 Italian, Maja Miličević Petrović, Adriano Ferraresi, and Silvia Bernardini

Università di Bologna

This paper investigates the collocationality of texts produced by L1 Italian learners of L2 English in two different tasks, essay writing and translation. The two tasks are seen as two modes of *constrained language* production (Kotze, 2020): they share the constraints of bilingual activation and semi-expert proficiency, while they differ on the text production constraint, which is independent/unmediated in L2 writing and dependent/mediated in translation. The tasks are compared in light of the *Revised Gravitational Pull Hypothesis*, a theoretical framework proposed by Halverson (2017) for modelling the translation process. We look at how the factors that make up the framework ("magnetism", referring to target/second language salience, "gravitational pull", referring to source/first language prominence, and "connectivity", referring to interlingual links), may impact more or less constrained L2 production. We specifically aim to establish, across production modes: (1) whether the salience of the English collocations used is higher/lower/equal; (2) whether Italian equivalents of the English collocations used are more/less/equally strong; (3) whether collocations are more/less/equally likely to result from direct English-Italian links.

The production by learners – CEFR C1 students in a translation MA degree course – was collected in an opportunistically built corpus that favoured ecological validity over balance: the writing component comprises 106 essays on 13 corpus linguistics topics, with one essay per student (total words: 224,968), while the translation corpus contains 131 translations of 37 Italian source texts, with 19 contributing students, and topics from several domains (psychology, marketing, economics; total words: 27,393). Collocations were extracted using syntactic dependencies as codified by the spaCy+UDPipe Universal Dependency Parser (https://github.com/TakeLab/spacy-udpipe). Sixteen relations were captured that involved adjectives, nouns, verbs or adverbs (e.g., *nmod* => noun pre-modification by another noun, as in *policy development*). Collocational strength was assessed through two association measures, mutual information and logDice, following an initial filtering-out of very rare and highly specialised collocations (method loosely inspired by Durrant and Schmitt 2009). Italian equivalents were identified via lexicographic evidence and machine translation, and were also assigned association scores. Connectivity was assessed through presence in bilingual dictionaries. An initial analysis revealed that lexical association scores are significantly higher in translations than in independent writing; translations also feature more collocations with direct cross-linguistic links, while source/first language seems to affect both modes similarly. We interpret these results as pointing to a need to evaluate L2 phraseological competence in function of the task(s) it was measured on, and we discuss the relevance of L1 properties for phraseological competence assessment.

References

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177.

Halverson, S. L. (2017). Gravitational pull in translation. Testing a revised model. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds), *Empirical Translation Studies: New Theoretical and Methodological Traditions*, De Gruyter, 9–46.

Kotze, H. (2020). Converging what and how to find out why. In L. Vandevoorde, J. Dams, & B. Defrancq (Eds), *New Empirical Perspectives on Translation and Interpreting*, Routledge, 333–371.

Collocation processing in writing and translation between Chinese and English: A corpus-based and keylogging analysis, Qiuqing Qin

Università di Bologna

Collocation facilitates language development and plays an essential role in producing fluent native-like language (Feng, 2020). Processing collocations not only reflects the collocational competence and linguistic fluency but reveals the psychological effort (Ferraresi & Bernardini, 2023; Durrant & Doherty, 2010; Ellis et al., 2008; Henriksen, 2013; Siyanova, 2008; Siyanova & Schmitt 2008). This contribution is a work-in-progress report concerning native Chinese speakers' use of collocations in L1 Chinese and L2 English text production. The study aims at analysing syntactic dependencies and corpus metrics (e.g., frequency, MI score, log Dice score), combining this data with information about pauses recorded by the keylogger Inputlog. The research questions are: (1) What is the correlation between collocations and pauses? (2) Does the correlation show more evidence in L1 than that in L2? (3) Does the correlation show more evidence in translation than that in writing?

18 MA Chinese students specialized in translation and English language with B2 as the L2 proficiency were recruited. All participants were asked to conduct a set of tasks in each language: a typing test, one source-based writing and one translation. Both qualitative and quantitative methods will be employed for data analysis to find out a correlation between collocations and their preceding and within pauses by integrating data from both product- and process-oriented perspectives. The contribution will address the major methodological challenges encountered and report on preliminary result.

References

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics, 19*(4), 443–477.

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, *6*(2), 125–155.

Ellis, N. C., Simpson-Vlach, R. & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 41*(3), 375–396.

Feng, H. (2020). *Form, meaning and function in collocation. A corpus study on commercial Chinese-to-English translation*. Routledge.

Ferraresi, A., & Bernardini, S. (2023). Comparing collocations in translated and learner language: In search of a method. *International journal of Learner Corpus Research*, *9*(1), 126–154.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics, 52*(3), 229–252.

Henriksen, B. (2013). Researching L2 learners’ collocational competence and development-a progress report. In C. Bardel, B. Laufer, & C. Lindqvist (Eds.), *L2 vocabulary acquisition, knowledge and use*, 29–56.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *The Canadian Modern Language Review, 64*(3), 429–458 .

Bottom-up meets top-down: Investigating the construct validity of phraseological complexity measures through a cross-methodological comparison, Nathan Vandeweerd, Fanny Forsberg Lundell, and Klara Arvidsson

Radboud University and Stockholm University

While Paquot's (2019) construct of phraseological complexity has proved a useful index of proficiency and development in L2 French (e.g., Vandeweerd et al., 2021), the practical usefulness of any complexity measure cannot on its own speak to the *validity* of the construct itself (Pallotti, 2015). Rather, it is necessary to root these measures in "collective understandings" of what the construct means (Purpura et al., 2015). To this end, we re-examined data from a previous study in which multi-word sequences were identified in L2 French texts using a manual bottom-up procedure (Forsberg & Bartning, 2010). Each sequence from the original study was coded for syntactic structure in order to determine the extent to which structures typically analysed in phraseological complexity studies (e.g., verb + noun, adjective + noun and verb + adverb collocations) capture the diversity of multi-word phenomena exhibited in learner texts. We also investigated whether more sophisticated sequences (i.e., those with a high pointwise mutual information) were more likely to be manually identified in the original study.

The results showed that while verb + noun, adjective + noun and adverb + verb collocations made up a large proportion of manually-identified units (36% of types), other syntactic structures were also frequently identified by the authors of the original study. These included, for example, verbs + prepositions (e.g., *commence à*, 'start to'; 16.4% of types), prepositional phrases (e.g., *au fur et à mesure*, 'as things progress'; 15.2% of types) as well as nouns modified by prepositional phrases (*qualtité de vie*, 'quality of life'; 5.5% of types). We also found a small but significant effect of PMI on the likelihood that a multi-word sequence was manually identified. Together, these results provide insights into the types of units to focus on in future studies of phraseological complexity in L2 French and shed more light on the construct of phraseological complexity more generally.

References

Forsberg, F., & Bartning, I. (2010). Can linguistic features discriminate between the communicative CEFR-levels?: A pilot study of written L2 French. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing* (pp. 133–157). European Second Language Association.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(11), 117–134.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145.

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, *65*(1), 37–75.

Vandeweerd, N., Housen, A., & Paquot, M. (2021). Applying phraseological complexity measures to L2 French: A partial replication study. *International Journal of Learner Corpus Research*, *7*(2), 197–229.